Financial Frictions and Variable Markups

Antonis Tsiflis*
University of Chicago

12 November 2022

Please click here for latest version

Abstract

How costly are financial frictions in the presence of variable markups? I analyze the interaction between financial frictions and variable markups and draw implications for the ability of financial frictions to explain income per capita differences across countries. For this purpose, I build a quantitative heterogeneous-agent model of producer dynamics. Intermediate producers are born with heterogeneous permanent productivity and are also subject to transitory productivity shocks. They engage in monopolistic competition. Their outputs are aggregated into a final good such that the demand elasticity faced by intermediate producers is decreasing in their relative output, making markups increasing in relative output. In order to hire capital, an intermediate producer must take an intra-period loan from financial intermediaries, but imperfect contractual enforcement limits the loan size to a multiple of the producer's collateralizable assets. Intermediate producers are born with no collateral, but they accumulate it over time using their profits. I calibrate the model using novel tax data from Pakistan. First, I match salient features of the producers' lifecycle and the distribution of sales across producers. Second, after estimating producer-level markups, I match the aggregate markup and the empirical relationship between market shares and markups. I find that financial frictions are more costly in the presence of variable markups: a 10% reduction in financial frictions increases output per capita by 39% more when markups are variable relative to when markups are constant. This additional benefit of reducing financial frictions stems from (1) productive producers with little accumulated collateral overcoming their collateral constraints faster by charging higher markups and thus (2) relatively high markups being restricted by the heightened competition.

*The empirical parts of this paper are joint work with Faraz Hayat. I am very grateful to my advisors Veronica Guerrieri, Simon Mongey, and Rohan Kekre for their guidance and encouragement. I also thank Nancy Stokey, Harald Uhlig, Chang-Tai Hsieh, Fernando Alvarez, Mikhail Golosov, Greg Kaplan, Robert Shimer, Joe Vavra, Thomas Winberry, and seminar participants at the University of Chicago for helpful comments.
Website: https://tsiflis.economics.uchicago.edu/, Email: atsiflis@uchicago.edu

1 Introduction

There are large differences in income per capita across countries, which remain largely unexplained by differences in the quantity and quality of factors of production (Jones, 2016). Developed countries are able to produce more from any given inputs than their developing counterparts. Although limited access to productive technologies by producers in poor countries might partly explain their lower productivity, a recent literature suggests that lower aggregate productivity is driven by a suboptimal allocation of factors of production across the economy's producers.

In light of the strong positive correlation between economic and financial development (see, for example, Levine (1997)), it is hardly surprising that a large literature has attributed misallocation to frictions in financial markets. By limiting the producers' access to financing, financial frictions can constrains producers to hiring less than their desired capital and thus capital is not allocated to its most productive uses. Self-financing can prevent this misallocation: producers can accumulate internal funds and thus rely less on external financing for their capital hiring. However, self-financing is ineffective if constrained producers earn limited profits, slowing down the accumulation of internal funds.

Another source of misallocation is markup variability across producers, which is a prominent feature of developing economies (see De Loecker et al. (2016) and this paper). In an efficient economy, relative product prices reflect relative production costs and thus resources are allocated to their most productive uses. However, in the presence of variable markups, the allocative role of prices is distorted, since relative product prices also reflect relative markups.

This paper analyzes how financial frictions and variable markups interact and draws implications for the ability of financial frictions to explain income per capita differences across countries. First, how does the presence of variable markups shape the output and efficiency effects of financial frictions? The distribution of markups affects the effectiveness of self-financing for overcoming financial constraints. The markups that producers charge determine their profitability and thus the funds available to them for self-financing their capital hiring. Therefore, the distribution of markups over the producers determines the economy's efficiency. Consider the productive producers who are young and thus have not yet accumulated enough assets to overcome their financial constraint. They are constrained to hiring less than their desired capital, limiting their size. If their small size translates to low markup, then their ability to accumulate assets would be further restricted. Second, how do financial frictions shape the distribution of markups in an economy? On the one hand, financial frictions limit the size of some producers, creating size dispersion in the economy. In so far as markups are increasing in size, the size dispersion translates to markup dispersion, which distorts the signaling role of prices. On the other hand, by keeping some producers small, financial frictions limit the competition for all other producers, who thus charge higher markups.

In analyzing these effects, this paper makes theoretical, empirical, and quantitative contributions. The theoretical contribution is to study financial frictions and variable markups in a unified framework. I develop a model of producer dynamics with two main features. First, financial frictions take the form of a collateral requirement that limits the amount of capital that producers can hire. Second, markups are endogenous and increase in the producers' relative output. On the empirical side, I use novel producer-level tax data from Pakistan to estimate markups for a more representative sample of producers than those in the literature, which are limited to public firms or specific sectors of the economy. Using the markup estimates, I establish facts about the variability of markups across producers and the relation between markups and producer characteristics. Finally, on the quantitative side, I calibrate key parameters of the model to match the empirical results and compute the overall effect of the interaction of financial frictions and variable markups on the economy's output.

In the theoretical part of the paper, I build a quantitative heterogeneous-agent model of producer dynamics. A final good is produced in a competitive sector using a continuum of differentiated intermediate goods as inputs. Intermediate producers are born with heterogeneous permanent productivity and are subject to transitory productivity shocks. They engage in monopolistic competition. Their outputs are aggregated into the final good such that the demand elasticity faced by intermediate producers is decreasing in their relative output, making markups increasing in relative output. The production technology available to intermediate producers requires both labor and capital. In order to hire capital, an intermediate producer must take an intra-period loan from financial intermediaries, but imperfect contractual enforcement limits the loan size (and thus the capital hired) to a multiple of the producer's collateralizable assets. The level of financial frictions determines the collateral required for hiring capital and thus the extent of distortion in the allocation of capital across producers. Intermediate producers are born with no collateral, but they accumulate it over time through retained profits.

I combine my model with empirical analysis of novel producer-level tax data from Pakistan, to quantitatively assess the importance of the interaction between financial frictions and variable markups. Financial frictions and markups are known sources of inefficiency in the developing world (Herrala and Turk-Ariss, 2013; De Loecker and Eeckhout, 2021), so the economy of Pakistan is a fitting environment to study their interaction. Based on balance sheet data, I estimate the production function of producers and infer producer-level markups. My dataset goes beyond the typical dataset used for markup estimation, which is either limited to public companies or specific sectors of the economy. Instead, I am able to estimate markups for both public and private businesses that cover all sectors of the economy of Pakistan. The implied markup distribution has significant dispersion which is correlated with market shares. I calibrate my model to match the distribution of sales across producers, the aggregate markup, and the empirical relationship

between market shares and markups. I also take advantage of the panel dimension of my data to calibrate my model to salient features of the producers' lifecycle.

I find that financial frictions are more costly in an economy with variable markups relative to one with constant markups. Specifically, I study a 10% relaxation of the financial frictions separately in each setting and examine how its effect varies with markup variability. In both settings, the relaxation of the financial frictions more than doubles the output and consumption per capita. However, the increase is about 39% larger in the variable markups' economy than in the constant markups' economy. This difference is also reflected in aggregate productivity, which increases by about 6% more in the variable markups' economy, and the aggregate markup, which slightly drops in the variable markups' economy but is by assumption constant in the constant markups' economy.

Variable markups amplify the positive effect of reducing the financial frictions through the interaction between variable markups and self-financing. In both economies, financial frictions prevent producers from achieving their optimal size, unless they have accumulated enough assets to overcome their collateral constraints. However, in the variable markups' economy, for as long as these constrained producers remain small, they face a high demand elasticity for their products. They therefore charge a low markup, which limits their profitability and thus their ability to overcome their collateral constraints. As a result, relaxing the economy's financial frictions not only relaxes the collateral constraints of the producers directly but also improves their self-financing ability indirectly through higher markups. The additional burden that low markups have on constrained producers also implies that the unconstrained producers face lower competition. They can therefore charge higher markups, which reduces the economy's efficiency. This implies that relaxing the financial frictions in the variable markups' economy has the additional benefit of encouraging competition between producers.

My paper builds on a large literature that studies the relation between financial and economic development (see Buera et al. (2015) for a survey). My model is most closely related to that of Midrigan and Xu (2014), to which the most important change is that I introduce variable markups through a Kimball (1995) aggregator.[1] This and other papers in the literature (Buera et al., 2011; Moll, 2014) have highlighted the importance of self-financing as a viable substitute for external financing, especially in sectors that do not require large up-front investments and in environments where constrained producers earn enough profits to support fast accumulation of internal funds. In my model, where markups are increasing in producer size, constrained producers have low markups and profits, slowing down the accumulation of internal funds. As a result, the variability of markups across producers amplifies the misallocation cost of financial frictions. It follows that financial development still has an important role to play in explaining differences in economic development.

[1] Boar and Midrigan (2019) use similar modeling building blocks, in order to study how policies attempting to alleviate the inefficiency of markups impact inequality.

This paper is also motivated by the growing literature using microdata and quantitative models to document the cross-sectional heterogeneity of markups and their implications for productivity and welfare. On the empirical side, the literature has used producer-level data to document that markups vary significantly in both developed and developing economies (see, for example, De Loecker et al. (2020) and De Loecker et al. (2016), respectively). I add to this literature by documenting the variability of markups across producers in Pakistan. On the theoretical side, Edmond et al. (2021) build a quantitative model to evaluate the cost of markups in general and markup variability in particular. They find that, depending on the market structure and level of aggregate markups, the welfare cost of markups can be large, with markup variability causing 1/4 to 1/2 of it.[2] In this paper, I show that the inefficiencies stemming from markups are amplified by financial frictions, while they also amplify the inefficiencies stemming from financial frictions.

2 Motivating evidence

Before describing the model, I provide evidence for the variability of markups. I use administrative tax data from Pakistan (described in the next section), in order to estimate the production function of firms and thus firm-level markups. I find that the markup distribution exhibits significant dispersion.

2.1 Data description

I use Pakistan's complete record of anonymized firm-level annual electronic income tax filings over the period 2014-2019. The filings consist of several forms with separate reporting criteria. Other than the standard income tax form, which is filed by all the firms, all private and public limited companies are required to report their profit and loss statements and balance sheets. Additionally, firms that pay any amount of taxes at source during the fiscal year are also required to make these additional filings. Examples of such taxes are custom duties paid during import clearance, taxes on export proceeds, and taxes paid in export processing zones. Other firms, such as partnerships and small individual businesses, while not required to file the additional forms, can voluntarily do so, in order to justify their reported profits. The income tax returns include information on firm formation year, profits, and sector. The presence of the formation year in my data allows me to infer the age of each firm at the time of filing. The profit and loss statements and balance sheets report the sales, stock of capital, and expenditures on variable factors of production.

The dataset contains the 835,204 firms that filed the annual income tax at least once over the period 2014-2019 (henceforth referred to as the "full sample"). My procedure for estimating markups relies on the data contained in the profit and loss statements and the balance sheets, and leverages within-

[2] For further references to the variable markups literature, see section 2 and Edmond et al. (2021).

Table 1: Firms by type and sector

Business Category	Count	Share	Sector	Count	Share
Panel A. Estimation sample					
Individual Businesses	73,097	0.83	Agriculture & Mining	3,516	0.04
Partnerships	6,125	0.07	Manufacturing	16,883	0.19
Companies	9,020	0.10	Retail & Wholesale	22,918	0.26
			Construction	1,013	0.01
			Services	16,672	0.19
			Other	27,240	0.31
Total	88,242	1.00	Total	88,242	1.00
Panel B. Full sample					
Individual Businesses	775,546	0.92	Agriculture & Mining	32,834	0.04
Partnerships	33,871	0.04	Manufacturing	196,219	0.23
Companies	25,787	0.03	Retail & Wholesale	105,178	0.14
			Construction	11,574	0.01
			Services	159,949	0.19
			Other	329,450	0.39
Total	835,204	1.00	Total	835,204	1.00

firm intertemporal variation. Consequently, I restrict the sample to the 88,242 firms that submit the additional filings in at least 2 of the 6 years, leading to 305,191 firm-year observations.

The leftmost three columns of panel A of table 1 show the composition of the sample by firm business type. Most firms (83%) are individual businesses, companies form 10% of the sample, and partnerships and associations of persons form the remaining 7% of the sample. The rightmost three columns of panel A of table 1 show the composition of the sample by business sector. Retail and wholesale is the largest sector, with about 26% of the firms operating in it, followed by manufacturing and services, each containing roughly 19% of the firms. About 31% of the firms belong to sectors in the "other" group, which among others includes commercial importers and exporters, general stores, and distributors. The remaining firms operate in construction, and agriculture & mining.

Since the estimation sample consists of only 11% of the full sample, I document the breakdown of the full sample by business category and sector in panel B of table 1, to verify the representativeness of the estimation sample. The estimation sample has a smaller share of individual businesses and a

Table 2: Distribution of firms across business types

	Estimation sample			Full sample		
	Indiv. Bus.	Partner.	Comp.	Indiv. Bus.	Partner.	Comp.
Agri. & Mining	0.72	0.11	0.17	0.83	0.10	0.07
Manufacturing	0.60	0.14	0.26	0.91	0.05	0.04
Retail & Wholesale	0.93	0.05	0.02	0.93	0.05	0.02
Construction	0.55	0.25	0.20	0.77	0.13	0.10
Services	0.79	0.70	0.14	0.89	0.05	0.06
Other	0.93	0.03	0.04	0.97	0.02	0.01

Table 3: Firm age and profits

	Age (years)		Profit (million PKR)	
	Estimation sample	Full sample	Estimation sample	Full sample
Mean	33.85	36.13	0.72	0.256
Std dev	17.65	15.56	0.64	0.459
p10	6	16	0.28	0
p25	24	17	0.41	0
p50	35	36	0.51	0
p75	46	46	0.79	0.415
p90	56	56	1.48	0.662
N	67,194	612,729	274,344	3,479,431

higher share of both companies and partnerships, which can be justified by the fact that individual businesses are not required to file the additional forms. Yet, individual business are still very well represented in the estimation sample. In addition, the estimation sample has a considerably larger share of retail & wholesale firms and a somewhat smaller share of manufacturing firms. In order to understand why retail & wholesale are over-represented in the estimation sample, table 2 shows the distribution of the firms of each sector across business types, for the estimation and full sample, respectively. Interestingly, the share of individual businesses drops in all sectors except retail & wholesale.

Table 3 compares the distributions of age and reported profit between the full and estimation samples. While the distribution of age is fairly similar, the firms in the estimation sample report higher profit on average, with more than half the firms in the full sample reporting zero profit.

I think that the novelty of the data used for the markup estimation and the calibration of the model is one of the contributions of this paper. First, while the widely available datasets used for markup estimation in developed economies, such as Compustat (De Loecker et al., 2020), consist of only public firms, my dataset contains records of both public and private firms. Second, survey data from developing countries typically focuses on manufacturing firms (De Loecker et al., 2016), whereas table 1 shows that my dataset covers several different sectors of the economy. Third, administrative tax records, which are typically audited, are less prone to measurement error than survey data. Fourth, credit constraints and markups are known sources of inefficiency in the developing world (Herrala and Turk-Ariss, 2013; De Loecker and Eeckhout, 2021), making the use of administrative data from a developing economy with GDP per capita of 1,500 U.S. dollars particularly fitting.

2.2 Markup estimation

I follow the approach of Hall (1988) to estimate markups using firms' profit and loss statements and balance sheets. I assume that firm i in period t has access to a production technology with productivity Ω_{it}, which uses V variable inputs $\{X^v | v = 1, 2, \ldots, V\}$, such as labor, materials, and capital K_{it}, which is also assumed to be variable:

$$Q_{it} = Q_{it}(X_{it}^1, \ldots, X_{it}^V, K_{it}, \Omega_{it}),$$

I assume that firms minimize cost, which implies that the first-order condition with respect to variable input X^v in the cost minimization problem is:

$$P_{it}^{X^v} = MC_{it} \frac{\partial Q_{it}}{\partial X_{it}^v},$$

where the $P_{it}^{X^v}$ is the price of input X^v and MC_{it} is the Lagrange multiplier in the cost minimization problem and equals the marginal cost of production. By rearranging the preceding equation and defining markups as the ratio of the output price P_{it} over marginal cost, $\mu_{it} \coloneqq \frac{P_{it}}{MC_{it}}$, I get the following expression for the markup:

$$\mu_{it} = \theta_{it}^{X^v} \left(\frac{P_{it}^{X^v} X_{it}^v}{P_{it} Q_{it}} \right)^{-1}, \tag{2.1}$$

where $\theta_{it}^{X^v}$ is the output elasticity of input X^v. Since cost and sales data is observable from the income tax returns, I only need to estimate the output elasticity to compute markups. To estimate the output elasticity, I make six additional assumptions:

1. Firms have a translog production function with common parameters.
2. Firm productivity is Hicks-neutral.
3. The logarithm of firm productivity, $\omega_{it} \coloneqq \ln \Omega_{it}$, can be decomposed into three additive components: a time-invariant component η_i, a time-varying productivity component that is

Table 4: Summary statistics

Variable	N	Mean	Std dev	p10	p25	p50	p75	p90
Sales	305,191	300.2	5,999	2.783	4.050	6.110	23.50	141.4
COGS	305,191	258.4	5,606	2.025	3.264	4.974	19.90	123.5
Capital	305,191	161.8	3,580	0.350	0.600	1.260	5.483	49.42

constant across firms γ_t, and an idiosyncratic and identically distributed shock ϵ_{it}; namely, $\omega_{it} = \eta_i + \gamma_t + \epsilon_{it}$.

4. The production function can be specified using deflated values of the output, variable inputs, and capital, instead of quantities. I make this assumption because the tax data reports elements of the balance sheet in local currency.

5. I capture the role of variable inputs in the production function using the reported value of cost of goods sold (COGS), which covers the direct cost of production and, among other categories, contains the cost of labor, energy, and materials. I prefer using such a composite variable input because the breakdown of the costs by individual variable inputs is poorly reported in the tax filings. De Loecker et al. (2020) deal with this issue in the same way.

6. Observations are independently distributed across firms. In the estimation process, when computing the standard errors, I cluster at the firm level allowing for arbitrary correlations between observations for a given firm at different time periods.

Under these assumptions, I can write a firm's output as

$$q_{it} = \beta_x x_{it} + \beta_k k_{it} + \beta_{xx} x_{it}^2 + \beta_{kk} k_{it}^2 + \beta_{xk} x_{it} k_{it} + \eta_i + \gamma_t + \epsilon_{it},$$

where the lower-case variables are the logarithms of the corresponding upper-case variables (in value terms), and X denotes the cost of goods sold.

I estimate this structural equation using the observed values of capital, cost of goods sold, and sales through a linear regression with time and firm fixed effects. Table 4 shows some moments of the distributions of the relevant variables, which are measured in millions of Pakistani rupees (PKR). All variables have a long right tail as evidenced by the mean being higher than the value at the 90th percentile.

The parameters from the production function estimation are given in table 5. My estimates suggest that while the production function is log-linear in capital, as evidenced by the near zero coefficients on higher order terms involving capital stock, it is not log-linear in the cost of goods sold. Given

Table 5: Production function estimation

Variable	Coefficient
Cost	0.110***
	(0.0418)
Capital	0.0539***
	(0.00833)
Cost - Sqr	0.0263***
	(0.00139)
Capital - Sqr	0.00104***
	(0.000135)
Capital x Cost	-0.00431***
	(0.000592)
Constant	7.653***
	(0.319)
Observations	305,191
R-squared	0.994
Firm FE	YES
Year FE	YES
Cluster	Firm Level

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

the parameter estimates, I calculate the output elasticity for firm i in period t as

$$\hat{\theta}_{it}^{X} = \hat{\beta}_x + 2 \times \hat{\beta}_{xx} x_{it} + \hat{\beta}_{xk} k_{it}.$$

For example, using the logarithms of the median cost of goods sold and capital from table 4 together with the regression estimates, I find that a firm with median cost of goods sold and capital has an output elasticity with respect to cost of goods sold of 0.86. Finally, I plug the estimated value $\hat{\theta}_{it}^{X}$ and the ratio of cost of goods sold to sales into equation (2.1), to get the markup of firm i in period t.

2.3 Markup analysis

Figure 1 shows the histogram of markups in Pakistan, as estimated using the procedure of the previous section. Markups are far from constant. The distribution has a single mode slightly above

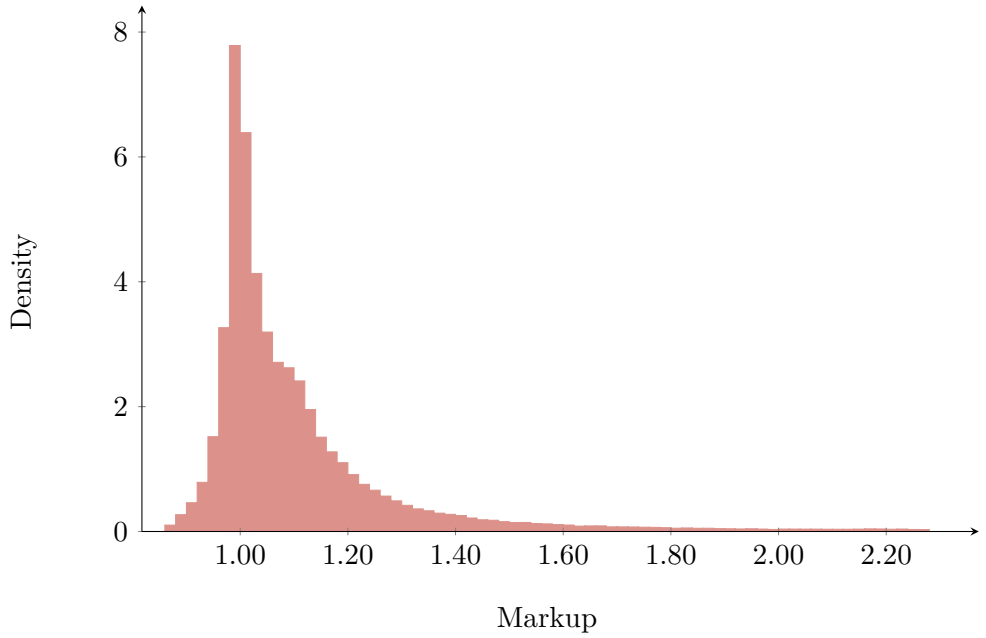

Figure 1: Pakistan's markup distribution

Notes: The bottom- and top-two percentiles of markups have been dropped for plotting purposes.

1 and exhibits a pronounced, long right tail.

Finally, I explore the relationship between markups and profitability, which will be important in my model. Higher markups need not imply higher profit; instead, they could be just recovering higher fixed costs. To check whether that is the case, I look at the correlation between a firm's profit to sales ratio and its markup. Through a linear regression of the logarithms of the profit to sales ratios on the logarithms of markups, I find that a 1% increase in markups is associated with a 0.74% increase in the profit to sales ratio (standard error of 0.02). Thus, there is a strong positive correlation between markups and profitability in my data.

3 Model

Time is discrete and denoted by subscript t. The economy is populated by four types of agents: a unit mass of financial intermediaries, a measure N^w of workers, a unit mass of intermediate good producers, and a unit mass of final good producers. The measure of workers is determined in equilibrium such that the wage rate is $w = 1$. Intermediate producers are indexed by superscript i.

With the exception of final good producers, whose problem is static, agents have access to one-period risk-free deposits at the financial intermediaries, a_t. There are also two perfectly competitive markets: one for labor, l_t, and one for capital, k_t.

3.1 Financial intermediaries

Perfectly competitive financial intermediaries receive deposits from workers and intermediate producers under a one-period contract paying a risk-free interest rate of r_t. They can freely transform deposits into capital, which they can then rent to intermediate producers. Since production depreciates capital at a rate δ, the equilibrium rental rate of capital is $r_t + \delta$.

3.2 Workers

Each worker is indexed by their deposits a_t. In each period, they inelastically supply one unit of labor.

Workers have preferences given by their lifetime utility

$$\sum_{t=0}^{\infty} \beta^t \log(C_t),$$

where β denotes the time discount factor and C_t denotes the workers' consumption of the final good. In each period, they choose how to allocate their total resources between consumption of the final good and savings:

$$C_t + a_{t+1} = w_t + (1 + r_t)a_t.$$

Total resources (right-hand side) consist of labor income from supplying labor at the wage w_t and the return on holding a_t of the risk-free asset with interest rate r_t. Workers allocate these resources between two uses (left-hand side): consumption of the final good C_t, whose price has been normalized to one, and saving through the risk-free asset a_{t+1}.

Workers face no uncertainty. In a stationary recursive equilibrium, their Euler equation implies that the interest rate must satisfy $r = \frac{1}{\beta} - 1$.

3.3 Final good producers

A unit mass of perfectly competitive producers combine the continuum of intermediate goods $\{y_t^i | i \in [0,1]\}$ to produce a final good Y_t. Their production function is implicitly defined by the Kimball aggregator:

$$\int_0^1 \mathcal{Y}\left(\frac{y_t^i}{Y_t}\right) di = 1. \tag{3.1}$$

I assume that $\mathcal{Y}(\cdot)$ takes the same functional form as in Klenow and Willis (2016):

$$\mathcal{Y}(q) \coloneqq 1 + (\bar{\sigma} - 1) \exp\left(\frac{1}{\epsilon}\right) \epsilon^{\frac{\bar{\sigma}}{\epsilon} - 1} \left[\Gamma\left(\frac{\bar{\sigma}}{\epsilon}, \frac{1}{\epsilon}\right) - \Gamma\left(\frac{\bar{\sigma}}{\epsilon}, \frac{q^{\epsilon/\bar{\sigma}}}{\epsilon}\right)\right],$$

where $\bar{\sigma} > 1$ and $\Gamma(u, z)$ is the incomplete gamma function

$$\Gamma(u, z) \coloneqq \int_z^\infty s^{u-1} \exp(-s) ds.$$

This is a parsimonious way of introducing a relationship between a producer's size and the elasticity of demand for their product. Taking the price of intermediate goods $\{p_t^i | i \in [0,1]\}$ as given, final good producers choose their intermediate inputs $\{y_t^i | i \in [0,1]\}$ to maximize profit

$$Y_t - \int_0^1 p_t^i y_t^i di$$

subject to the production technology (3.1). The solution to this problem gives the inverse demand functions for the intermediate goods

$$p_t^i = \mathcal{Y}'\left(q_t^i\right) \left[\int_0^1 \mathcal{Y}'\left(q_t^j\right) q_t^j dj\right]^{-1}, \qquad \forall i \in [0,1], \tag{3.2}$$

where $q_t^i \coloneqq y_t^i / Y_t$ denotes the relative output of producer i. The elasticity of demand as a function of relative output $\sigma(q)$ is given by

$$\sigma(q) = \bar{\sigma} q^{-\epsilon/\bar{\sigma}}.$$

The parameters $\bar{\sigma}$ and ϵ control the level of the demand elasticity and its variability with relative output q, respectively. Specifically, $\epsilon/\bar{\sigma}$ is the elasticity of the demand elasticity with respect to relative output, or the super-elasticity of demand for short. Taking the limit $\epsilon \to 0$ generates the CES case $\mathcal{Y} \to \mathcal{Y}^{CES} \coloneqq q^{\frac{\bar{\sigma}-1}{\bar{\sigma}}}$, which has constant elasticity of demand $\sigma(q) = \bar{\sigma}$. When $\epsilon > 0$, the elasticity of demand is decreasing in relative output.

3.4 Intermediate producers

Similarly to workers, intermediate producers have logarithmic period utility and time-discount factor β. However, unlike workers, each intermediate producer faces a probability ξ of exiting at the end of each period. This uncertainty is resolved at the beginning of each period. Exiting producers are replaced by an equal measure of new producers. When describing each producer's problem, for notational simplicity I drop the i superscript that indexes the producer.

Each intermediate producer is characterized by their productivity, consisting of a permanent component z and a transitory component e_t, and their holdings of the risk-free asset a_t. The permanent component of productivity is drawn upon entry from a normal distribution $G(z)$ with mean such that $\mathbb{E}[\exp(z)] = 1$ and standard deviation σ_z. The transitory component e_t follows a finite-state Markov process, whose initial state is drawn from the process' stationary distribution. Entering producers start with zero wealth. Each intermediate producer produces a differentiated good and engages in monopolistic competition with the other intermediate producers.

Intermediate producers have access to a production technology that employs labor l_t and capital k_t as factors of production:

$$y_t = \exp(z + e_t)^{1-\eta} \left(l_t^{\alpha} k_t^{1-\alpha} \right)^{\eta}. \tag{3.3}$$

The parameter α determines the output elasticity of labor relative to capital and η controls the returns to scale.

In order to hire any capital, producers must take an intra-period loan from financial intermediaries. I assume that contractual enforcement is imperfect and allows producers to appropriate a fraction $(1-\theta)$ of the loan, where $\theta \in [0, 1]$. Financial intermediaries prevent such appropriation by requiring a $(1 - \theta)$ collateral per unit loaned. Therefore, producers can only hire capital up to a multiple $1/(1 - \theta)$ of their collateralizable net worth, which consists of their asset holdings:

$$k \leq \frac{1}{1-\theta} a. \tag{3.4}$$

The level of financial development, which is governed by $\theta \in [0, 1]$, affects the constraint's tightness. Under completely undeveloped markets ($\theta = 0$), intermediate producers can hire capital only up to their current asset holdings, $k \leq a$. Under the opposite extreme of fully developed markets ($\theta = 1$), there is no constraint on how much capital intermediate producers can hire.

Taking as given the wage rate w_t and the rental rate of capital $r_t + \delta$, producers hire labor and capital, and choose their price p_t, so as to maximize their profit, defined as revenue $p_t y_t$ net of the wage bill and spending on capital:

$$\pi_t \coloneqq p_t y_t - w_t l_t - (r_t + \delta) k_t.$$

They do so subject to their production technology (3.3), the inverse demand for their output by final good producers (3.2), and the collateral constraint (3.4).

Intermediate producers also choose how to allocate their profit and the return on asset holdings $(1 + r)a_t$ between consumption of the final good C_t and new purchases of the risk-free asset a_{t+1}:

$$C_t + a_{t+1} = \pi_t + (1 + r)a_t.$$

If a producer exits at the end of the current period, then they choose zero savings, $a_{t+1} = 0$, and consume all their resources.

3.5 Equilibrium

Let $n_t(z, e, a)$ be the period-t measure of intermediate producers with permanent productivity z, transitory productivity e, and assets a. I denote the domains of these three variables by $\mathcal{Z}$, E, and A, respectively. In any period t, there is a unit mass of producers in the economy and thus

$$\int_{\mathcal{Z} \times E \times A} dn_t(z, e, a) = 1.$$

Let $\Pi : E \times E \mapsto [0,1]$, so that $\Pi(e,e')$ gives the one-period probability that the transitory productivity of an intermediate producer transitions from e to e', $\Pi^{stat} : E \mapsto [0,1]$ be the stationary distribution of the transitory productivity, $g : \mathcal{Z} \mapsto [0,\infty)$ be the density of the distribution of permanent productivity, and $a^{int} : \mathcal{Z} \times E \times A \mapsto A$ be the policy function for assets of an intermediate producer. Letting $\tilde{A}$ denote a compact subset of A and $\tilde{Z}$ denote a compact subset of $\mathcal{Z}$, the law of motion for the measure of intermediate producers is

$$\begin{aligned} n_{t+1}(\tilde{Z},e',\tilde{A}) = &\int_{\mathcal{Z}\times A} \sum_{e\in E} \Pi(e,e') \mathbb{1}\left\{z \in \tilde{Z}, a^{int}(z,e,a) \in \tilde{A}\right\} (1-\xi) dn_t(z,e,a) \\ &+ g(\tilde{Z}) \Pi^{stat}(e') \mathbb{1}\left\{0 \in \tilde{A}\right\} \xi \end{aligned} \tag{3.5}$$

The first term on the right-hand side adds up the intermediate producers in period t with permanent productivity in $\tilde{Z}$ that do not exit, transition to transitory productivity e', and choose next period assets in $\tilde{A}$, $a^{int}(z,e,a) \in \tilde{A}$. The second term captures the share of all ξ entrants that draw a permanent productivity in $\tilde{Z}$ and transitory productivity e'. This term is non-zero only in the case where $\tilde{A}$ contains zero, since entrants always start with zero assets.

A stationary recursive equilibrium consists of an interest rate r, policy functions for workers' consumption, $C^w(a)$ and holdings of the risk-free asset, $a^w(a)$, a policy function for final good producers' input use, $y^f(z,e,a)$, policy functions for non-exiting intermediate producers' consumption, $C^{int}(z,e,a)$, holdings of the risk-free asset, $a^{int}(z,e,a)$, labor use, $l^{int}(z,e,a)$, and capital use, $k^{int}(z,e,a)$, measures of intermediate producers, $n(z,e,a)$, and a measure of workers N^w such that

1. all agents' optimization problems, as described in sections 3.2, 3.3, and 3.4, are satisfied
2. the labor market clears

$$N^w = \int_{\mathcal{Z}\times E\times A} l^{int}(z,e,a) dn(z,e,a)$$

   where the left-hand side is the labor supply by the workers of total mass N^w each supplying one unit of labor and the right-hand side adds up the labor demand by the intermediate producers

3. the capital (or risk-free asset) market clears

$$a^w + (1-\xi) \int_{\mathcal{Z}\times E\times A} a^{int}(z,e,a) dn(z,e,a) = \int_{\mathcal{Z}\times E\times A} k^{int}(z,e,a) dn(z,e,a)$$

   where a^w is the equilibrium holdings by workers, the second term on the left-hand side adds up the holdings by non-exiting intermediate producers, and the right-hand side adds up the demand for capital by intermediate producers

4. the law of motion for the measure of intermediate producers in (3.5) is satisfied

4 Analytical results

Consider the cost minimization problem of an intermediate producer with permanent productivity z, transitory productivity e, and assets a. Since this is a static problem, I drop the time subscripts for notational simplicity. The optimal choices of capital and labor satisfy the following first-order conditions (FOCs):

$$w = MC\ \eta\alpha\frac{y}{l}, \tag{4.1}$$

$$r + \delta + \lambda = MC\ \eta(1-\alpha)\frac{y}{k}, \tag{4.2}$$

where MC denotes the producer's marginal cost and λ denotes the multiplier on the collateral constraint (3.4). The first equation is the FOC with respect to labor and equates the marginal cost of labor w with the product of the marginal cost MC and the marginal product of labor $\eta\alpha y/l$. Similarly, the second equation is the FOC with respect to capital, but in this case the marginal cost of capital $r+\delta$ is augmented by the multiplier on the collateral constraint. The marginal cost is given by

$$MC = y^{\frac{1-\eta}{\eta}} \exp(z+e)^{\frac{\eta-1}{\eta}} \left(\frac{r+\delta+\lambda}{\eta(1-\alpha)}\right)^{1-\alpha} \left(\frac{w}{\eta\alpha}\right)^{\alpha}.$$

The direct dependence on output captures the fact that decreasing returns to scale make production more costly as output increases. In addition, the marginal cost depends on output through the multiplier on the collateral constraint λ. For unconstrained producers, for whom the collateral constraint is not binding, the multiplier is zero, but it is positive for constrained producers. Specifically,

$$\lambda = \max\left\{\frac{1-\alpha}{\alpha} y^{\frac{1}{\eta\alpha}} \exp(z+e)^{\frac{\eta-1}{\eta\alpha}} \left(k^{max}\right)^{-\frac{1}{\alpha}} w - (r+\delta), 0\right\},$$

where $k^{max} \coloneqq \frac{1}{1-\theta}a$ is the maximum capital allowed by the collateral constraint. The multiplier is increasing in output and the marginal cost is increasing in the multiplier. Intuitively, a constrained producer cannot hire more capital and thus can only increase output by hiring more labor. Due to the decreasing returns to labor, the marginal cost is increasing in output. In the case of a constrained producer, the multiplier is decreasing in assets a and thus the marginal cost is decreasing in assets too. The reason is that a producer with higher assets can (and does) hire more capital and thus does not run as much into the decreasing returns to labor.

The producer sets their price at a markup m over marginal cost, $p = mMC$, where the markup is a decreasing function of the elasticity of demand σ:

$$m = \frac{\sigma}{\sigma - 1} = \frac{\bar{\sigma}}{\bar{\sigma} - q^{\frac{\epsilon}{\bar{\sigma}}}}.$$

Since the elasticity of demand is decreasing in relative output q, the markup is increasing in relative output.

Combining the above elements leads to the following proposition.

Proposition. *For an intermediate producer,* $\frac{\partial q}{\partial a} \geq 0$ *and* $\frac{\partial m}{\partial a} \geq 0$*, with strict inequalities if the producer is constrained.*

An increase in the assets of a constrained producer allows them to hire more capital. This reduces their marginal cost and thus their price, allowing them to capture a larger market share. As a result, they face a lower elasticity of demand and thus they charge a higher markup.

5 Quantitative analysis

In order to make further progress in understanding the interaction of financial frictions and markups, I numerically solve for the equilibrium of my economy. In particular, I calibrate my model to the economy of Pakistan and perform comparative statics with respect to the level of financial development, θ. I then uncover the role of variable markups by performing the same comparative statics in an economy with constant markups and comparing the results.

5.1 Parameterization

I choose the period length of my model to be one year. I set the discount rate β to 0.92, which implies a real interest rate of about 8.7%. I assume that capital depreciates at a rate $\delta = 0.06$, that the parameter controlling the output elasticity of labor relative to capital in the intermediate producers' production function is $\alpha = 2/3$, and that the probability of exit for intermediate producers is $\xi = 0.1$.

The remaining parameters of the model with constant markups (that is, with the exception of ϵ, which controls the super-elasticity of demand and is set to zero in the constant markups economy) are jointly calibrated to the economy of Pakistan.[3] For this purpose, I choose empirical targets calculated using either the data described in section 2.1 or, when this is not possible, aggregate data from external sources. Table 6 summarizes the parameter values chosen and table 7 displays the values that the calibration target take in both the data and the model.

I assume that the intermediate producers' transitory productivity follows a discretized AR(1) process with persistence ρ and normally distributed shocks with mean zero and standard deviation σ_ϵ. These productivity parameters, together with the standard deviation of the permanent productivity component σ_z, control most closely the dispersion of output and output growth across producers

[3] I choose to use the constant markups version of my economy for most of the calibration because solving the economy with variable markups is significantly more computer-time intensive.

Table 6: Parameterization

		Benchmark
Discount factor	β	0.92
Capital depreciation	δ	0.06
Production function's labor power	α	0.67
Exit probability	ξ	0.1
Transitory productivity's persistence	ρ	0.85
Transitory productivity shocks' s.d.	σ_ϵ	1.90
Permanent productivity s.d	σ_z	0.97
Span of control	η	0.92
Collateral constraint	θ	0.76
Average demand elasticity	$\bar{\sigma}$	10.09
Super-elasticity	$\epsilon/\bar{\sigma}$	0.32

and the autocorrelation of the producers' output. I therefore target the cross-sectional standard deviation of sales and sales growth, as well as the one-year firm-level autocorrelation of sales, in my data.

To pin down the span of control η, I use the aggregate profit share in my data of 0.14 as a target. The financial development parameter θ determines debt issuance, so I target the debt-to-output ratio in Pakistan's manufacturing sector. Given that debt is not reliably reported in my data, I use the value of loans granted to the manufacturing sector as reported by the State Bank of Pakistan. Coupled with the gross domestic product of manufacturing as reported by the World Bank, I get a debt-to-output ratio of 0.69.

The average demand elasticity for the output of intermediate producers, $\bar{\sigma}$, is set to 10.09, in order to match the cost-weighted aggregate markup in Pakistan's manufacturing sector of 1.11. Under variable markups, the elasticity of demand for the output of intermediate producers and therefore the markup distribution are determined by not only the average demand elasticity $\bar{\sigma}$ but also the super-elasticity $\epsilon/\bar{\sigma}$. Specifically, $\bar{\sigma}$ controls the level of markups and ϵ determines how the sales distribution translates to the markup distribution. My model implies the following relationship between producer-level markups μ_t^i and market shares $p_t^i y_t^i / Y_t$:

$$\frac{1}{\mu_t^i} + \log\left(1 - \frac{1}{\mu_t^i}\right) = \text{constant} + \frac{\epsilon}{\bar{\sigma}} \log\left(\frac{p_t^i y_t^i}{Y_t}\right).$$

I estimate this regression in my data and set ϵ such that $\epsilon/\bar{\sigma}$ matches its regression estimate of 0.32.

Table 7: Calibration targets

Moment	Data	Model
Std. dev. of sales	1.90	2.18
Std. dev. of sales growth	0.66	0.68
1-year sales' autocorrelation	0.94	0.94
Aggregate profit share	0.14	0.24
Debt-to-output	0.69	0.88
Average markup	1.11	1.11
Super-elasticity	0.32	0.32

5.2 The effect of financial frictions and markups

I now show how financial frictions and markups enter the intermediate producers' optimality conditions, as well as how they are aggregated and affect the aggregate outcomes of the economy.

Combining the first-order condition for the producers' labor choice (4.1) with the fact that they set their price at a markup over marginal cost implies that the markup m^i is a wedge that reduces the labor share of revenue (left-hand side) below the production elasticity of labor, $\eta\alpha$:

$$\frac{wl^i}{p^i y^i} = \frac{\eta\alpha}{m^i}.$$

Aggregating over the producers gives a similar expression for the aggregate labor share:

$$\frac{wL}{Y} = \frac{\eta\alpha}{M},$$

where the aggregate labor wedge

$$M = \int_0^1 m^i \frac{l^i}{L} di$$

is the aggregate markup, defined as the labor-weighted average of the producer-level markups.

Using the first-order condition for the capital choice (4.2), a similar expression for the capital share of revenue can be derived:

$$\frac{(r+\delta)k^i}{p^i y^i} = \frac{\eta(1-\alpha)}{m^i \tilde{\lambda}^i},$$

where

$$\tilde{\lambda}^i := \frac{r+\delta+\lambda^i}{r+\delta}$$

is the ratio of the shadow cost of capital to the cost of capital absent the collateral constraint. In this case, the wedge between the capital share of revenue (left-hand side) and the production elasticity

of capital, $\eta(1-\alpha)$, is driven by not only the markup but also how binding the collateral constraint is, as captured by $\tilde{\lambda}^i$. Aggregating over the producers gives an expression for the aggregate capital share:

$$\frac{(r+\delta)K}{Y} = \frac{\eta\alpha}{\Lambda},$$

where the aggregate wedge

$$\Lambda = \int_0^1 m^i \tilde{\lambda}^i \frac{k^i}{K} di$$

is the capital-weighted average of the producer-level wedges.

Table 8 reports the model-implied moments of the distribution of the markup (or labor wedge) and the capital wedge, weighted by each producer's labor and capital share, respectively. The average markup is 1.21, implying that the aggregate labor share is $1/1.21 = 17\%$ lower than the share implied by the production elasticity of labor. The capital wedge is larger on average and more dispersed. Since $\tilde{\lambda}^i \geq 1$, the capital wedge $m^i\tilde{\lambda}^i$ is by definition larger than the markup. If all producers were unconstrained, then the two wedges would coincide. However, some producers are constrained, leading to a mass of relatively high capital wedges. As a result, the average capital wedge is significantly higher than the average markup wedge and the capital wedge distribution is more dispersed than the markup distribution. The average capital wedge is 1.54, implying that the aggregate capital share is $1/1.54 = 35\%$ lower than the share implied by the production elasticity of capital.

Figure 2 plots the shadow cost of capital, $r + \delta + \lambda^i$, against assets, for two intermediate producers with different productivities. Conditional on productivity, producers with higher assets face a laxer collateral constraint and thus the shadow cost of capital is decreasing in assets. Conditional on assets, more productive producers are more constrained (have a higher shadow cost of capital), since their optimal scale of production is higher and thus want to employ more capital. This relationship is also reflected in the mass of constrained producers that have high markups in figure 3, which plots separately the markup distribution of constrained and unconstrained firms. These producers are very productive and thus charge high markups despite being constrained.

Figure 4 plots the markup policy of intermediate producers for two different levels of productivity. At high levels of assets, producers are unconstrained and thus they can charge their optimal markup. As observed in figure 2, the required level of assets for a producer to be unconstrained is increasing in productivity, due to the increase in the optimal scale of production. Conditional on the level of assets, producers with higher productivity have lower marginal cost and can therefore set lower prices, capture a larger market share, and charge higher markups.

Figure 5 illustrates the saving policy of intermediate producers for a given permanent productivity z but two different levels of idiosyncratic productivity e. At low levels of assets, producers choose

Table 8: Distribution of wedges

	Markup	Capital wedge
Mean	1.21	1.54
Std dev	0.07	0.93
Min	1.01	1.01
p1	1.06	1.09
p5	1.1	1.13
p10	1.12	1.16
p25	1.15	1.21
p50	1.2	1.31
p75	1.25	1.58
p90	1.3	2.1
p95	1.33	2.6
p99	1.4	4.65
Max	1.51	305.51

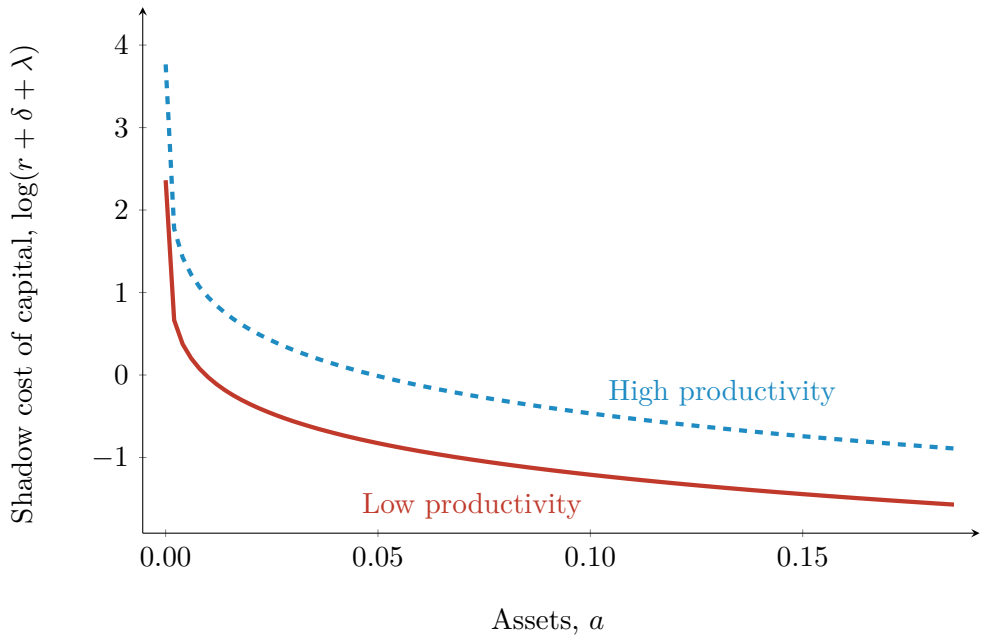

Figure 2: Intermediate producers: Shadow cost of capital

Notes: Both curves correspond to producers with the highest permanent productivity z. The "High productivity" producer also has the highest idiosyncratic productivity e, while the "Low productivity" producer has the fourth highest idiosyncratic productivity.

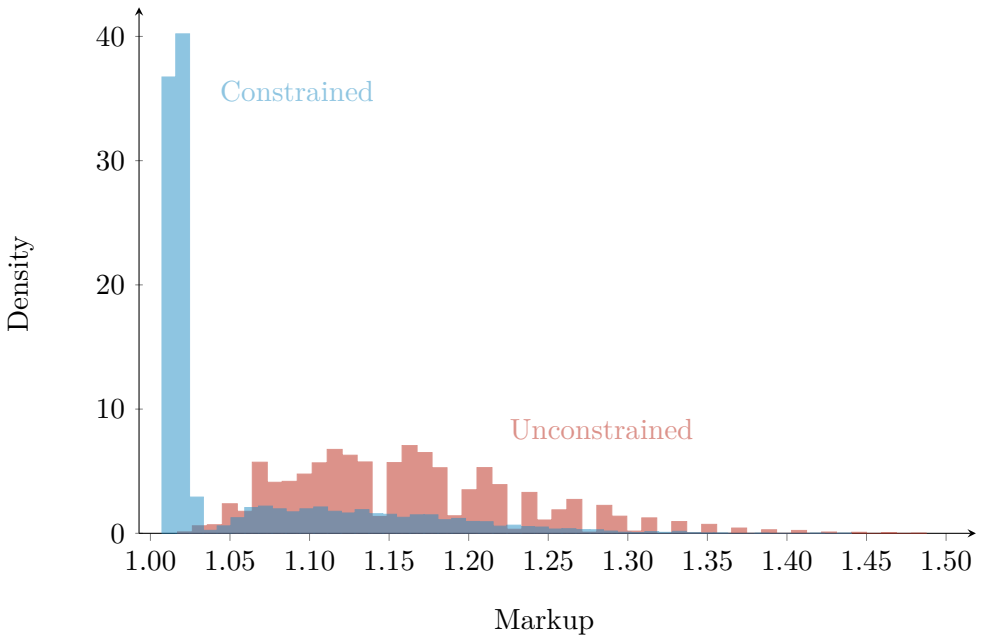

Figure 3: Intermediate producers: Markup distribution by constraint

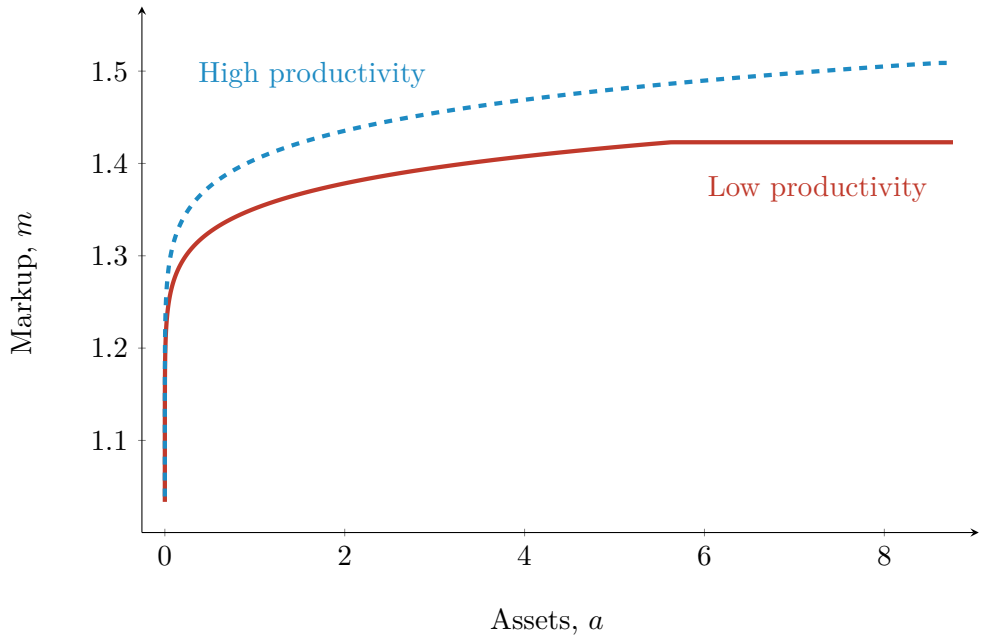

Figure 4: Intermediate producers: Markup policy

Notes: Both curves correspond to producers with the highest permanent productivity z. The "High productivity" producer also has the highest idiosyncratic productivity e, while the "Low productivity" producer has the second highest idiosyncratic productivity.

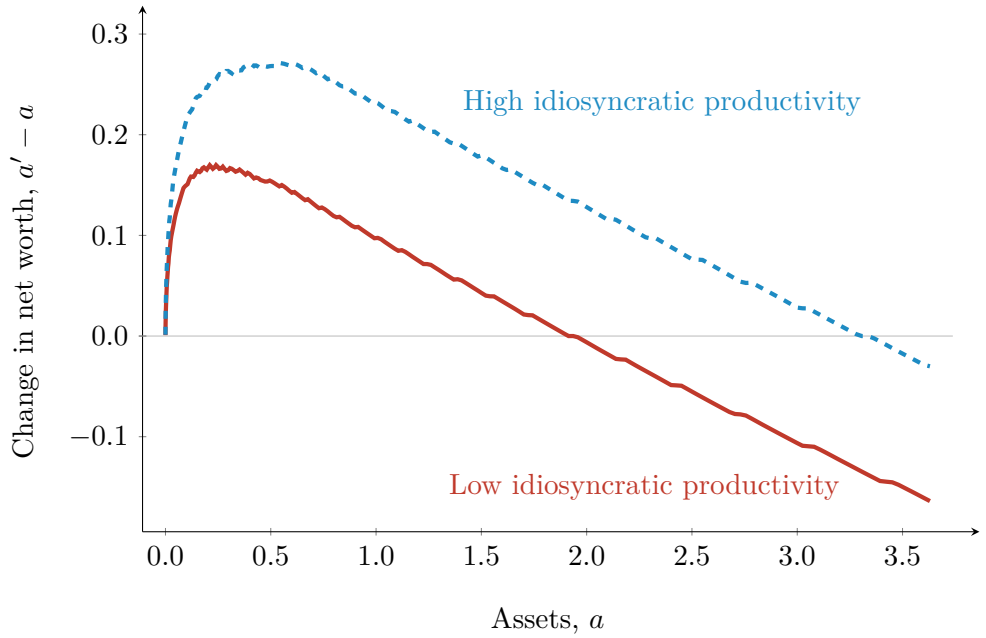

Figure 5: Intermediate producers: Saving policy

Notes: Both curves correspond to producers with the highest permanent productivity z. The "High productivity" producer also has the highest idiosyncratic productivity e, while the "Low productivity" producer has the second highest idiosyncratic productivity.

to build up their asset stock, in order to relax future collateral constraints. As the asset stock increases, future collateral constraints become less binding and thus producers save less. Indeed, at high levels of assets, they reduce their asset stock. Given their smoothing motive and positive probability of exit, producers will not choose to accumulate enough assets to remain unconstrained under any realization of the future.

5.3 Surviving producer's lifecycle

Figure 6 plots the lifecycle of one million intermediate producers. Specifically it shows the time series of their average assets and the fraction of the survivors that are constrained, starting from the time they are born. The typical producer accumulates assets in the first years of their life, in order to overcome their collateral constraint. By the tenth year of their life, almost all producers have managed to become unconstrained. Therefore, in equilibrium, the set of constrained producers consists of young producers that have not yet been able to accumulate enough assets in order to become unconstrained.

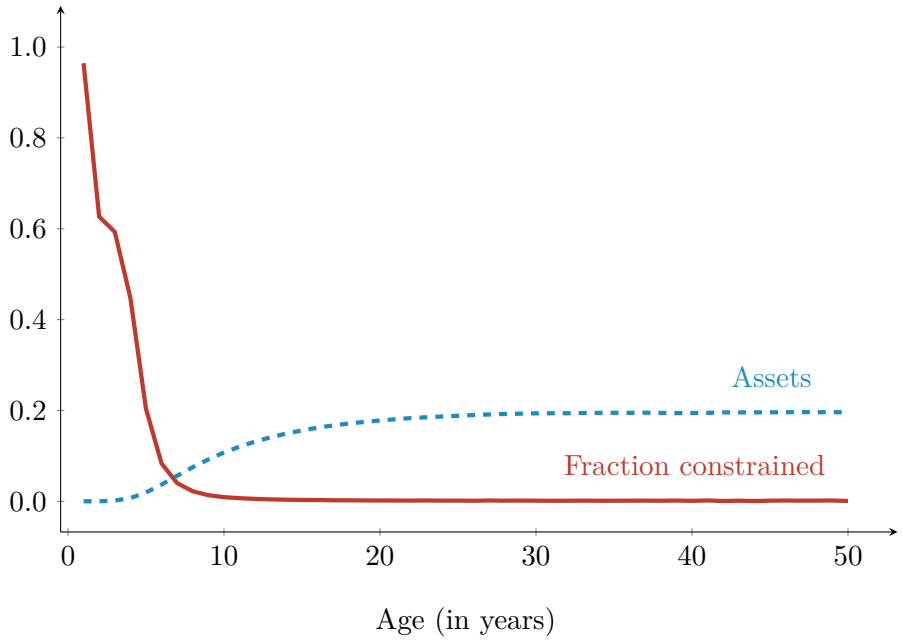

Figure 6: Surviving intermediate producer's lifecycle

Notes: The figure is based on a lifecycle simulation of a cohort of one million intermediate producers. At each age, the figure shows the average assets held by the surviving producers and the fraction of the surviving producers that are constrained.

5.4 Comparison with efficient allocation

The aggregate production function can be written as

$$Y = Z\left(L^{\alpha}K^{1-\alpha}\right)^{\eta},$$

where the aggregate productivity Z is given by

$$Z^{-1} = \left(\int_0^1 \left(\frac{q^i}{\tilde{z}^i}\right)^{\frac{1}{\eta}} (\tilde{\lambda}^i)^{-\alpha} di\right)^{\eta(1-\alpha)} \left(\int_0^1 \left(\frac{q^i}{\tilde{z}^i}\right)^{\frac{1}{\eta}} (\tilde{\lambda}^i)^{1-\alpha} di\right)^{\eta\alpha}$$

and $\tilde{z}^i \coloneqq \exp(z^i + e^i)^{1-\eta}$ denotes the total factor productivity of producer i. Financial frictions enter the expression for aggregate productivity directly through $\tilde{\lambda}^i$ and generate efficiency losses. At the same time, both financial frictions and variable markups distort the allocation of output across producers, as captured by q^i, leading to further efficiency losses.

5.5 How costly are financial frictions in the presence of variable markups?

I now compare the effect of a relaxation of financial frictions in an economy with variable markups relative to one with constant markups. Specifically, I perform comparative statics with respect to degree of financial development θ and present the results in table 9. For each economy, I show the

Table 9: Comparative statics with respect to financial development

	ϵ =3.23			CES			
	θ_{low}	θ_{high}	%Δ	θ_{low}	θ_{high}	%Δ	%Δ-in-Δ
Consumption	0.04	0.09	115.43	0.24	0.49	102.29	12.84
Debt (gross)	0.03	0.09	158.54	0.21	0.51	146.94	7.89
Debt (gross) to capital	0.61	0.68	10.52	0.6	0.68	12.32	-14.61
Debt (gross) to output	0.77	0.92	19.24	0.78	0.94	21.26	-9.48
Capital	0.06	0.13	133.93	0.34	0.75	119.86	11.74
Labor	0.02	0.05	117.98	0.15	0.3	103.65	13.82
Output	0.04	0.1	116.82	0.26	0.54	103.65	12.7
Capital share	0.18	0.2	7.89	0.19	0.2	7.96	-0.81
Labor share	0.51	0.51	0.54	0.55	0.55	0.0	105543.07
Profit share	0.31	0.29	-5.58	0.26	0.24	-5.86	-4.84
Frac. constrained	0.92	0.88	-4.33	0.9	0.85	-5.23	-17.18
Output per capita	0.04	0.09	111.27	0.23	0.41	79.86	39.32
Consumption per capita	0.04	0.09	109.91	0.21	0.38	78.66	39.73
Investment per capita	0.0	0.01	127.94	0.02	0.03	94.17	35.86
Agg. markup	1.21	1.21	-0.65	1.11	1.11	0.0	-Inf
Agg. capital wedge	1.67	1.54	-7.31	1.62	1.5	-7.37	-0.75
Agg. productivity	1.11	1.15	3.52	1.2	1.24	3.32	5.78

CES: θ_{low} =0.68, θ_{high} =0.76

Kimball: θ_{low} =0.68, θ_{high} =0.76

results for two levels of θ, θ_{low} and θ_{high}, and the percentage difference between them. The last column of the table shows the percentage difference in the percentage differences. The value of θ_{high} corresponds to the value implied by the calibration, while the value of θ_{low} corresponds to a 10% drop from that value. Reducing the level of financial frictions from θ_{low} to θ_{high} corresponds to an about 20% increase in aggregate debt-to-output.

Focusing on the steady states corresponding to high financial development, θ_{high}, we see that the output per capita in the constant markups economy is more than four times higher than in the variable markups economy. This can be explained by the constant markups economy having more than five times higher capital and six times higher labor. The higher output per capita is also a result of the aggregate productivity being about 8% higher in the constant markups economy. The lower inefficiency is also reflected in the aggregate wedges, with the variable markups economy

having about 9% higher aggregate markup and 2.6% higher aggregate capital wedge. The share of producers that are constrained is also slightly higher in the variable markups economy.

In both economies, the degree of financial frictions plays an important role in the determination of the equilibrium. Most of the producers are constrained. Specifically, in the more financial developed economies more than 85% of the producers are constrained, while in the less financial developed economies this percentage rises to more than 90%. The effect of reducing the financial frictions is similar when taking into account the intensive margin too: aggregate capital wedge drops by about 7%. The less constrained economies also have more than double the output of the respective more constrained economies. This is a result of the more than doubling of aggregate capital and labor, as well as a more than 3% increase in aggregate productivity.

Comparing the effect of relaxing the collateral constraints between the constant and variable markups economies, I find that the presence of variable markups amplifies the benefit of reducing the financial frictions. The percentage change in output per capita is 39% larger in the variable markups economy than in the constant markups economy. I now turn to explaining this difference.

5.6 Decomposing the effect of each aggregate wedge

In this section, I use the economy's aggregate equilibrium conditions, in order to better understand the mechanism through which a change in financial frictions affects the aggregate variables and how this mechanism differs with the level of markup variability. Specifically, given the prices $\{r, w\}$, which are simple functions of the model parameters, and the aggregate wedges $\{M, \Lambda, Z\}$, the aggregate variables of the economy $\{Y, K, L\}$ are determined by a system of equations consisting of the definitions of the three wedges:

$$\begin{aligned} Y &= Z\left(L^{\alpha}K^{1-\alpha}\right)^{\eta}, \\ \frac{wL}{Y} &= \frac{\eta\alpha}{M}, \\ \frac{(r+\delta)K}{Y} &= \frac{\eta(1-\alpha)}{\Lambda}. \end{aligned}$$

This system of equations is log-linear and thus, after talking logarithms, can be solved for $\{\log Y, \log L, \log K\}$ as functions of the aggregate wedges:

$$\begin{aligned} \log Y &= \frac{1}{1-\eta}\log Z - \frac{\eta\alpha}{1-\eta}\log M - \frac{\eta(1-\alpha)}{\eta\alpha}\log\Lambda + C_Y, \\ \log L &= \frac{1}{1-\eta}\log Z - \frac{1-\eta(1-\alpha)}{1-\eta}\log M - \frac{\eta(1-\alpha)}{\eta\alpha}\log\Lambda + C_L, \\ \log K &= \frac{1}{1-\eta}\log Z - \frac{\eta\alpha}{1-\eta}\log M - \frac{1}{\alpha}\log\Lambda + C_K, \end{aligned}$$

where the $\{C_Y, C_L, C_K\}$ are functions only of the production function parameters, η and α, and the input prices, $r+\delta$ and w. Aggregate output, labor, and capital are all increasing in the aggregate

Table 10: Decomposition by aggregate wedge

	ϵ =3.23				CES			
	Total	Z	M	Λ	Total	Z	M	Λ
Y	0.78	0.44	0.05	0.3	0.71	0.41	0.0	0.3
L	0.79	0.44	0.06	0.3	0.71	0.41	0.0	0.3
K	0.86	0.44	0.05	0.37	0.79	0.41	0.0	0.37

productivity of the economy $\log Z$ and decreasing in the aggregate markup $\log M$ and the aggregate capital wedge $\log \Lambda$.

I use this representation of the economy to understand how important each wedge is for translating a change in the level of financial frictions to a change in aggregate output, labor, and capital. The production function parameters and input prices are independent of the variability of markups ϵ and the level of financial frictions θ. Therefore, all differences in aggregate output, labor, and capital between economies with different level of financial frictions θ and/or markup variability ϵ can be attributed to differences in the aggregate wedges. Table 10 revisits the comparative statics of Table 9. It decomposes the increase in $\{\log Y, \log L, \log K\}$, caused by the improvement in the level of financial development θ, to each aggregate wedge.

The effect of the change in the aggregate capital wedge is the same in the two economies. As discussed earlier, the aggregate capital wedge is given by

$$\Lambda = \int_0^1 m^i \tilde{\lambda}^i \frac{k^i}{K} di.$$

In the CES economy, all producers charge the same markup: $m^i = m$, $\forall i$. Therefore, the change in the aggregate capital wedge caused by a change in θ is a result of changes in the extent to which producers are constrained, $\tilde{\lambda}^i$, and the allocation of capital across producers, k^i/K. However, when markups are variable, the change in the aggregate capital wedge can also be caused by a change in the covariance between markups and the extent to which producers are constrained. Despite this difference, I find that the change in θ changes the aggregate capital wedge by the same amount in the two economies.

Instead, the larger increase in output observed in the variable markups' case is attributable to the other two wedges. In the CES case, the aggregate markup is constant by assumption and thus has no effect. However, in the variable markups economy, the reduction in the financial frictions reduces the aggregate markup and this increases output, labor, and capital. In addition, aggregate productivity increases by more in the variable markups economy, boosting the effect of financial development on output, labor, and capital.

5.7 Change in the lifecycle of producers

I now investigate further the effect on financial development on aggregate efficiency and how it depends on markup variability. In both the variable and the constant markups economy, the reduction in financial frictions allows producers to overcome their collateral constraint faster. In order to compare the magnitude of this effect in the two economies, I focus on the lifecycle of a cohort of producers that enter the economy at the same time. Table 11 shows the fraction of the surviving cohort that is constrained in the low financial frictions economy relative to the high financial frictions economy, five periods after birth. Each row corresponds to a different level of permanent productivity. For the three highest levels of permanent productivity, the reduction of the financial frictions reduces the fraction of constrained producers by more in the variable markups economy than in the constant markups economy. For example, in the highest permanent productivity group, only 48% of the producers that are constrained when financial frictions are high remain constrained when financial frictions are low, compared to 61% in the constant markups economy. The opposite is true for the lowest permanent productivity groups: the reduction in financial frictions helps producers more in the constant markups economy than in the variable markups economy. It follows that the reduction of financial frictions is particularly important in the variable markups case.

The reason that the reduction in financial frictions benefits the most productive producers particularly in the variable markups economy can be understood by looking at the interaction of financial frictions and variable markups at the producer level. Independently of the variability of markups, constrained producers are restricted in size. This harms their profitability and thus their ability to build their asset stock and overcome their collateral constraint. However, when markups are variable, the constrained producers' small size implies that they also charge low markups. As a result, their profitability and their ability to overcome their constraint is harmed further. Therefore, reducing the financial frictions is more beneficial in the presence of variable markups. This is particularly true for high productivity producers, who (conditional on assets) have a higher optimal scale of production and thus are more likely to be constrained.

Table 11: Relative fraction constrained by permanent productivity in period $t = 5$

z index	ϵ =3.23	CES	Δ
1	0.89	0.69	0.19
2	0.88	0.68	0.21
3	0.87	0.66	0.2
4	0.78	0.71	0.07
5	0.63	0.68	-0.04
6	0.65	0.69	-0.04
7	0.75	0.66	0.09
8	0.67	0.65	0.02
9	0.54	0.58	-0.04
10	0.55	0.58	-0.03
11	0.48	0.61	-0.13

5.8 Change in the markup distribution

Given the importance of the aggregate markup in explaining the difference in the increase in output between the variable markups and constant markups economy, I now analyze the effect that the reduction of the financial frictions has on the markup distribution in the variable markups economy. Table 12 shows moments of the (unweighted) markup distribution for the two levels of financial frictions used in the comparative statics. Although the changes in the markup distribution are small in magnitude, I find that the markups at the top of the distribution fall with financial development. As financial frictions are reduced, young producers are less constrained and are able to capture larger market shares. This creates more competition and thus lower market shares and markups for older producers. Indeed, some older producers were able to charge high markups because they had accumulated enough assets to be unconstrained, even if they were not particularly productive. Preventing such high markups is beneficial for the economy.

Table 12: Markup distribution moments by θ

	θ_{low}	θ_{high}
Mean	1.06	1.07
Std dev	0.07	0.07
Min	1.01	1.01
p5	1.01	1.01
p10	1.02	1.01
p25	1.02	1.02
p50	1.02	1.02
p75	1.1	1.11
p90	1.18	1.18
p95	1.23	1.22
Max	1.54	1.51

θ_{low} =0.68, θ_{high} =0.76

6 Conclusion

I analyze the interaction between financial frictions and variable markups and draw implications for the ability of financial frictions to explain income per capita differences across countries. For this purpose, I build a quantitative heterogeneous-agent model of producer dynamics with two main elements. First, financial frictions take the form of imperfect contractual enforcement. In order to hire capital, producers must take an intra-period loan, but imperfect contractual enforcement limits the loan size (and thus the capital hired) to a multiple of the producer's collateralizable assets. Second, producers produce differentiated goods and engage in monopolistic competition, leading to markups. These goods are aggregated into a final good such that the demand elasticity faced by intermediate producers is decreasing in their relative output, making markups increasing in relative output. I combine the model with empirical analysis of novel producer-level tax data from Pakistan. Specifically, I calibrate the model to match the empirical distribution of sales across producers, the aggregate markup, the empirical relationship between market shares and markups, and salient features of the producers' lifecycle.

I find that financial frictions are more costly in an economy with variable markups relative to one with constant markups: a 10% reduction in financial frictions increases output per capita by 39% more when markups are variable relative to when markups are constant. Variable markups amplify the positive effect of reducing the financial frictions through the interaction between variable markups and self-financing. Independently of the variability of markups, financial frictions prevent

producers from achieving their optimal size, unless they have accumulated enough assets to overcome their collateral constraints. However, in the presence of variable markups, for as long as these constrained producers remain small, they face a high demand elasticity for their products. They therefore charge a low markup, which limits their profitability and thus their ability to overcome their collateral constraints. As a result, relaxing the economy's financial frictions not only relaxes the collateral constraints of the producers directly but also improves their self-financing ability indirectly through higher markups. The additional burden that low markups have on constrained producers also implies that the unconstrained producers face lower competition. They can therefore charge higher markups, which reduces the economy's efficiency. This implies that relaxing the financial frictions in the variable markups' economy has the additional benefit of encouraging competition between producers.

Overall, financial frictions can play a central role in determining the aggregate productivity of an economy. This role is elevated in the presence of markup heterogeneity across producers. In light of the importance of aggregate productivity in accounting for differences in output per capita across countries, financial development appears to be a prime determinant of economic development.

References

Boar, C. and V. Midrigan (2019): "Markups and inequality," Working paper, National Bureau of Economic Research.

Buera, F. J., J. P. Kaboski, and Y. Shin (2011): "Finance and development: A tale of two sectors," *American economic review*, 101, 1964–2002.

——— (2015): "Entrepreneurship and Financial Frictions: A Macrodevelopment Perspective," *Annual Review of Economics*, 7, 409–436.

De Loecker, J. and J. Eeckhout (2021): "Global market power," Working paper.

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

De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016): "Prices, markups, and trade reform," *Econometrica*, 84, 445–510.

Edmond, C., V. Midrigan, and D. Y. Xu (2021): "How costly are markups?" Working paper.

Hall, R. E. (1988): "The relation between price and marginal cost in US industry," *Journal of political Economy*, 96, 921–947.

Herrala, R. and R. Turk-Ariss (2013): "Credit Constraints, Political Instability, and Capital Accumulation," Working paper, International Monetary Fund.

Jones, C. I. (2016): "The facts of economic growth," in *Handbook of macroeconomics*, Elsevier, vol. 2, 3–69.

Kimball, M. S. (1995): "The quantitative analytics of the basic neomonetarist model," *Journal of Money, Credit and Banking*, 27, 1241–1277.

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

Levine, R. (1997): "Financial development and economic growth: views and agenda," *Journal of economic literature*, 35, 688–726.

Midrigan, V. and D. Y. Xu (2014): "Finance and misallocation: Evidence from plant-level data," *American economic review*, 104, 422–58.

Moll, B. (2014): "Productivity losses from financial frictions: Can self-financing undo capital misallocation?" *American Economic Review*, 104, 3186–3221.