Productivity, Investment, and Wealth Dynamics under Financial Frictions*

Alvaro Aguirre† Matias Tapia† Lucciano Villacorta†

Abstract

We develop a framework to provide direct microeconomic evidence of the mechanisms underlying macroeconomic models of firm dynamics with financial frictions. Using administrative data, we non-parametrically estimate firm-level productivity and its effect on firms' investment and saving decisions in a unified framework that explicitly allows for financial frictions. The productivity process is largely non-linear, with larger persistence for more productive firms. We uncover heterogeneous responses of investment and wealth accumulation to productivity shocks. Our estimates are consistent with the presence of both collateral-based and earning-based constraints and the existence of a self-financing channel emphasized in quantitative macro models.

JEL classification: C33, E23, O11, L0

Keywords: Investment, wealth, self-financing, financial frictions, productivity.

*We are grateful to Daniel Ackerberg, Manuel Arellano, Stephane Bonhomme, Richard Blundell, Paco Buera, Andrea Caggese, Emmanuel Farhi, Manuel Garcia Santana, Virgiliu Midrigan, Ben Moll, Josep Pijoan-Mas, Yongs Shin, Chad Syverson, Alonso Villacorta, Gianluca Violante, Fabrizio Zilibotti and attendees at the Conference on econometric methods and empirical analysis of microdata in honor of Manuel Arellano, STEG Workshop on Firms, Frictions, and Spillovers 2022, Midwest Macro Meetings, the BSE Summer Forum, on Financial Shocks, Channels, and Macro Outcomes, SED, LACEA-LAMES, the 26th International Panel Data Conference, the 2020 World Congress of the ES, EEA 2020, Santiago Macro Workshop, PUC Chile, CEMFI, CB of Chile, UDP, and U Chile. Diego Huerta and Cristian Valencia provide excellent research assistance. This study was developed within the scope of the research agenda conducted by the Central Bank of Chile (CBC). The CBC has access to anonymized information from various public and private entities by virtue of collaboration agreements signed with these institutions. The views expressed do not represent the views of the CBC or its board members.

† Central Bank of Chile. E-mail: aaguirre@bcentral.cl, mtapia@bcentral.cl, lvillacorta@bcentral.cl

1 Introduction

Over the last decade, a rich literature has used macro models with heterogeneous agents to quantify the importance of financial frictions at the firm level for aggregate productivity, capital, and income [see Buera et al., 2015]. An important insight provided by these models is that the quantitative effects of financial frictions are driven by the joint distribution of firm wealth and productivity and how this distribution evolves over time. If firm productivity and wealth are not well aligned, financial frictions can distort the relationship between investment and productivity, generating misallocation of capital and other production factors across firms with potentially important macro implications. Another crucial insight from these models is that firms can overcome financial frictions over time by accumulating wealth in response to persistent productivity shocks. This endogenous "self-financing channel" implies that wealth and productivity will align over time and has the potential to mitigate the aggregate adverse effects of financial frictions [e.g. Moll, 2014].[1]

However, a detailed analysis, using micro-data, of the firm-level empirical predictions and mechanisms that underlie these models is currently absent from the literature. This is a relevant issue, as there is scant direct evidence on the individual decisions that lie at the center of the forces driving the self-financing channel and its macroeconomic implications.[2] In particular, micro-level features like the characteristics of the firm productivity process and the individual policy functions for investment and wealth accumulation govern the evolution of the joint distribution of productivity and wealth and the actual relevance of self-financing on an aggregate scale. This paper attempts to bridge up this gap, providing firm-level evidence on the joint dynamics of productivity, investment, and wealth accumulation.

We present a novel methodology to estimate firm-level production functions in the face of financial constraints and use it to characterize the nature of the firm productivity process and its effect on the firm's decisions. To the best of our knowledge, this is the first paper that estimates, using micro-data, the relevant policy functions for investment and wealth accumulation that emerge in macro models with financial frictions. We use these functions to empirically document the response of investment and wealth accumulation to productivity shocks at the micro level, study how these responses vary along the wealth and productivity distribution and explore the strength of the self-financing mechanism in reducing misallocation.

[1]Quantitative models, in Banerjee and Moll [2010], and Midrigan and Xu [2014], suggest that self-financing is strong enough to rapidly undo the impact of financial frictions on misallocation. These findings have lent support to the idea that the macro implications of financial frictions are less significant than thought earlier.

[2]Buera et al. [2021] "macro models have tended to rely on strong structural assumptions, e.g., assumptions on functional forms and distributions of unobservables, and on somewhat stylized calibration strategies, and thus economists often view it as disconnected from micro empirical research."

To do so, we take advantage of a rich data set of manufacturing firms obtained from a census of administrative records of formal firms in Chile from 2006 to 2016. Besides including data on inputs and output at the firm level, the database provides information on the balance sheet of firms, allowing us to characterize wealth and directly analyze the role of financial frictions. The data set has the advantage of including a panel of firms of different sizes and characteristics, mostly private, in the context of an emerging economy. As we can follow individual firms over time, we can directly observe their wealth accumulation and investment decisions and relate them to the evolution of the estimated productivity process. This data set is an ideal laboratory to study the firm-level dynamics that lie at the foundations of the macroeconomic models used in the literature [see discussion in Diggs and Kaboski, 2022].[3]

We provide four key findings *First*, as opposed to the standard linear AR(1) productivity process with constant persistence embedded in most quantitative macro models, we find a highly nonlinear productivity process revealing heterogeneity in the persistence of productivity shocks, which depends on the previous level of productivity and the magnitude and the sign of new shocks. This has important implications for the strength of self-financing, as higher (lower) persistence provides stronger (weaker) incentives for firms to accumulate wealth after a positive productivity shock (see Moll [2014]). We find that persistence is typically increasing in productivity, ranging from 0.95 for high-productivity firms to as low as 0.75 for low-productivity firms. However, persistence can change in the face of extreme events. For example, extremely negative shocks to an ex-ante highly productive firm can reduce the influence of previous productivity shocks on current productivity.

Second, we find heterogeneous patterns in the response of firm investment to productivity shocks. We are the first paper to document a nonlinear relationship between investment and productivity at the firm level. We find larger responses in investment to productivity shocks at higher productivity levels. Such a larger response could be explained by our finding that more productive firms also display higher productivity persistence. Also, we find that the investment reaction to positive productivity shocks is increasing in wealth for all productivity levels, providing support for models with collateral constraints. For instance, the investment elasticity to a productivity shock rises from 0.05 to 0.10 for low-productivity firms, from 0.23 to 0.35 for medium-productivity firms, and from 0.4 to 0.6 for high-productivity firms when we move along the wealth distribution. Moreover, we find that the investment of productive firms reacts to a positive productivity shock even at the lower end of the wealth distribution (the elasticity is close to 0.4 for a high-productivity firm at the lower end of the wealth

[3] "Perhaps the biggest obstacle in researching financial frictions in developing countries is data availability. Ideally, data would consist of information on the firm ability and wealth over several years."(Diggs and Kaboski, 2022)

distribution as opposed to 0.05 for a low-productivity firm with the same low wealth), which could also lend support to the existence of earning-based constraints [as in Lian and Ma, 2020, Drechsel, 2022].

Third, we are the first paper to document that the effect of productivity on wealth accumulation is heterogeneous in wealth and productivity. The elasticity of positive income shocks to savings is high for low-wealth firms and weakens significantly as we move upwards along the wealth distribution, which is in line with the notion of self-financing. For high-productivity firms, such elasticity goes from one (which indicates a full transmission of income to savings) to 0.45 as we move up in the wealth distribution. For low-productivity firms, such elasticity ranges from 0.6 to 0.2.

Fourth, we use our estimated empirical model to compute the speed of convergence of the marginal product of capital (MPK) between firms with the same productivity but different levels of initial wealth. While our results show that convergence in MPK between firms does occur, it is slow, as differences in MPK persist for more than three decades, even for very productive firms. This suggests that the self-financing channel might be less strong than suggested elsewhere in the literature [e.g. Banerjee and Moll, 2010, Midrigan and Xu, 2014].

Our approach Our empirical analysis is guided by the macroeconomic models that study financial frictions and the self-financing channel [e.g. Buera and Shin, 2011, Moll, 2014, Midrigan and Xu, 2014]. But, an important aspect of our tractable econometric framework is that it uncovers the firms' productivity process and its effects on firms' investment and wealth accumulation decisions without relying on a structural estimation. In contrast to fully-specified structural approaches, which require the specification of particular functional forms for preferences, financial frictions, and especially the distribution of productivity, we adopt a non-parametric approach where we recover productivity from the firm production function and estimate nonlinear firm's policy rules that are compatible with a large class of heterogeneous-agent models with financial frictions. This setup provides a great degree of flexibility, allowing us to incorporate different forms of financial frictions, including both collateral and earnings-based constraints, and to study non-linear responses to productivity shocks. In line with the predictions of theoretical models, the marginal effect of productivity is allowed to be heterogeneous across firms and contingent on the level of firm wealth and productivity. Therefore, we can characterize the complete distribution of micro-level investments and wealth accumulation propensities in response to productivity shocks.

Our empirical framework shares the spirit of the empirical consumption-household income framework [e.g. Blundell et al., 2008, Kaplan and Violante, 2010, Arellano et al., 2017], which exploits panel data to estimate how consumption decisions respond to unobserved household income shocks. A crucial econometric difference between these frameworks lies in the estimation of the net income process. In the household framework, income shocks and their effect on consumption are extracted directly from the household income data after removing demographic characteristics that are assumed to be orthogonal to the income shocks. By contrast, to estimate the unobserved firm productivity process and its effect on investment and savings, we need to estimate the production function parameters, where the regressors are endogenous and correlated with unobserved productivity.

The first step of our analysis builds on previous literature that estimates production functions and productivity at the firm level. This literature relies on a proxy variable approach to recover productivity using the firm's input decisions [see Ackerberg et al., 2015, for a review]. For instance, Olley and Pakes [1996] recover productivity by inverting an investment demand function, whereas Levinsohn and Petrin [2003] invert the firm demand function for intermediate inputs. The proxy variable approach uses observed differences in input demands to control for differences in unobserved productivity in a production function regression. We extend this estimation method to allow for financial frictions, as otherwise proxy methods deliver biased estimates of the production function and the productivity process. Intuitively, financial frictions generate differences in input demands for equally productive firms that the proxy variable method misinterprets as differences in unobserved productivity. Instead, our method compares input demands for firms with similar levels of wealth. Additionally, the proxy method is not well-suited to identify and estimate flexible empirical policy functions, as it does not allow for unobservables besides productivity in the policies. This is an empirically restrictive assumption since it rules out the possibility of idiosyncratic shocks and measurement error.

Combining the insights of the self-financing channel with recent developments in nonlinear panel models with latent variables [Hu and Schennach, 2008, Arellano et al., 2017], we propose a sequential identification scheme to identify the production function, the productivity process, and the investment and wealth accumulation policy functions. Due to the self-financing channel, more productive firms have higher incentives to invest and accumulate wealth than less productive firms with the same level of wealth. We use that conditional correlation between investment and wealth accumulation to reveal differences in unobserved productivity across firms and control for it in the production function regression.[4] Once the production function parameters are identified, the productivity process is identified from

[4]From an instrumental variable perspective, both policy functions can be thought of as noisy measures of unobserved productivity. If the production function and both policies are related only through productivity and observed state variables, we show that the wealth accumulation policy provides a valid external instrument for investment, which is used as a proxy variable with noise in a production function regression — that also includes the stock of wealth to control for collateral constraints.

the time-series dependence structure of the firm net income process (the production income net of the compensation to endogenous inputs). Finally, once the productivity process is identified, the policy rules that depend on productivity are identified using non-parametric instrumental variables arguments given the exclusion restrictions provided by our dynamic model. In that sense, we are the first paper to provide the conditions for the identification of the empirical functions underlying the quantitative macro models with financial frictions.

Another advantage of our approach (compared to a full structural estimation) is econometric transparency (see the discussion in Andrews et al. [2020]). First, we formally discuss the identification of the nonlinear policy functions that emerge from heterogeneous agent models with financial frictions. Second, our IV estimator is transparent, as it directly connects our estimates with the relevant moments and variation in the data that "drive" the estimator. Although the empirical model cannot provide direct policy counterfactuals, its estimated parameters may be used directly or indirectly to calibrate structural models that are able to do so. Our production function and productivity estimates can be used to parametrize the firm's production function and the productivity process directly in a structural model, while our empirical policy rules can be used as matching targets for other key parameters related to preferences, adjustment costs to capital, and financial constraints.

We also uncover new empirical results on the estimates of the firm production function and productivity process as we find significant differences once we control for financial frictions. We show that applying standard methods without controlling for financial frictions underestimates the marginal effect of capital (the constrained input) in the production function due to the negative correlation between capital and financial frictions and underestimates the productivity of constrained firms as they show larger investment gaps with respect to their optimal levels. As a result of the underestimation of the capital parameter and productivity, those methods overestimate the labor parameter to fit the production function. Consistent estimation of these objects is crucial as they are key inputs in structural models that quantitatively study the role of financial frictions and the self-financing channel.

Related literature Our paper makes contributions to different streams in the literature. Our initial motivation is the macro-finance literature that studies the aggregate effects of financial frictions. We are closer to the set of papers focusing on collateral constraints and self-financing [e.g. Banerjee and Moll, 2010, Buera and Shin, 2011, Buera et al., 2011, Buera and Shin, 2013b, Caggese and Cuñat, 2013, Moll, 2014, Midrigan and Xu, 2014, Buera et al., 2015], as we guide our empirical specification by the general implications of these models, i.e., self-financing by incumbents can undo financial frictions and allow firms to invest closer to their optimum. As mentioned, our contribution is to provide novel direct evidence and an identification strategy on firms' wealth accumulation and investment decisions, which in these papers are an endogenous outcome of calibrated structural models built under different assumptions. As suggested by Hopenhayn [2014], this may be the source of the disparity of magnitudes reported for the aggregate effects of financial frictions. Our estimates may help to discipline these models and provide further insights into their underlying mechanisms. We provide estimates of key elasticities and unlike these papers, we exploit microeconomic data on both real and financial variables. Ours is the first paper to provide direct, firm-level evidence of the self-financing channel.

This paper also connects to two strands of research in corporate finance. One area of literature, starting with Fazzari et al. [1987], tries to identify financially constrained firms through their investment sensitivity to cash flows beyond profitability, typically captured by Tobin's Q or other observable characteristic. A second related area discusses the determinants of firms' cash holding decisions and associates them to firm characteristics such as growth opportunities and risk management.[5] In our framework, the investment and wealth accumulation policy functions are two key outcomes, and we identify unobservable productivity not only to control for profitability but also to estimate non-linear and interaction effects with our measure of collateral. Furthermore, since we follow the structural macro models, we focus on net wealth instead of cash flows.

The paper also connects with the empirical literature that estimates production functions at the firm level using the proxy variable approach [Olley and Pakes, 1996, Levinsohn and Petrin, 2003, Ackerberg et al., 2015, Doraszelski and Jaumandreu, 2013, 2018, Gandhi et al., 2020, Shenoy, 2020]. Our paper differs from these papers in several aspects. First, our paper is the first paper to study the biases that appear when the proxy method is used to estimate the production function under collateral constraints. Second, we use the insights and mechanisms presented in macro models to propose a novel strategy that is robust to financial frictions. In this sense, our paper is the first paper that uses the self-financing channel to identify the firm productivity process and the firm production function. We allow for more flexible policy rules, including transitory shocks, unlike the proxy variable approach. Finally, a key difference is the identification and estimation of the policy functions.

The rest of the paper is organized as follows. Section 2 presents a model of firm dynamics with collateral constraints to motivate the ingredients of the empirical model and shed light on the biases of the proxy variable approach. Section 3 introduces the empirical model and its assumptions. Section 4 establishes identification, while Section 5 describes the estimation methods. Section 6 describes the data and presents the main empirical results.

[5] See, for example, Opler et al. [1999], and Almeida et al. [2004]

2 A Simple Model with Financial Frictions

We start with a stylized structural model inspired by the macro literature that studies financial frictions and the self-financing channel [see Buera et al., 2015, for a detailed analysis]. The model motivates the ingredients of the empirical policy rules we take to the data, provides the mechanisms and assumptions that identify the empirical model, and illustrates the biases of the proxy variable approach under financial constraints.

Consider a price-taking firm with initial wealth A_{it}, capital K_{it} and productivity Z_{it} that solves the following dynamic problem to maximize the discounted value of distributed profits D_{it} choosing labor L_{it}, investment I_{it} and next period wealth A_{it+1}:

$$
\begin{aligned}
V(A_{it}, K_{it}, Z_{it}) &= \max_{A_{it+1}, I_{it}, L_{it}} D_{it} + \beta E\left[V(A_{it+1}, K_{it+1}, Z_{it+1}) | Z_{it}\right], \\
\text{s.t.} \qquad D_{it} + g(A_{it+1}) &= Y_{it} - W L_{it} - (r+\delta) K_{it} + (1+r) A_{it}, \\
Y_{it} &= Z_{it} K_{it}^{\beta_k} L_{it}^{\beta_l} \\
K_{it+1} &= I_{it} + (1-\delta) K_{it}.
\end{aligned}
$$

where Y_{it} is the value added produced by firm i. Investment, which determines the next period's capital, is decided before observing next period's productivity, while labor is decided contemporaneously with productivity.[6] As preferences are linear, $g(\cdot)$ is assumed to be convex, ruling out corner solutions.[7] The firm discounts future flows at β, capital depreciates at δ, and the firm pays interest rate r for its debt, implicitly defined by $K_{it} - A_{it}$.

The log of productivity z_{it} follows a Markovian process with distribution P_z.

$$
P_z\left(z_{it} \mid z_{it-1}, z_{it-2}, \ldots\right) = P_z\left(z_{it} \mid z_{it-1}\right) \tag{1}
$$

where $E\left[z_{it} \mid z_{it-1}\right] = \varphi\left(z_{it-1}\right)$ is a continuous function of z_{it-1}. The quantitative macro literature typically assumes normality for P_z and linearity for $\varphi\left(z_{it-1}\right)$.

Financial Constraint Following Buera et al. [2015] we consider the following specification

$$
K_{it+1} \leq \kappa(A_{it}, Z_{it}) \tag{2}
$$

[6] This assumption implies that it takes a full period for new capital to be delivered and installed.

[7] Assuming linear preferences is not needed in our empirical framework but simplifies the illustrative analysis in this section. Including the convex function g introduces an incentive to smooth assets over time, ruling out solutions in which firms retain either all or none of their earnings. This specification combines ease of analysis with the general qualitative implications of models that introduce concavity in preferences.

Equation 2 implies that debt is limited by the repayment capacity of the firm, a combination of its productivity and current wealth. This captures the idea that financial friction depends on the profitability of the firm and its financial status. The first term in 2 is an *asset-based collateral constraint*, as net worth determines the part of the balance sheet that is owned by the firm and can be pledged as collateral. The second term in 2 represents *earning-based constraints*, as persistent productivity determines the flow of current and future cash flows, which are the main factors in earning-covenants and earning-based lending (see Lian and Ma [2020]).[8] Equation 2 nests the standard linear constraints [Moll, 2014, Midrigan and Xu, 2014, Buera et al., 2015, Drechsel, 2022], and nonlinear constraints as in Gopinath et al. [2017]. We left the function κ unrestricted and consider policy functions that are compatible with any financial constraint that is a function of wealth and productivity.

Optimality Conditions The FOC with respect to investment can be written as:

$$C_k E(Z_{it+1}|Z_{it})^{\frac{1}{1-\beta_l}} \left(I_{it} + (1-\delta)K_{it}\right)^{\frac{\beta_k}{1-\beta_l}-1} = \beta(r+\delta) + \mu(A_{it}, Z_{it}), \tag{3}$$

where C_k is a constant. The last term on the right-hand side is the wedge caused by financial frictions, and is the multiplier of the collateral constraint (2), which is decreasing in A_{it}. Given the wedge, MPKs will not equalize across firms, so the equilibrium allocation of capital, a function of the current distribution of A_{it} and Z_{it}, is not efficient. Equation (3) generates an investment policy function (in logs) that depends nonlinearly on wealth and productivity.

$$i_{it} = h(z_{it}, k_{it}, a_{it}) \tag{4}$$

Finally, in an environment with collateral constraints, the firm must decide on wealth accumulation, which is crucial to finance future investments. The FOC is given by:

$$g'(A_{t+1}) = \beta\left(1 + r + E_t\left[\kappa_A \mu(A_{t+1}, Z_{t+1})\right]\right) \tag{5}$$

Hence, even if the constraint does not bind today, wealth accumulation is desirable if the constraint can bind in the future. When the constraint binds, an additional dollar of retained earnings allows the firm to increase investment in κ_A dollars. The marginal benefit of wealth is then the expected MPK net of borrowing costs, the value of the multiplier. As productivity is persistent, high-productivity firms have higher expected future MPKs, gen-

[8] Lian and Ma [2020], and Ivashina et al. [2022] provide substantial evidence that both types of financial constraints are prevalent in developed and developing countries, whereas Aguirre [2017], Brooks and Dovis [2020] and Drechsel [2022] show that both constraints are quantitatively important.

erating a positive correlation between productivity and wealth accumulation. That is, for a given level of wealth, more productive firms will accumulate more wealth today to expand future investment. As emphasized by Moll [2014] higher productivity persistence increases the incentives to accumulate wealth. Similarly to investment, we can define this general relationship as the wealth accumulation policy function

$$a_{it+1} = g(z_{it}, k_{it}, a_{it}) \tag{6}$$

The extent of financial frictions and the strength of self-financing are reflected in the responses of investment and wealth accumulation to persistent productivity shocks and how these responses vary with available collateral and current productivity.

This simple setup illustrates the **goal of this paper**: to flexibly characterize, using microdata, the firm productivity process in (1) without parametrizing its distribution, and document its impact on firm decisions estimating the firm policy functions in (4) and (6) without relying on approximations and distributional assumptions.

Biases in proxy variable estimators in the presence of financial frictions. In this paper, we follow the industrial organization literature and recover firm productivity without using distributional assumptions by estimating the firm production function. However, the model above provides insights into the biases of the estimates that use standard empirical methods that do not account for financial frictions and how these biases can distort the interpretation of the production function and the productivity process. We illustrate our argument in the context of the influential paper by Olley and Pakes [1996], henceforth OP. However, the same logic applies to Levinsohn and Petrin [2003], as long as financial frictions affect the demand for materials as in Mendoza and Yue [2012]. The analysis also applies to Ackerberg et al. [2015], and Wooldridge [2009]. These biases are problematic, as having consistent estimates of the production function and the productivity process is key for structural macro models that quantitatively study financial frictions.

Here we discuss the intuition behind the biases, and we provide a detailed explanation using the macro model described at the beginning of this section in Online Appendix 1. Consider the log of the value-added production function described above:

$$y_{it} = \beta_l l_{it} + \beta_k k_{it} + z_{it} + \varepsilon_{it}, \tag{7}$$

where ε_{it} is measurement error in value added.[9] The challenge in the estimation of β_l and

[9]We focus on a model with perfect competition where output prices are homogeneous across firms as in Olley and Pakes [1996], Levinsohn and Petrin [2003], Ackerberg et al. [2015], and Gandhi et al. [2020]. For

β_k is that z_{it} is an unobservable variable that is potentially correlated with the regressors k_{it} and l_{it}, creating an endogeneity problem in the OLS regression of y_{it} on k_{it} and l_{it}. The OP approach uses the investment policy function as an auxiliary equation to obtain information on z_{it}. In the absence of collateral constraints, by controlling for investment in the production function, OP eliminates the endogeneity problem and provides consistent estimates of β_l and β_k. Intuitively, OP interprets differences in firm investment in the data as differences in unobserved firm productivity.

However, in the model with financial frictions described above, the investment function does not only depend on productivity and initial capital, but also on net worth and its effect on credit access. Therefore, under borrowing constraints, equally productive firms with different wealth might have different investment, capital, and output. In consequence, if the implied heterogeneity in output is driven by heterogeneity in capital due to differences in credit access, the OP approach will wrongly assign such dispersion to productivity, as it will misinterpret differences in investment as solely coming from productivity gaps. As a result, the OP productivity proxy captures an important part of the effect of capital on output, underestimating its true marginal impact, and downward biasing the estimated coefficient for capital. Conversely, as long as financial frictions are relatively less severe for labor, the labor coefficient is upwardly biased. OP interprets a financially constrained firm with low investment as a low-productivity firm that hires "too many" workers and produces "too much" output relative to its proxy-OP productivity. Hence, it will assign a large role to labor in production, overestimating the labor elasticity. Furthermore, these biases in estimated factor elasticities can have significant effects on the measure of returns to scale. Additionally, OP will underestimate the productivity of financially constrained firms, as they have larger investment gaps with respect to their optimal levels. Therefore, the estimated productivity distribution across firms will also be biased. We empirically illustrate these biases and how our methodology corrects them, both with actual as well as from simulated data using a model with financial frictions.

3 General Empirical Framework

This section discusses the empirical model and its stochastic assumptions. The model consists of the production function in (7) and the empirical counterparts of the productivity process in (1) and the policy functions in (4) and (6).[10]

production function estimation with monopolist competition, see De Loecker [2011].

[10] We consider a Cobb-Douglas production function since it is the specification used in the structural models that study the self-financing channel. Our approach can accommodate more flexible production functions as long as productivity is Hicks-neutral.

As we want to recover the productivity distribution from the data, we consider a very flexible specification for the productivity process using a quantile model:

$$z_{it} = Q_z\left(z_{t-1}, \eta_{it}\right) \tag{8}$$

with innovation $\left(\eta_{it} \mid z_{t-1}\right)$ uniform distributed $U\left(0,1\right)$ and persistence:

$$\rho\left(z_{t-1}, \tau\right) = \frac{\partial z_{it}}{\partial z_{it-1}} = \frac{\partial Q_z\left(z_{t-1}, \tau\right)}{\partial z_{it-1}} \tag{9}$$

with τ being the quantile of the shock. Equation 8 is a direct non-parametric model for $P_z\left(z_{it} \mid z_{it-1}\right)$, that leaves the dependence structure of z unrestricted beyond the Markovian assumption as opposed to the normal distribution usually assumed in macro models. Also, compared to the linear AR(1) traditionally used in the literature, it allows for nonlinear persistence with two main properties. First, for a given shock η_{it} the relationship between z_{it} and z_{it-1} depends on z_{it-1}. Therefore, persistence after a given innovation can vary across firms with different productivity levels z_{it-1}. Second, for a given z_{it-1} the relationship between z_{it} and z_{it-1} depends on η_{it}. This implies that unusually large shocks can change the relationship between current and past productivity, canceling the cumulative effects of past shocks. This modeling approach was introduced by Arellano et al. [2017] to model persistent income shocks to households. We are the first paper that uses it to estimate productivity and production functions at the firm level. As it is standard in the literature, shocks η_{it+1} and ε_{it} are not part of the firm's information set when making decisions at t. The assumptions on the stochastic processes underlying both shocks are explained below.

Following the model in Section 2, capital k_{it} is a dynamic but predetermined input, decided in $t-1$ when the firm chose i_{it-1}, while labor l_{it} is a flexible input. The specification of the empirical policy rules follows the model discussed earlier, but each policy function is augmented by a stochastic shock:

$$i_{it} = h_t\left(z_{it}, k_{it}, a_{it}, v_{it}\right), \tag{10}$$

$$a_{it+1} = g_{t+1}\left(z_{it}, a_{it}, k_{it}, w_{it+1}\right). \tag{11}$$

where h_t and g_{t+1} are the nonlinear reduced-form policy rules of investment and wealth that can be derived in a firm-dynamics model with financial frictions as the one discussed earlier. The terms v_{it} and w_{it+1} capture other unobserved factors besides z_{it} that affect the evolution of investment and wealth. For example, in the context of our earlier stylized model, shocks to collateral constraints could affect the investment policy function. Firms may face temporary idiosyncratic shocks that affect the relationship between debt, productivity, and collateral (i.e., $\kappa(Z_{it}, A_{it}, v_{it})$). In the case of the wealth accumulation policy function, stochastic shocks can come from unexpected fluctuations in the valuation of firms' assets. If these occur in the interim between the distribution of dividends (when equation 5 is solved) and when the firm uses wealth as collateral to borrow (when equation 4 is solved), they will appear as unplanned changes in the valuation of collateral.[11] More generally, the inclusion of shocks can bridge the transition from the insights of the stylized model to an empirical model that deals with actual microdata and issues such as measurement errors that emerge from the use of different datasets.

We also assume that h_t and g_{t+1} are monotonic in v_{it} and w_{it+1}, respectively. The specification in (10) nests a number of nonlinear empirical investment functions studied in the literature [e.g. Olley and Pakes, 1996, Cooper and Haltiwanger, 2006, Gala et al., 2020]. The two major innovations of our framework are (i) the inclusion of wealth a as an additional state variable in (10), to control for the existence of collateral constraints and (ii) the explicit modeling of wealth dynamics in (11) and its relationship with productivity (the self-financing channel). An additional important difference to Gala et al. [2020] and Cooper and Haltiwanger [2006] is that we explicitly include z_{it} as a state variable, whereas those papers replace z_{it} for value-added, which is an endogenous variable.[12]

Investment and Wealth Accumulation Propensities The nonlinear functions h_t and g_{t+1} allow for heterogeneous effects of productivity shocks on investment and wealth accumulation, depending on the collateral and productivity of the firm. Our objects of interest are the following average derivative effects with respect to z_{it}:

$$\Phi_{it}^h = \Phi^h\left(a_{it}, k_{it}, z_{it}\right) = E_{v_{it}}\left[\frac{\partial h_t\left(z_{it}, k_{it}, a_{it}, v_{it}\right)}{\partial z}\right] \tag{12}$$

$$\Phi_{it+1}^g = \Phi^g\left(a_{it}, k_{it}, z_{it}\right) = E_{w_{it+1}}\left[\frac{\partial g_{t+1}\left(z_{it}, k_{it}, a_{it}, w_{it+1}\right)}{\partial z}\right] \tag{13}$$

where the expectations are taken with respect to the idiosyncratic shocks in the policies. Φ_{it}^h and Φ_{it+1}^g measure the average propensities of investment and wealth accumulation in

[11]The inclusion of v_{it} in the investment function represents a departure from the unobservable scalar assumption required by the proxy variable. It is important to recall that under the proxy variable approach, the investment function is of no interest by itself, as it is only an auxiliary equation to recover the production function.

[12]As Gala et al. [2020] argue in footnote 10, including z_{it} instead of y_{it} requires the estimation of the production function, which adds a number of econometric problems, most significantly, endogeneity. One of the contributions of our paper is to address this issue and consistently estimate the production function and the correct investment equation as a function of unobserved productivity.

response to productivity shocks. These are key objects to understand the nature of financial frictions, and the way in which firms adjust to them.

Estimating these propensities can shed light on the specific nature of the credit constraints. For example, if firms face collateral constraints, Φ_{it}^{h} should be increasing in a_{it} for a given z_{it}, as the investment of wealthier firms can respond more to productivity shocks. Similarly, the existence of earning-based [Drechsel, 2022] or forward-looking constraint [Buera et al., 2015] would imply that Φ_{it}^{h} is increasing in z_{it} for a given level of wealth, as more productive firms can leverage up more in function of their future flows. Of course, both types of constraints can be active, reflecting the heterogeneous nature of contracts and credit relationships faced by firms. As mentioned earlier, these propensities also provide evidence of the response of firms to financial frictions. For instance, if the self-financing channel exists as in Moll [2014], Φ_{it+1}^{g} is always positive and (weakly) decreasing in a_{it} for a given productivity and increasing in z_{it} for a given value of current wealth.

Finally, as it is standard in the macro literature, we model the labor input as a non-dynamic input in the sense that current choices are not affected by past values:

$$l_{it} = n_t \left(z_{it}, a_{it}, k_{it}, w_{l,it} \right), \tag{14}$$

where equation (14) is the empirical labor decision. An extension from the stylized model in Section 2 is that our empirical specification allows for potential effects of financial frictions over hiring decisions, as represented by the inclusion of a_{it} in the policy function. Once again, the term $w_{l,it}$ represents an i.i.d. shock that is independent of the state variables a_{it}, k_{it}, and z_{it}. This shock can capture exogenous transitory shocks to wages in the model in Section 2 or optimization errors as discussed in Ackerberg et al. [2015].

To complete the model description, we formally make the following assumptions, using the notation $x_i^t = (x_{i1}, \ldots, x_{it})$ for any variable x_{it}.

Assumption 1. *(Conditional Independence). For all* $t \geq 1$*:*

(i) ***Output Shock:*** ε_{it+s} *for all* $s \geq 0$ *is independent over time and independent of* $a_i^{t-1}, z_i^{t-1}, i_i^{t-1}, k_i^{t-1}, l_i^t$, y_i^{t-1} *and* η_{it+s}. *Also* ε_{i1} *is independent of* z_{i1}, a_{i1} *and* k_{i1}, *and* $E\left[\varepsilon_{it}\right] = 0$.

(ii) ***Productivity Shock:*** η_{it+s} *for all* $s \geq 0$ *is independent over time and independent of* $a_i^{t-1}, z_i^{t-1}, i_i^{t-1}, k_i^{t-1}, l_i^{t-1}$, *and* y_i^{t-1}.

(iii) ***Policy Functions Shocks:*** v_{it} *and* w_{it+1} *are mutually independent, independent over time, and also independent of* $z_{i1}, a_{i1}, k_{i1} \left(\varepsilon_{is}, \eta_{is}\right)$ *for all* s *and of* v_{is} *and* w_{is+1} *for all* $s \neq t$.

Assumption 2. *(First Order Markovian). For all* $t \geq 1$*:*

(i) a_i^{t+1} is independent of $\left(a_i^{t-1}, k_i^{t-1}, z_i^{t-1}\right)$ conditional on (a_{it}, k_{it}, z_{it})
(ii) i_i^t is independent of $\left(a_i^{t-1}, k_i^{t-1}, z_i^{t-1}\right)$ conditional on (a_{it}, k_{it}, z_{it})

Parts (i) and (ii) of Assumption 1 state that current and future productivity and production shocks, which are independent of past productivity and production shocks are also independent of the current and past wealth and capital stocks, investment, and labor decisions. The initial wealth stock a_{i1}, initial capital stock k_{i1}, and initial productivity z_{i1} are arbitrarily dependent. Allowing for a correlation between a_{i1}, k_{i1}, and z_{i1} is important, as wealth and capital accumulation upon entry in the sample may be correlated with past persistent productivity shocks. Part (iii) requires investment and wealth shock to be mutually independent, independent over time, and independent of production components. Assumption 1 implies that ε_{it}, v_{it} and w_{it+1} are independent of the state variables $(k_{it}, a_{it,}z_{it})$ and mutually independent conditional on $(l_{it}, k_{it}, a_{it,}z_{it})$. Hence, Assumption 1 provides the exclusion restrictions necessary for identification, while Assumption 2 is a first-order Markov condition on wealth and capital dynamics. Assumption 2 is a standard assumption both in macro models as well as in the empirical literature that estimates production functions.

4 Identification

Given the goal of this paper, it is crucial to show that the nonlinear model we aim to estimate can actually be identified from the micro-data we have at hand. Recently, Hu and Schennach [2008] and Arellano et al. [2017] have established conditions under which nonlinear dynamic models with latent variables are non-parametrically identified under conditional independence restrictions. We build on these papers and use the insights of the self-financing channel to provide non-parametric identification of the empirical model introduced in Section 3. In particular, the goal of this section is to show that β_k, β_l, $Q_z\left(z_{t-1}, \eta_{it}\right)$, h_t, g_{t+1} are identified from data on $(y_{it},\ k_{it},\ l_{it},\ i_{it},\ a_{it},\ a_{it+1})$ given that $(z_{it},\ w_{it+1},\ v_{it},\ \varepsilon_{it})$ are not observed by the econometrician and z_{it} is correlated with $(l_{it},\ a_{it},\ k_{it})$. To establish the identification of the non-parametric model, we use the following high-level conditions that we connect with the insights of the macro model discussed in Section 2.

Let $X_{it} = (a_{it,}k_{it}, l_{it})$ be the covariates of the model and let $f\left(a \mid b\right)$ be a generic notation for the conditional density $f_{A|B}\left(a \mid b\right)$.

Condition 1. *Almost surely in covariate values X_t: (i) the joint density $f\left(y_t, i_t, a_{t+1}, z_t \mid X_t\right)$ is bounded, as well as all its joint and marginal densities; (ii) the characteristic function of ε_{it} has no zeros on the real line; (iii) for all $z_{1t} \neq z_{2t}$, $Pr\left[f\left(i_{it} \mid z_{1t}, X_t\right) \neq f\left(i_{it} \mid z_{2t}, X_t\right)\right] > 0$; (iii) $f\left(a_{t+1} \mid z_t, X_t\right)$ is complete in z_{it}. (iv) for $\tilde{y}_{it} = y_{it} - \beta_l l_{it} - \beta_k k_{it}$, $f\left(\tilde{y}_{it} \mid \tilde{y}_{it-1}\right)$,*

$f\left(z_{it} \mid \tilde{y}_{it-1}\right)$, $f\left(z_{it} \mid \tilde{y}_i^T\right)$ *are complete and the distribution of* $f\left(z_{it} \mid a_i^t, k_i^t, \tilde{y}_i^T\right)$ *is complete in* $\left(a_i^{t-1}, k_i^{t-1}, \tilde{y}_i^T\right)$.

Condition 1-(i) requires bounded densities. Condition 1-(ii) is a technical assumption previously used in the literature.[13] The normal distribution and many other standard distributions satisfy this condition. Condition 1-(iii) requires that $f\left(i_{it} \mid z_{it}, X_{it}\right)$ be non-identical at different values of z_{it}. This condition is weaker than the assumption in Olley and Pakes [1996] and Ackerberg et al. [2015], where the realization of investment has to be monotonic in z_{it}. Here we require that two firms with the same level of current wealth and capital but different productivity levels have different investment probabilities. Accordingly, the macro model with earning-based constraints sketched in Section 2 fulfills this condition. Lian and Ma [2020] and Ivashina et al. [2022] provide substantial evidence that earning-based lending is prevalent in developed and developing countries. Also, models with only asset-based constraints fulfill this condition as long as the firms can borrow as much as they want, paying a premium in the interest rate that depends on the collateral as in Cavalcanti et al. [2021].

Condition 1-(iv) is a completeness condition commonly assumed in the literature on non-parametric instrumental variables [Newey and Powell, 2003].[14] Intuitively, we need enough variation in the density $f\left(a_{it+1} \mid z_{it}, a_{it}, k_{it}\right)$ for different values of z_{it}. This requires a statistical dependence between wealth accumulation a_{it+1} and productivity z_{it} conditioned on the observed state variables. This requirement is met by the self-financing channel in the model described in Section 2, where conditional on the same level of current wealth, highly productive firms should accumulate more wealth to relax the friction in the future than less productive firms. In instrumental variable terminology, this is a relevance condition that ensures that a_{it+1} is a valid instrument for z_{it}. For example, suppose $\left(a_{it+}, z_{it}, a_{it}, k_t\right)$ follows a multivariate normal distribution. Then, the completeness condition will require that $E\left[a_{it+1} z_{it} \mid a_{it}, k_{it}\right] \neq 0$, which is ensured by the self-financing channel. Condition 1-(v) requires that z_{it} and z_{it-1} are statistically dependent, which is ensured by the Markovian assumption.

These conditions lead to the following theorem, which sequentially combines the results in Hu and Schennach [2008] and Arellano et al. [2017].

Theorem 1. *If Assumption 1, Assumption 2 and condition 1 (i)-(v) hold, then* β_k, β_l, $Q_z\left(z_{t-1}, \eta_{it}\right)$, h_t, g_{t+1} *are identified from data on* y_{it}, k_{it}, l_{it}, i_{it}, a_{it} *for* $T \geq 4$.

[13]This condition is used for the i.i.d shock of the household income in Arellano et al. [2017] and for the i.i.d shock in the firm production function in Hu et al. [2020].

[14]The distribution of $\tilde{y}_{it} \mid \tilde{y}_{it-1}$ is complete if $E\left[\phi\left(\tilde{y}_{it}\right) \mid \tilde{y}_{it-1}\right] = 0$ implies that $\phi\left(\tilde{y}_{it}\right) = 0$ for all ϕ in some space.

The sketch of identification is sequential. First, we identify the production function parameters β_k and β_l. Once we have identified the production function, we define the firm net income process $\tilde{y}_{it} = y_{it} - \beta_k k_{it} - \beta_l l_{it}$ and use its autocorrelation structure to establish the identification of the productivity process. Finally, we show the identification of the policy functions h_t and g_t that depend on the latent productivity that we have identified from the firm net income. Below we discuss the sketch of the sequential identification and leave the details for Online Appendix 2.

Production Function From *Assumption* 1, ε_{it}, v_{it}, and w_{it+1} are independent conditional on $(l_{it}, k_{it}, a_{it,} z_{it})$, which can be interpreted as the exclusion restrictions in a nonlinear IV setting. Using this conditional independence assumption, we can write the conditional distribution of the observed variables $f(y_t, i_t \mid a_{t+1}, X_t)$, which is a data object, in terms of some elements of the model that we aim to identify:

$$f(y_t, i_t \mid a_{t+1}, X_t) = \int f(y_t \mid z_t, k_t, l_t)\, f(i_t \mid z_t, X_t)\, f(z_t \mid a_{t+1}, X_t)\, dz_t \tag{15}$$

We notice that equation (15) can be framed into the setup studied in Hu and Schennach [2008]. Given *condition* 1(i)-(iv), Theorem 1 of Hu and Schennach [2008] can be applied to our setting to show that the distribution of the production function $f(y_t \mid z_t, k_t, l_t)$ is identified from the data, which leads to the identification of the production function parameters [see Hu et al., 2020]). A novelty of our approach is that our model with financial frictions provides a second policy rule (the self-financing channel) that connects the latent productivity with an observed variable a_{it+1} that is not directly linked to the production function regression (i.e a_{it+1} is not an input in the production function regression), so we can use it as an instrument. We formally discuss the identification of the production function parameters when the policy functions are nonlinear in Online Appendix 2 (see Proposition 1), and to build intuition, we discuss below the case with linear policies.

Intuition. To build intuition for identification, let's consider the case where the policy functions are normally distributed: $i_{it} = h_z z_{it} + h_a a_{it} + v_{it}$ and $a_{it+1} = g_z z_{it} + g_a a_{it} + w_{it+1}$.[15] Both policies give us information on z_{it}. Similar to the proxy approach, we can invert the investment function: $z_{it} = \pi_1 i_{it} + \pi_2 a_{it} + \pi_4 v_{it}$ where $\pi_1 = 1/h_z$, $\pi_2 = -h_a/h_z$ and $\pi_4 = -1/h_z$ and replaced into the production function:

$$y_{it} = \beta_l l_{it} + \beta_k k_{it} + \pi_1 i_{it} + \pi_2 a_{it} + \tilde{\varepsilon}_{it} \tag{16}$$

where $\tilde{\varepsilon}_{it} = \varepsilon_{it} + \pi_4 v_{it}$. In the absence of investment shocks, a simple OLS regression between

[15]For simplicity of exposition, we consider the case where $h_k = g_k = 0$.

y_{it} on l_{it}, k_{it}, i_{it} and a_{it} identifies β_l and β_k, as in the proxy variable approach. The difference with the proxy variable is that the regression controls for a_{it}. Hence, rather than looking for unconditional differences in investment across firms to control for differences in productivity, we are considering differences in investment across firms with the same collateral constraints. In the more general case with investment shocks (i.e $v_{it} \neq 0$), z can not be expressed only as a function of observables and parameters. Therefore, even after controlling for current wealth, one cannot disentangle variation in investment coming from z_{it} from variation in other shocks.[16]

The self-financing channel is key for identification. According to the model in Section 2, for a given level of state variables a_{it} and k_{it}, more productive firms should increase investment more and simultaneously accumulate more wealth to reduce the constraint in the future. Therefore, the covariance between i_{it} and a_{it+1}, conditioned on current wealth a_{it}, allows us to isolate the variation in i_{it} due to variation in z_{it} from the variation in i_{it} due to variation in v_{it}. Hence, a_{it+1} can be used as an instrument for investment in equation (16,) given the conditional independence assumption - wealth does not have a direct effect in the production function- and the relevance condition (completeness) implied by the self-financing channel $g_z \neq 0$. A regression between $E\left[y_{it} \mid a_{it+1}, l_{it}, k_{it}, a_{it}\right]$, and $\left[l_{it}, k_{it}, E\left[i_{it} \mid a_{it+1}, k_{it}, l_{it}, a_{it}\right], a_{it}\right]$ identifies $\{\beta_l, \beta_k\}$.

Productivity Process Once we have identified β_k, β_l, and given that the productivity is Hicks-neutral, we can write the firm net-income process $\tilde{y}_{it} = y_{it} - \beta_k k_{it} - \beta_l l_{it}$ as an additive model with two independent latent variables (given *Assumption* 1).[17]

$$\tilde{y}_{it} = z_{it} + \varepsilon_{it} \tag{17}$$

Given that z_{it} is Markovian and ε_{it} is i.i.d over time, equation (17) has a similar structure to the household income process model with non-linear Markovian persistent shocks studied in Arellano et al. [2017]. To identify the productivity process, we rely on the fact that the net-income process in (17) has a Hidden-Markov structure (by *Assumption* 1) where $\{\tilde{y}_{it-2}, \tilde{y}_{it-1}, \tilde{y}_{it}\}$ are independent given z_{it-1}. The additivity of the net-income process and *condition* 1-(v) allow us to identify the joint distribution of $(\varepsilon_{i2,\cdots,}\varepsilon_{iT-1})$ and the joint distribution of $(z_{i2,\cdots,}z_{iT-1})$ from the autocorrelation structure of $(\tilde{y}_{i1,\cdots,}\tilde{y}_{iT})$ for $T \geq 3$ and identify $Q_z\left(z_{t-1}, \eta_{it}\right)$ for $T \geq 4$. For example, in the linear case $z_{it} = \rho_z z_{it-1} + \eta_{it}$,

[16] This violates the unobservable scalar assumption required by the proxy approach and, therefore, the production function model can not be consistently estimated using OLS since $E\left(i_{it}\tilde{\varepsilon}_{it}\right) \neq 0$

[17] For identification and estimation of production functions with non-neutral productivity see Doraszelski and Jaumandreu [2018] and Villacorta [2018].

we can use equation (17) to express the model in terms of the observed net income process $\tilde{y}_{it} = \rho_z \tilde{y}_{it-1} + \eta_{it} + \varepsilon_{it} - \rho_z \varepsilon_{it-1}$ and use three waves of the net-income process $\{\tilde{y}_{it-2}, \tilde{y}_{it-1}, \tilde{y}_{it}\}$ to identify ρ_z from an IV regression using $\tilde{y}_{it-2}$ as an instrument for $\tilde{y}_{it-1}$ (given that $\tilde{y}_{it-1}$ and ε_{it-1} are correlated). Then, the variance of the productivity shock and the variance of the measurement error in income are identified from $E\left(\tilde{y}_{it}\tilde{y}_{it-1}\right) = \rho_z E\left(\tilde{y}_{it-1}\tilde{y}_{it-1}\right) - \rho_z \sigma_\varepsilon^2$ and $E\left(\tilde{y}_{it}\tilde{y}_{it}\right) = \rho_z^2 E\left(\tilde{y}_{it-1}\tilde{y}_{it-1}\right) + \sigma_\eta^2 + \left(1-\rho_z^2\right)\sigma_\varepsilon^2$. In proposition 2 (in appendix) we extend this argument and use the Hidden-Markovian structure of $\{\tilde{y}_{it-2}, \tilde{y}_{it-1}, \tilde{y}_{it}\}$ to establish identification of a non-parametric productivity process.

Policy Functions Once $\left(z_{i1} \mid \tilde{y}_i^T\right)$ is identified, we use *Assumption* 1 and *Assumption* 2 to construct the following IV moment restriction, which allows us to relate the conditional distribution of observable variables $f\left(a_1, k_1 \mid \tilde{y}^T\right)$, $f\left(a_{t+1} \mid a^t, k^t, \tilde{y}^T\right)$, and $f\left(i_t \mid a^t, k^t, \tilde{y}^T\right)$ which are data objects, to the distribution of the policy rules we want to identify (see proposition 3).

$$f\left(a_1, k_1 \mid \tilde{y}^T\right) = E\left[f\left(a_1, k_1 \mid z_1\right) \mid \tilde{y}_i^T = \tilde{y}^T\right] \tag{18}$$

$$f\left(a_{t+1} \mid a^t, k^t, \tilde{y}^T\right) = E\left[f\left(a_{t+1} \mid z_t, a_t, k_t\right) \mid a_i^t = a^t, k_i^t = k^t, \tilde{y}_i^T = \tilde{y}^T\right] \tag{19}$$

$$f\left(k_{t+1} \mid a^t, k^t, \tilde{y}^T\right) = E\left[f\left(k_{t+1} \mid z_t, a_t, k_t\right) \mid a_i^t = a^t, k_i^t = k^t, \tilde{y}_i^T = \tilde{y}^T\right] \tag{20}$$

where the expectation in (18) is taken with respect to the density of z_{i1} given $\tilde{y}_i^T$ for fixed values of a_1 and k_1 and the expectation in (19) and (20) are taken with respect to the density of z_{it} given $\tilde{y}_i^T$, k_i^t, and a_i^t for a fixed value of a_{t+1} and k_{t+1}, respectively. Equation (18) is analogous to a nonlinear IV problem where z_{i1} is the endogenous regressor and $\tilde{y}_i^T$ is the vector of instruments. The difference with a standard nonlinear IV is that the "endogenous regressor" in the moment condition in (18) is a latent variable. However, this is not a problem since we have identified $\left(z_{i1} \mid \tilde{y}_i^T\right)$ using the production function. Provided that the distribution of $\left(z_{i1} \mid \tilde{y}_i^T\right)$ is complete (*condition* 1(v)), the unknown density $f\left(a_1, k_1 \mid z_1\right)$ is identified from (18). Similarly, equations (19) and (20) can be interpreted as nonlinear IV restrictions where a_{it} and k_{it} are the controls (they are arguments in the wealth function and investment functions), and the vector $\tilde{y}_i^T$ contains the excluded instruments. Given *condition* 1(v) and *Assumption* (2), the distributions $f\left(a_{t+1} \mid z_t, a_t, k_t\right)$ and $f\left(k_{t+1} \mid z_t, a_t, k_t\right)$ for $t > 2$ are identified recursively from equations (19) and (20). The identification of $f\left(a_{t+1} \mid z_t, a_t, k_t\right)$ and $f\left(k_{t+1} \mid z_t, a_t, k_t\right)$ allows us to recover the policy functions $g_{t+1}()$ and $h_t()$. Here, we are using the autocorrelation structure of $\tilde{y}_i^T$, conditioned on current values of a_{it} and k_{it}, to construct instruments (lagged and lead values of the firm's net income process) to identify the policies. For example, in the linear case $a_{it+1} = g_z z_{it} + g_a a_{it} + g_k k_{it} + w_{it+1}$, we can use equation (17) to express the model in terms of the observed net income process $a_{it+1} = g_z \tilde{y}_{it} + g_a a_{it} + g_k k_{it} + w_{it+1} - g_z \varepsilon_{it}$, and identify the linear policy functions from an IV regression using $\tilde{y}_{it-1}$ as an instrument for $\tilde{y}_{it}$ (given that $\tilde{y}_{it}$ and ε_{it} are correlated) and controlling for a_{it} and k_{it}.

5 Empirical Strategy

In this section, we discuss two approaches to estimate different versions of the model. First, we consider a parsimonious model where at least one of the policies is a quasi-linear function in productivity. For this model, we propose a novel procedure that consists of an IV regression within the proxy variable framework, following the identification strategy presented in section 4. Second, we consider a more flexible model that allows for unrestricted nonlinear effects of productivity, and we consider a flexible estimation method well suited for nonlinear models with latent variables.

5.1 Parsimonious policy functions

Proxy-IV The identification of β_l and β_k using the IV-proxy method strategy requires that at least one of the two policy functions is polynomial of degree one in z_{it} and separable in z_{it} and the shock. The other policy function and the distribution of the shocks are left unrestricted. Consider the following investment function:

$$i_{it} = h\left(z_{it}, k_{it}, a_{it}, v_{it}\right) = h_1\left(k_{it}, a_{it}\right) + h_2\left(k_{it}, a_{it}\right) z_{it} + v_{it}, \tag{21}$$

It is important to notice that model (21) is flexible enough to capture heterogeneous effects of productivity on investment depending on the level of collateral. Meanwhile, the wealth accumulation policy function is left unrestricted. As in the proxy variable approach, we can invert equation (21) in terms of productivity:

$$z_{it} = \pi_1\left(k_{it}, a_{it}\right) + \pi_2\left(k_{it}, a_{it}\right) i_{it} + \omega_{it} \tag{22}$$

where $\pi_1\left(k_{it}, a_{it}\right) = -h_1\left(k_{it}, a_{it}\right)/h_2\left(k_{it}, a_{it}\right)$, $\pi_2\left(k_{it}, a_{it}\right) = 1/h_2\left(k_{it}, a_{it}\right)$ and $\omega_{it} = -v_{it}/h_2\left(k_{it}, a_{it}\right)$. Replacing (22) in the production function:

$$y_{it} = \beta_l l_{it} + \phi\left(k_{it}, a_{it}\right) + \pi_2\left(k_{it}, a_{it}\right) i_{it} + \omega_{it+1} + \varepsilon_{it}, \tag{23}$$

where $\phi(k_{it}, a_{it}) = \beta_k k_{it} + \pi_1\left(k_{it}, a_{it}\right)$. As we emphasized in section 4, an OLS regression of (23) does not provide a consistent estimator of β_l since $E(\omega_{it} \mid i_{it}) \neq 0$. However, given *Assumption 1*, a_{it+1} can be used as an instrument for i_{it} in equation (23). Therefore, we propose the following two-stage procedure:

First Stage: Estimate (23) with an IV estimator using $\pi_2(k_{it}, a_{it})\, a_{it+1}$ as the instrument for $\pi_2(k_{it}, a_{it})\, i_{it}$. The IV regression delivers a consistent estimator of β_l, $\phi(k_{it}, a_{it})$ and $\pi_2(k_{it}, a_{it})\, a_{it+1}$. For instance, in the linear case where $g_2(k_{it}, a_{it}) = 1$, a_{it+1} will be the instrument for i_t.

Second Stage: Combining equation (22) with the Markovian model for a linear productivity process $z_{it} = \rho_z z_{it-1} + \eta_{it}$:

$$z_{it} = \rho_z \pi_1(k_{it-1}, a_{it-1}) + \rho_z \pi_2(k_{it-1}, a_{it-1})\, i_{it-1} + \rho_z \omega_{it-1} + \eta_{it}, \tag{24}$$

Replacing equation (24) into the production function:

$$y_{it} - \beta_l l_{it} = \beta_k k_{it} + \rho_z \pi_1(k_{it-1}, a_{it-1}) + \rho_z \pi_2(k_{it-1}, a_{it-1})\, i_{it-1} + \rho_z \omega_{it-1} + \eta_{it} + \varepsilon_{it}, \tag{25}$$

using *assumption 1* we can define the following moment condition from equation (25)

$$E(\omega_{it-1} + \eta_{it} + \varepsilon_{it} \mid k_{it}, k_{it-1}, a_{it-1}, a_t) = 0, \tag{26}$$

The moment condition in (26) allows us to identify β_k. We refer to this novel estimator as *Proxy-IV*. Once β_l and β_k are estimated we can estimate the productivity process and the policy functions following the IV strategy discussed in section 4 for the simple cases where productivity and the policies are linear functions.

5.2 Flexible policy functions

To estimate a more flexible model that allows for nonlinear persistence in productvity and nonlinear interactions between z_{it} and observed state variables in the policies, we bring the following nonlinear specifications to the data:

(i) For productivity, we implement the following quantile specification:

$$z_{it} = Q_z(z_{t-1}, \eta_{it}) = \sum_{r=1}^{R} \alpha_r^Q(\tau) \phi_r^Q(z_{it-1})$$

where τ represents the τth conditional quantile of z_{it} given z_{it-1}, ϕ_r^Q is a dictionary of functions and α_r^Q the parameters associated which are quantile-specific, allowing the effect of z_{it-1} on z_{it} to change with the shocks. The quantile model is a direct non-parametric model for the conditional distribution of productivity, as it does not assume normality or impose separability in the productivity process, leaving the dependence structure of z_{it} unrestricted beyond the Markovian assumption. We also implement a more parsimonious model that is nonlinear in past productivity but separable in the shock.

(ii) For the policy functions, we use these nonlinear specifications:

$$i_{it} = \sum_{r=1}^{R} \alpha_r^h \phi_r^h \left(z_{it}, k_{it}, a_{it}, \delta_t^h\right) + v_{it}$$

$$a_{it+1} = \sum_{r=1}^{R} \alpha_r^g \phi_r^g \left(z_{it}, k_{it}, a_{it}, \delta_t^g\right) + w_{it+1}$$

$$l_{it} = \sum_{r=1}^{R} \alpha_r^n \phi_r^n \left(z_{it}, k_{it}, a_{it}\right) + w_{l,it+1}$$

where ϕ_r^h, ϕ_r^g and ϕ_r^n are dictionaries of functions and α_r^h, α_r^g and α_r^n are the associated parameters. Note that ϕ_r^h, ϕ_r^g and ϕ_r^n are anonymous functions without an economic interpretation, as they are just building blocks of flexible models. The objects of interest will be summary measures of the derivative effects constructed from the model, like the propensities discussed in Section 3. We follow the proxy variable literature and model the functions as high-order polynomials to allow for flexible interactions between productivity and observed state variables. In our baseline specification of the nonlinear model, we specify stationary policy functions with additive errors that are normally distributed to have a more parsimonious model to take to the data, but, as we showed in Section 4, the model is non-parametrically identified with time-varying functions, non-additive errors and without parametric distributions.

Stochastic EM Estimation Algorithm (SEM) We adapt the stochastic EM algorithm in Arellano and Bonhomme [2016], and Arellano et al. [2017] to our firm's framework to estimate the nonlinear model. See details in Online Appendix 3.

6 Data and Empirical Results

6.1 Data

We use administrative records generated by Chile's tax collection agency (*Servicio de Impuestos Internos* - SII). The records cover all firms that operate in the formal sector. Each firm is assigned a unique identifier by SII, so they can be tracked across time preserving anonymity. We use the information contained in income tax form F22, submitted annually by firms. The data set contains information on *firms* (as opposed to *plants*) of all sizes and sectors, although we focus on the manufacturing sector. Firms are defined as productive units that generate revenue, utilize production factors and operate under a unique tax ID. Form F22 has firm-level information on annual sales, expenditures on intermediate materials, a proxy for the capital stock ("immobile assets") and the firm's wage bill, as well as the firm's economic sector. We combine this information with tax form 1887, which reports monthly information on workers employed by the firm, and therefore allows us to calculate a measure of annual employment adjusted by the number of months per worker. Crucially, form F22 provides information on the firm's balance sheets. In particular, we can build a measure of net worth, defined as the difference between reported total assets and total liabilities[18]. This allows us to combine the information on the production side traditionally used in the literature on production functions with information on the firm's self-reported wealth and its evolution across time.

To clean up the raw data, we follow several steps. First, we drop observations with zero or missing information for the capital proxy, sales, expenditures on intermediate inputs, employment, or net worth. Second, we focus on firms that have at least five workers. Third, we build a measure of annual investment by using the annual change in the capital stock and assuming a 10% depreciation rate.[19] The final dataset has 4867 firms in the manufacturing sector between 2005 and 2016. As discussed earlier, the data set provides a panel of firms of different sizes and characteristics in the context of an emerging economy. Although we do not have information on whether firms are publicly traded, the relatively small coverage of the Chilean stock market (768 firms across all sectors) implies that almost all of our firms must be private. Having information on balance sheets is an advantage relative to most databases used in the literature on production function estimations, either from surveys or administrative records, which typically provide detailed information on the production side but do not account for assets or wealth. Moreover, we can directly observe wealth accumulation and investment decisions at the individual level, as well as the dynamics of output, inputs, and the estimated productivity process. The combination of financial statements and

[18] In particular, we use code 123 of form F22, "Total del Pasivo", for total liabilities. This variable is the combination of all the liabilities of the firm, as the tax form does not provide a decomposition between financial liabilities, credit from suppliers, etc. Similarly, total assets come from code 122, "Total de Activos", which combines all assets, including financial instruments as well as our capital proxy, "Activo Inmovilizado", code 647. Net worth is calculated simply as the difference between both. This means that our measure of physical capital (code 647) is equal to net worth (code 122 - code 123) plus total liabilities (code 123) net of non-capital assets (code 122 - code 647).

[19] As an alternative, we also use the information on tax form F29, which has monthly data on investment in machinery and equipment. The behavior of both investment series is very similar.

information on the production side is not unique to our dataset. Long and detailed information is available for a large number of countries in datasets such as Compustat, Amadeus, and Orbis. However, relative to those sources, our dataset has the advantage of including a heterogeneous set of firms operating in an emerging economy. In that sense, this might be a better setup to study the effects of financial frictions that are likely to be less relevant in the developed world, in particular for relatively large firms.[20]

Table 1 presents some descriptive statistics of the data. As expected, there is a large degree of heterogeneity between firms. Sales for firms in the 90th percentile are 40 times larger than those in the 10th percentile, while differences in capital or investment are even larger. While the average firm has 91 workers, the median firm has only 20, and firms in the 10th percentile have 7. There is also a large variation in net worth, both in levels as a ratio to capital. This highlights that the data contains a diverse set of firms, some of them quite small and with very low levels of wealth. While our data still has omissions (as it can not account for firms in the informal sector), it seems fit to provide a rich characterization of the behavior of heterogeneous firms over time, and the potential role of financial frictions in the context of a developing country and enriches the evidence previously available in the literature, in the spirit of the discussion in Kaboski [2021] and Diggs and Kaboski [2022].[21]

	Mean	p10	p50	p90
Value Added (million CLP)	1647.4	39.7	188.0	1536
Capital (million CLP)	2393.9	7.90	90.5	1197.9
Number of Employees	91.73	7	20	150
Investment (million CLP)	549.7	0.7	16.1	270.7
Net Worth (million CLP)	868.0	5.1	37.2	365.0
Capital to Output ratio	2.19	0.06	0.46	2.43
Net Worth to Capital Ratio	4.76	0.05	0.41	3.79

Table 1: Sample Descriptive Statistics

6.2 Empirical Results

We now use the data to implement the methodology discussed in Section 5. We begin by estimating firm-level production functions, accounting for the presence of financial frictions,

[20] Other datasets, such as the Enterprise Surveys conducted by the World Bank, are similar to ours in that they also include firms of all sizes in developing countries, although, by their nature, they are less suited to follow a specific firm across several consecutive years, as we do here.

[21] "Perhaps the biggest obstacle in researching financial frictions in developing countries is data availability. Ideally, data would consist of information on the firm ability and wealth over several years. Additionally, data may not include representative coverage of all firms. To have a full understanding of the firm side of an economy, it is necessary to include businesses across sectors and wealth, privately and publicly owned, and formal and informal firms."(Diggs and Kaboski, 2022)

and then use those estimates to study the properties of the productivity process. In the second part of the section, we present a novel empirical characterization of investment and wealth accumulation policy functions, highlighting the role of non-linearities and providing an empirical analysis of the self-financing channel.

In line with the earlier discussion, our estimates for the production function and the underlying productivity process, robust to financial frictions, differ significantly from those of the proxy variable approach. We also show that the productivity process of firms is highly non-linear, with larger persistence for highly productive firms. However, extremely large productivity shocks can change productivity. These findings are in stark contrast with the linear AR(1) productivity process typically assumed in the literature. Regarding policy functions, we find large heterogeneity in the sensitivity of investment and wealth accumulation to productivity shocks. We present novel evidence on the nonlinear relationship between investment and productivity, with larger investment responses to productivity shocks in more productive firms. Our results suggest that both collateral constraints and earning-based constraints are present. Finally, we also show novel evidence of the existence of self-financing, with a very large savings propensity in low-wealth, productive firms. However, the impact of self-financing appears to be limited, as convergence in the MPK between constrained and unconstrained firms is slow.

6.2.1 Production Functions

We start by comparing our two novel estimators that control for financial frictions (Proxy-IV and SEM) with OP -the proxy variable approach in Olley and Pakes [1996]- which uses investment as an auxiliary equation to recover productivity, and LP -the proxy variable approach of Levinsohn and Petrin [2003]-, which uses intermediate inputs to recover productivity.

As discussed earlier, we expect OP to underestimate the capital elasticity and overestimate the labor elasticity. By a similar argument, we expect the same type of bias in other methodologies relying on a proxy variable approach, such as LP. In contrast, by controlling for wealth in the policy functions, our estimators can discriminate between productivity and the effects of collateral constraints. In addition, by relying on the co-movements between wealth accumulation and investment decisions, after controlling for the current stock of net wealth, our estimators can disentangle productivity shocks from transitory shocks that can temporarily affect investment and saving decisions.

Table 2 presents the results of the full estimation of the production function parameters (β_l, β_k) using the four methodologies. The general pattern is consistent with the presence of financial constraints and with our earlier theoretical predictions.

The estimate of β_l is 0.67 for OP and, as expected, decreases significantly to 0.44 in

Proxy-IV and 0.46 in SEM. This estimated labor elasticity is in line with the labor share computed from aggregate data in Chile. Conversely, the opposite pattern holds for capital: the estimate of β_k is 0.35 for OP and increases to 0.42 for Proxy-IV and 0.43 for SEM. Similar biases appear in LP, which suggest that financial frictions are also present in demand for intermediate goods as in Mendoza and Yue [2012] and Bigio and La'o [2020].[22]

It is worth noting that the estimates of the production function parameters are very similar in proxy-IV and SEM. Even though we show that the model is non-parametrically identified from data, in order to devise tractable estimation methods, we impose some restrictions on the empirical model. Having robust results with both estimators suggests that neither the parametric assumption on the policy shocks in SEM nor the quasi-linear policy in IV affects the estimation of the production function parameters.[23]

These differences in input elasticities have relevant implications for the degree of returns to scale at the firm level, a crucial parameter to understand aggregate dynamics. In particular, OP results are consistent with constant returns to scale, while Proxy-IV and SEM both imply decreasing returns to scale with a span of control around 0.87. This figure lies on the upper end of the range used in the related literature.[24] This lower span of control relative to OP implies a larger entrepreneurial income share that can be retained by firms, allowing for faster wealth accumulation to overcome financial constraints.

To complement our results, we simulate data from an extended version of our stylized model to confirm the biases of the proxy variable approach and the robustness of our proposed estimators. In line with the empirical estimates, we set $\beta_k = 0.43$ and $\beta_l = 0.44$ in the calibrated model.[25] Table 3 presents the estimates for simulated data. As expected, OP delivers biased estimates, whereas Proxy-IV and SEM recover the true underlying parameters.[26] Therefore, data generated from a quantitative model, explicitly including financial frictions, supports our insights regarding the biases of traditional methodologies, as well as our novel estimators.

[22] Our proxy variables estimates of the production function are similar to ones in Gandhi et al. [2020]. Their proxy variable estimates with the Chilean data are 0.77 for β_l and 0.33 for β_k.

[23] For the proxy-IV, we assume that one of the policies is a quasi-linear function of degree 1 in productivity and leave the other policy function and the distribution of productivity and shocks completely unrestricted. In SEM, we allow all the policies to be nonlinear in all state variables, including productivity, but we parameterize the distribution of the policy shocks.

[24] For instance, Buera and Shin [2013a] use 0.79 while Midrigan and Xu [2014] use 0.85.

[25] See Online Appendix 4 for calibration details.

[26] As the model does not include intermediate inputs as required by the LP estimator, we only use the OP, Proxy-IV, and SEM estimators.

	OP	LP	Proxy-IV	SEM
β_l	0.67	0.81	0.44	0.46
	0.008	*0.007*	*0.01*	*0.003*
β_k	0.35	0.33	0.42	0.43
	0.05	*0.04*	*0.01*	*0.007*
σ_ϵ	0.68	0.62	0.22	0.20
Observations	13516	13516	13516	13516
Firms	4867	4867	4867	4867

Table 2: Production Function Estimates from Microdata

Note: The table shows the Production function estimates from administrative data for Chile, using alternative methodologies: OP - Olley and Pakes [1996]-, LP- Levinsohn and Petrin [2003]-, and two estimators that control for financial friction, Proxy-IV and SEM.

6.2.2 Productivity Process: Distribution

Figure 1 depicts the productivity distribution across firms for the proxy variable approach (OP, LP) and our more general model (SEM) that controls for financial frictions. In OP, the standard deviation of productivity is 0.18, significantly lower than 0.40 under SEM (see Table 4). We also find that the gap between ours and OP's productivity estimates, i.e., the fraction by which true productivity is underestimated, is increasing in the firm's productivity. For instance, the coefficient of a linear regression between $z_{it}^{SEM} - z_{it}^{OP}$ and z_{it}^{SEM} is 0.7.

The fact that OP dampens productivity differences between firms is once again consistent with the presence of financial frictions: OP underestimates the productivity of constrained, high-productivity firms. Conversely, the productivity of unconstrained but low-productivity firms, which can invest comparatively more, is overestimated. Hence, by ignoring firm wealth, OP estimates a more compressed distribution relative to methods that are robust to frictions. From figure 1 we can see that using the non-parametric quantile model (SEM-Quantile) delivers a very similar distribution than the one estimated using the nonlinear model with normal errors, so assuming normality for productivity seems to be a reasonable assumption.

6.2.3 Productivity Process: Persistence

As discussed earlier, the persistence of productivity is key for self-financing, as it relates to the incentives for wealth accumulation (Moll, 2014). If positive productivity shocks are not expected to last, incentives for wealth accumulation are weaker.

	OP	Proxy-IV	SEM
β_l	0.505	0.443	0.442
β_k	0.397	0.424	0.431

TABLE 3: Production Function Estimates Using Simulated Data

Note: Production function estimates from simulated data using alternative methodologies: OP - Olley and Pakes [1996]-, and two estimators that control for financial friction, Proxy-IV and SEM. The model used to generate data is described in Online Appendix 4.

To highlight the importance of controlling for financial frictions when estimating productivity persistence, we first show the estimates of a linear AR(1) model for productivity using the proxy variable approach and our proposed estimator. Table 4 presents the results for productivity persistence when we fit a linear model. The first row displays the autocorrelation of the estimated productivity, ρ_z. The estimated value of ρ_z raises from 0.53 under OP to 0.87 in proxy-IV and 0.85 under SEM, respectively. This implies that OP could underestimate the incentives for self-financing.

Non-linearities. In most quantitative macro models, productivity is assumed to follow an AR(1) process like the one in Table 4. As discussed in Section 3, one of the contributions of this paper is to uncover firm productivity without relying on either linearity or distributional assumptions. Figure 2 shows that the productivity process appears to be highly non-linear. Therefore, the traditional assumption of linearity might be at odds with the data. To disentangle the role played by past and new productivity shocks on the nonlinear persistence and also the role of a normal parametric assumption, we estimate two different models. The left panel of Figure 2 depicts the estimated persistence for different levels of initial productivity (horizontal axis) in a model that is non-linear in past productivity but separable in new shocks $z_{it} = \varphi(z_{it-1}) + \eta_{it}$ with η_{it} normally distributed. This model allows persistence to be heterogeneous across firms but does not allow new shocks to change the current persistence. The micro-data reveals high heterogeneity in productivity persistence with a positive monotonic relationship in past productivity. That is, firms at the lower end of the productivity distribution have smaller persistence (around 0.75), whereas ex-ante very productive firms display a very high persistence (close to 0.95). Therefore, highly productive firms will very likely remain at the upper end of the productivity distribution in the future. This novel result has important implications for the study of financial frictions and bodes

well for self-financing, as it suggests that highly-productive firms have both the ability and the incentives to build up collateral in order to converge toward their optimal capital.
We analyze this notion more formally in the next section when we embed the non-linear productivity process in the estimation of the policy functions of firms and evaluate whether wealth and investment decisions change with productivity.

The right panel (panel b) of Figure 2 displays the estimated persistence of the more flexible model $z_{it} = Q_z(z_{t-1}, \eta_{it})$. As discussed in Section 3, this model allows the persistence to change with past productivity and new shocks and do not restrict the conditional distribution of productivity. Thus, for a given value of a new productivity shock, the relationship between z_{it} and z_{it-1} depends on z_{t-1}, and for a given value of z_{it-1} the relationship between z_{it} and z_{it-1} may change in the face of extremely large (negative or positive) shocks. The 3-d graph displays the estimated persistence for different values of past productivity and new shocks. On the two horizontal axes, we report the percentile of past productivity and the percentile of the innovation (the shock) of the quantile process. A value at the lower end of the innovation distribution represents a very large negative shock, whereas a value at the upper end represents a very large positive shock. As before, we uncover a huge heterogeneity in persistence across firms that are in line with the results of the parsimonious model with normal errors . For the most common types of events (shocks) of a size close to the median of the shock distribution (the middle section of the right horizontal axis), the relationship between past productivity (left horizontal axis) and persistence (vertical axis) is positive and qualitatively consistent with the result in the left panel, implying that for *median shocks*, persistence is higher for ex-ante highly productive firms. Also, persistence increases in past productivity if firms experience shocks that align with their previous productivity. For instance, a low-productive firm that experiences a negative shock displays a persistence close to 0.75, whereas a high-productive firm that experiences a very positive shock displays a persistence close to 0.95. However, persistence can change abruptly in the face of extreme events.[27] For example, productivity persistence in very productive firms drops from almost one to 0.7 in the wake of an extremely adverse shock. A similar thing happens at the bottom of the productivity distribution after an extremely favorable shock. This means that large, infrequent shocks, besides having a direct effect on impact, can also alter the existing relationship between past and current productivity, canceling the cumulative effect of past shocks and permanently altering the trajectory of productivity. Therefore, in the aftermath of an unusually large shock, the incentives to self-finance can change drastically.

[27] This is consistent with similar results for household income shocks (Arellano et al. [2017])

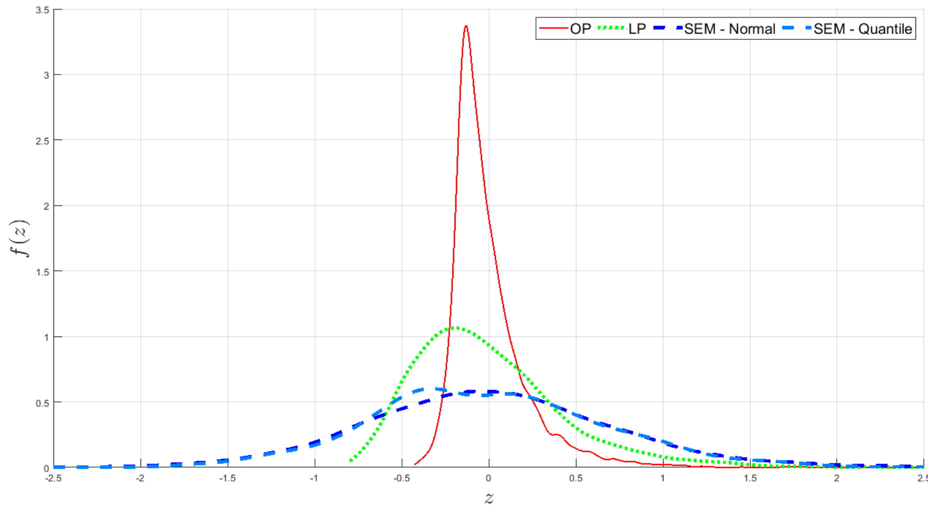

Figure 1: Estimated distribution of productivities

Note: The figure shows the estimated distribution of firm-level productivities using administrative microdata for Chile, under alternative methodologies: OP - Olley and Pakes [1996]-, LP- Levinsohn and Petrin [2003]-, and the SEM algorithm using Normal shocks and the SEM algorithm using a quantile model.

	OP	Proxy-IV	SEM
ρ_z	0.53	0.87	0.85
	0.01	*0.01*	*0.01*
σ_η	0.18	0.30	0.39
Observations	13516	13516	13516
Firms	4867	4867	4867
R^2	0.37	-	0.70

Table 4: Estimated Parameters of the Productivity Process

Note: The table shows the estimated parameters for the firm-level productivity process from administrative microdata for Chile, using alternative methodologies: OP - Olley and Pakes [1996]-, and the two estimators that control for financial frictions, Proxy-IV and SEM.

6.2.4 Policy Functions

We now present the estimated policy functions, one of the main goals of our empirical exercise. Given our interest in understanding the role of financial frictions and the self-

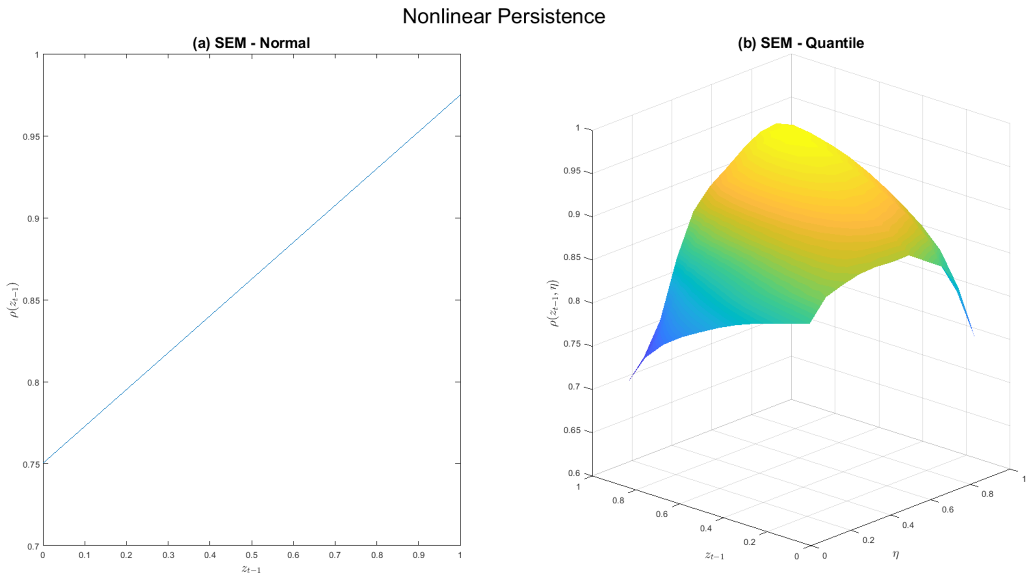

FIGURE 2: Estimated persistence of productivity

Note: The figure shows the estimated nonlinear persistence of firm-level productivity using administrative microdata for Chile. The first plot displays the persistence of estimated productivity using the model with separable errors along the distribution of past productivity, whereas the second plot displays the persistence of estimated productivity using the quantile model where the size and the sign of the shock might affect the persistence depending on the past value of productivity.

financing channel, we pay special interest to the estimation of policy functions and the analysis of the economic forces that underlie them.

6.2.5 Investment Policy Function

Nonlinear propensities As mentioned earlier, this is the first paper that estimates the investment policy function of a model with financial frictions nonparametrically using microdata without relying on approximations. The estimated propensities from the investment model inform us about how the behavior of firms in response to the same productivity shock varies with different levels of the state variables. Panels (a) and (b) of Figure 3 display the estimated average derivative effect of productivity on investment $\hat{\Phi}_t^h(a, k, z)$ (the investment propensity). The three-dimensional graphs show how the investment response to productivity changes for different combinations of $\frac{A}{K}$ and z. In panel (a), the propensities are evaluated at a fixed level of low k, whereas panel (b) presents the results for a fixed level of high k.[28] For a fixed capital level, evaluating the propensity for different wealth-to-capital ratios $\frac{A}{K}$

[28]Confidence intervals are presented in Online Appendix 5.

shows how the investment response to productivity shocks varies with the financial situation of the firm. Note that firms with lower $\frac{A}{K}$ are, by definition, highly-leveraged firms.

Estimated investment responses to productivity in Figure 3 are very heterogeneous, with values ranging from around 0.05 to 0.6. Propensities are the smallest in low-productivity firms with low levels of wealth. Arguably, these firms are less able to adjust investment after a positive and persistent productivity shock as they might be collateral constrained and can not rely too much on current and future earnings. However, investment propensities increase as we move along both the wealth and productivity distribution. In general, the sensitivity of investment to productivity shocks increases with z. This is, for given levels of wealth and capital, investment responses are larger for ex-ante, more productive firms. This is consistent with the characteristics of the estimated non-linear productivity process described in the previous subsection and the fact that persistence increases with productivity. Moreover, the higher propensity for high-productivity firms also appears to be consistent with the empirical implications of models of financial constraints in which firm productivity can affect firm lending contracts and borrowing opportunities[29], in which firms can use their future cash flows as collateral and expand their investment. For instance, the investment propensity of a high-productivity firm located at the bottom of the $\frac{A}{K}$ distribution is high (around 0.4). Despite these firms possessing few assets to be pledged as collateral (relative to their level of debt), they can strongly react to a positive productivity shock, a result that is at odds with a model in which only assets can be used as collateral. We take this as evidence that earning-based lending plays a role in the financing of private firms in Chile.[30]

Another important message from figure 3 is that investment propensities are increasing in wealth regardless of productivity. For instance, the investment propensity rises from 0.05 to 0.10 for low-productivity firms, from 0.23 to 0.35 for medium-productivity firms, and from 0.4 to 0.6 for high-productivity firms as we move along the wealth distribution. Interestingly, the biggest change (in levels) occurs for the most productive firms. This result might reflect that, even with earning-based constraints, productive firms with low wealth are still more financially constrained in relative terms and, therefore, benefit the most from an additional unit of wealth.

An interesting pattern that emerges from Figure 3 is that the positive relationship between investment propensity and wealth varies with the firm's productivity. For high-

[29]This is the case in models with earning-based constraints as in Lian and Ma [2020], and Drechsel [2022], or forward-looking constraints as in Aguirre [2017], and Brooks and Dovis [2020].

[30]The higher investment propensity for highly productive firms may also reflect a form of conditional convergence, as their current capital might be further away from their optimal capital relative to low-productivity firms. However, in the absence of earning-based constraints, under financial frictions, the investment of a low $\frac{A}{K}$ firm might not adjust, even if it is very productive.

productivity firms, the investment propensity converges to its maximum value of 0.6 for values of $\frac{A}{K}$ around 0.4, suggesting that productive firms with a debt-to-capital ratio lower than 0.6 are no longer constrained, whereas for low-productivity firms, the propensity is still increasing in $\frac{A}{K}$ for higher values of $\frac{A}{K}$, suggesting those firms still face relevant constraints. This has important implications for quantitative models with collateral constraints, as it implies that the collateral constraint parameter might be heterogeneous across firms depending on the productivity level.

Comparing panels (a) and (b) of figure 3, we can notice that the propensity is decreasing in k, and that this is stronger for highly-productive firms[31].

To have a taste on how propensities behave using the actual combinations of state variables in the data, we compute the propensity of each of the firms in our sample and plot it against the wealth-to-capital ratio $\frac{A}{K}$ in Figure 4 panels (a)-(c). We use our estimated productivity variable to cluster firms in three different "productivity groups": (i) low-productivity firms with productivity below the 50 percentile of the productivity distribution, (ii) median-productivity firms with productivity between the 50 and 75 percentile, and (iii) high-productivity firms with productivity above the 75 percentile. The data replicates the patterns suggested by the estimated policy functions. Investment propensities are increasing in $\frac{A}{K}$ and in z. As we can see, there is a positive relationship between the investment propensity and $\frac{A}{K}$ for all productivity levels, although the marginal impact of $\frac{A}{K}$ is decreasing in $\frac{A}{K}$. Moreover, propensities are more prominent for more productive firms.

For example, the investment propensity of low-productivity firms with little $\frac{A}{K}$ (panel (a)) is close to 0.1 on average. We can also see that for some firms with low productivity and very low wealth-to-capital ratios, the propensity is close to zero. However, the propensity increases up to 0.3 as we move along the distribution of $\frac{A}{K}$. Panels (b) and (c) show that the propensities for median- and high- productivity firms start at 0.25 and 0.45, respectively, much higher than the propensities of low-productivity firms with a similar level of $\frac{A}{K}$. As discussed earlier, a potential explanation is that these firms are more capable of adjusting investment because they can rely on current and future earnings. However, collateral constraints are also important for these firms, as propensities increase for firms with higher levels of $\frac{A}{K}$. The positive relation between investment propensities and $\frac{A}{K}$ for high-productivity firms suggests that a combination of earnings and wealth are essential for all firms with low levels of wealth. For high levels of $\frac{A}{K}$, propensities are roughly constant, as these firms are probably not constrained, and investment responses are close to optimal.

[31] This again is consistent with a notion of conditional convergence, in which highly-productive firms with low capital are further away from their optimal capital level than low-productivity firms with the same capital.

6.2.6 Wealth Accumulation Policy Function

Panels (c) and (d) of Figure 3 display the estimated average derivative effect of productivity on wealth accumulation (the nonlinear propensity) $\hat{\Phi}_{t+1}^{g}(a,k,z)$ using SEM. As before, this method allows the wealth accumulation policy function to be non-linear in productivity z. Hence, the three-dimensional graph presents how savings propensities change for different combinations of wealth and productivity. In almost all cases, the average derivative effect of productivity on savings decreases as wealth grows, consistent with the notion that self-financing is more important for firms with low wealth. Similarly, for a given combination of wealth and productivity, propensities are increasing in capital, consistent with the theoretical impact of larger leverage.

Regarding non-linearities, for a given level of capital, propensities are largest in firms that are highly productive but hold little wealth. In fact, the savings propensity to productivity shocks in firms on the upper end of the productivity distribution and the lower end of the wealth distribution is close to 1. This is, earnings shocks for highly productive but severely constrained firms are almost entirely saved, as the value of alleviating the constraint is comparatively large. As discussed earlier, this effect is reinforced by the larger persistence of productivity for highly-productive firms, which provides more incentives to wealth accumulation for productive firms as the theoretical mechanism in Moll [2014]. Consistent with the self-financing channel, the propensity decreases as we move along the wealth distribution since high-wealth firms are less constrained and have fewer incentives to save.

The savings propensity is also heterogeneous in productivity, as it is significantly lower for low-productivity firms, which are probably less constrained and have fewer incentives to save. However, at low wealth levels, even low-productivity firms save a considerable share of the earnings associated with a productivity shock when wealth is low (the propensity is 0.45). This propensity decreases to 0.2 as wealth increases.

We see similar patterns when we characterize saving propensities using the actual combination of all state variables that we see in the data (including estimated productivity) in figure 4 panels (d)-(f). Propensities are positive for all firms in the data and are increasing in productivity and decreasing in wealth. Again, the propensity is higher for high-productivity firms with low levels of wealth. As we discussed above, even for high-productivity firms that can also rely on earnings, the magnitude of the increment in the investment propensity as wealth increases is higher for high-productivity firms (see figure 4-(c)), so these firms have strong motives to save and accumulate wealth (see figure 4-(f)). The higher wealth accumulation propensity for very productive firms is consistent with the insights of a model where collateral and earning-based constraints interact. In a model with collateral and earning-based constraints, the marginal effect of wealth on investment is increasing in productivity:

an increase in wealth reduces borrowing constraints directly through the standard collateral constraint channel, generating an increase in investment and production, which in turn reduces borrowing constraints through the earnings-based constraint channel. This indirect channel is more potent for high-productivity firms than for low-productivity firms since their earnings increase more with the initial increase in wealth. The latter creates a higher incentive for high-productivity firms with low levels of wealth to increase savings and accumulate wealth in response to a positive productivity shock (see di Giovanni et al. [2022]).

6.2.7 Quantitative Implications: MPKs convergence

To get a more direct appraisal of the implications of our estimated policy functions for the self-financing channel, we use our data and estimates to look at the convergence of the marginal product of capital (MPK) between constrained and unconstrained firms in the spirit of the exercise in Banerjee and Moll [2010].

To do so, we use the data and our estimates of firm productivity and the production function to calculate the initial MPK of two firms that share the same level of initial productivity but have different levels of initial wealth and capital. We then use the estimated policy functions to simulate the evolution of their capital, labor, and wealth across time, assuming that productivity is constant and there are no additional shocks. Using the estimated production function parameters, we calculate the evolution of the MPK associated with the simulated capital and labor path.

Results are presented in Figure 5. For each row, the graphs plot the evolution of the marginal product of capital for a firm that starts on the lower end of the wealth distribution (10th percentile) vis-a-vis firms with the same level of productivity z but larger levels of initial wealth (50th percentile in the first column, 75th percentile in the second, 90th in the third). We report the convergence in MPKs between a constrained and unconstrained firm for three different productivity scenarios. The first row depicts firms in the 10th percentile of the productivity distribution, while the 50th and 90th productivity deciles are presented in the second and third rows.

Consistent with the self-financing channel, low-wealth, constrained firms are able to increase their capital stock across time, such that the marginal product of capital converges towards that of firms with similar firm productivity z but higher levels of initial wealth a_0. Convergence, however, is relatively slow, and marginal productivity gaps persist for decades. For example, across all three productivity levels, the marginal product of capital in a firm with initial wealth in the 10th percentile of the wealth distribution is close to three times larger than in a firm in the 90th wealth percentile. While this gap closes steadily across the years, marginal products in low-wealth firms are still at least twice as large as those of

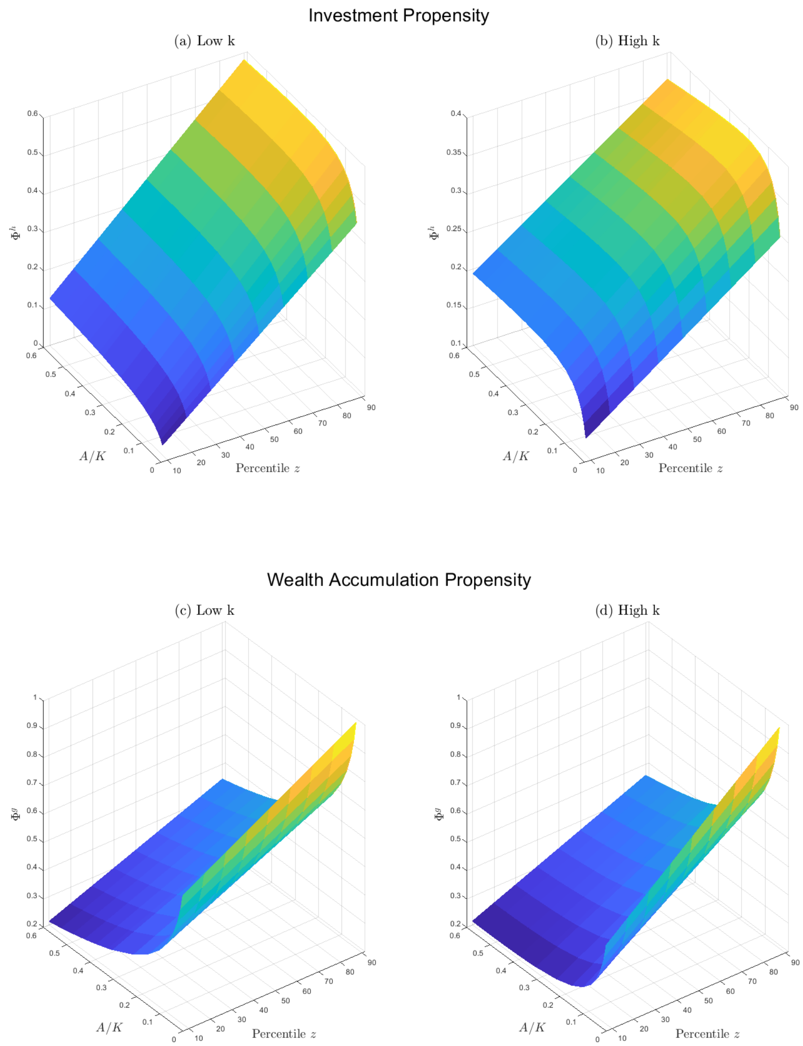

FIGURE 3: Nonlinear model: Investment and Wealth accumulation propensities to productivity

Notes: The figure exhibits the estimated derivative effect of productivity in the investment policy function (panels a and b) and the estimated derivative effects of productivity in the wealth accumulation policy (panels c and d) function using the SEM method. The estimated model is highly non-linear, so the figure displays the marginal effect for different values of productivity and the wealth to capital ratio for two different values of capital.

high-wealth firms after one decade. The speed of convergence in our data is much slower than in Banerjee and Moll [2010], where, for a similar initial gap, differences in marginal product between constrained and unconstrained firms vanish in less than a decade. For example, even for firms in the 90th percentile of the productivity distribution, convergence in

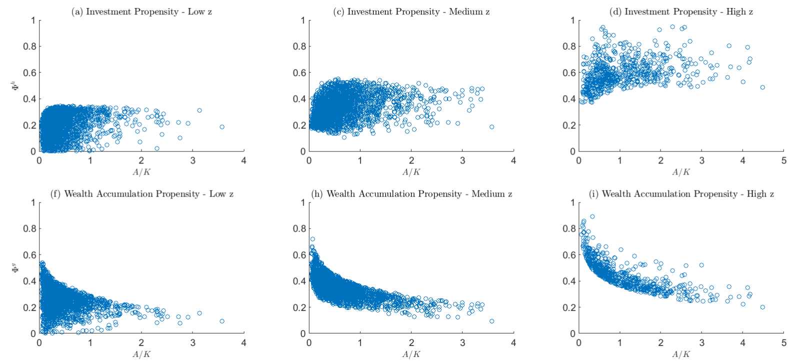

FIGURE 4: Investment and Wealth accumulation propensities in response to productivity

Notes: The figure exhibits how the investment and wealth accumulation propensity varies along the distribution of $\frac{A}{K}$ in the microdata for different productivity values. Each point represents the propensity of each particular firm evaluated at its actual value of a, k, and z. Figures (a), (b), and (c) are the investment propensities for low-, median- and high-productivity firms. Figures (d), (e), and (f) is the wealth accumulation propensities for low-, median- and high-productivity firms.

the marginal product of capital between firms in the 10th and 90th wealth percentiles takes almost 40 years, although half of the initial gap disappears after ten years.

Therefore, our results indicate that while the self-financing channel plays an important role in reducing productivity gaps and the extent of misallocation in this context, it cannot offset the persistence of significant differentials in marginal productivity over the medium term.

7 Conclusions

We provide an empirical analysis of wealth accumulation and investment dynamics in firms that operate under financial frictions and how these decisions relate to the unobservable firm's productivity process. We present a novel framework, robust to financial frictions, to jointly model and estimate the firm wealth accumulation dynamics, its investment decisions, and the unobservable productivity process.

Our results, using Chilean manufacturing data, show that the productivity process seems to be largely non-linear, with larger persistence for more productive firms, while persistence can change significantly in the face of extreme events. This differs significantly from the assumptions for the productivity process used in the structural macro literature.

We use our setup to provide a detailed analysis of the firm´s policy functions. We find that the behavior of firms is consistent with the predictions of macro models with financial frictions, although there are significant non-linearities. Results suggest that both collateral and earnings-based constraints are present in the data. We also find support for the existence

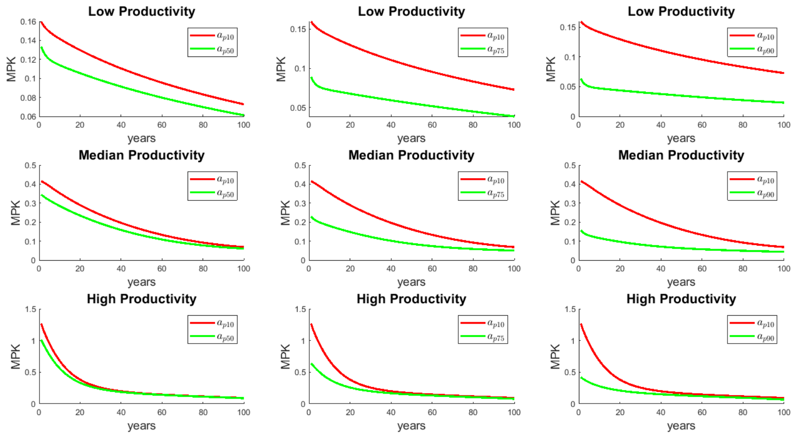

FIGURE 5: Convergence in the marginal product of capital across firms

Notes: The figure exhibits the simulated evolution of the marginal product of capital for firms with different levels of initial productivity and wealth. Low-wealth firms (10th percentile) are depicted in red, while high-wealth firms (50th percentile in column 1, 75th in column 2, and 90th in column 3) are depicted in green. The first row presents firms in the 10th percentile of the productivity distribution, while the second and third rows present figures in the 50th and 90th productivity deciles. The simulation uses the estimated production function and investment and wealth accumulation policy functions, holding firm productivity constant.

of an active self-financing channel, although its ability to overcome misallocation appears to be limited. This novel microeconomic evidence, as well as the methodology used to generate it, can provide support and guidance to the quantitative macroeconomic models that have studied the role of financial frictions and their aggregate implications.

References

Daniel A Ackerberg, Kevin Caves, and Garth Frazer. Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451, 2015.

Alvaro Aguirre. Contracting institutions and economic growth. *Review of Economic Dynamics*, 24:192–217, 2017.

Heitor Almeida, Murillo Campello, and Michael Weisbach. The cash flow sensitivity of cash. *Journal of Finance*, 59:1777–1804, 2004.

Isaiah Andrews, Matthew Gentzkow, and Jesse M Shapiro. Transparency in structural research. *Journal of Business & Economic Statistics*, 38(4):711–722, 2020.

Manuel Arellano and Stéphane Bonhomme. Nonlinear panel data estimation via quantile regressions. *Econometrics Journal*, 19(3):C61–C94, 2016.

Manuel Arellano, Richard Blundell, and Stéphane Bonhomme. Earnings and consumption dynamics: a nonlinear panel data framework. *Econometrica*, 85(3):693–734, 2017.

Abhijit Banerjee and Benjamin Moll. Why does misallocation persist? *American Economic Journal: Macroeconomics*, 2:189–206, 2010.

Saki Bigio and Jennifer La'o. Distortions in production networks. *The Quarterly Journal of Economics*, 135(4):2187–2253, 2020.

Richard Blundell, Luigi Pistaferri, and Ian Preston. Consumption inequality and partial insurance. *American Economic Review*, 98(5):1887–1921, 2008.

Wyatt Brooks and Alessandro Dovis. Credit market frictions and trade liberalizations. *Journal of Monetary Economics*, 111:32–47, 2020.

Francisco J Buera and Yongseok Shin. Self-insurance vs. self-financing: A welfare analysis of the persistence of shocks. *Journal of Economic Theory*, 146(3):845–862, 2011.

Francisco J Buera and Yongseok Shin. Financial frictions and the persistence of history: A quantitative exploration. *Journal of Political Economy*, 121(2):221–272, 2013a.

Francisco J Buera and Yongseok Shin. Financial frictions and the persistence of history: A quantitative exploration. *Journal of Political Economy*, 121(2):221–272, 2013b.

Francisco J Buera, Joseph P Kaboski, and Yongseok Shin. Finance and development: A tale of two sectors. *The American Economic Review*, 101(5):1964–2002, 2011.

Francisco J Buera, Joseph P Kaboski, and Yongseok Shin. Entrepreneurship and financial frictions: A macro-development perspective. *Annual Review of Economics*, 2015.

Francisco J Buera, Joseph P Kaboski, and Robert M Townsend. From micro to macro development. 2021.

Andrea Caggese and Vicente Cuñat. Financing constraints, firm dynamics, export decisions, and aggregate productivity. *Review of Economic Dynamics*, 16(1):177–193, 2013.

Tiago V Cavalcanti, Joseph P Kaboski, Bruno S Martins, and Cezar Santos. Dispersion in financing costs and development. Technical report, National Bureau of Economic Research, 2021.

Russell W Cooper and John C Haltiwanger. On the nature of capital adjustment costs. *The Review of Economic Studies*, 73(3):611–633, 2006.

Jan De Loecker. Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. *Econometrica*, 79(5):1407–1451, 2011.

Julian di Giovanni, Manuel García-Santana, Priit Jeenas, Enrique Moral-Benito, and Josep Pijoan-Mas. Government procurement and access to credit: Firm dynamics and aggregate implications. 2022.

Savita Diggs and Joseph Kaboski. Smoothing financial frictions for structural change. *Policy Brief, STEG Pathfinding Papers*, 5, 2022.

Ulrich Doraszelski and Jordi Jaumandreu. R&d and productivity: Estimating endogenous productivity. *The Review of Economic Studies*, 80(4):1338–1383, 2013.

Ulrich Doraszelski and Jordi Jaumandreu. Measuring the bias of technological change. *Journal of Political Economy*, 126(3):1027–1084, 2018.

Thomas Drechsel. Earnings-based borrowing constraints and macroeconomic fluctuations. *DP16975*, 2022.

Steven Fazzari, R Glenn Hubbard, and Bruce C Petersen. Financing constraints and corporate investment. Technical report, National Bureau of Economic Research, 1987.

Vito D Gala, Joao F Gomes, and Tong Liu. Investment without q. *Journal of Monetary Economics*, 116:266–282, 2020.

Amit Gandhi, Salvador Navarro, and David A Rivers. On the identification of gross output production functions. *Journal of Political Economy*, 128(8):2973–3016, 2020.

Gita Gopinath, Şebnem Kalemli-Özcan, Loukas Karabarbounis, and Carolina Villegas-Sanchez. Capital allocation and productivity in south europe. *The Quarterly Journal of Economics*, 132(4):1915–1967, 2017.

Hugo A Hopenhayn. Firms, misallocation, and aggregate productivity: A review. *Annu. Rev. Econ.*, 6(1):735–770, 2014.

Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

Yingyao Hu, Guofang Huang, and Yuya Sasaki. Estimating production functions with robustness against errors in the proxy variables. *Journal of Econometrics*, 215(2):375–398, 2020.

Victoria Ivashina, Luc Laeven, and Enrique Moral-Benito. Loan types and the bank lending channel. *Journal of Monetary Economics*, 126:171–187, 2022.

Joseph P Kaboski. Financial frictions, financial market development, and macroeconomic development. *Unpublished Working Paper*, 2021.

Greg Kaplan and Giovanni L Violante. How much consumption insurance beyond self-insurance? *American Economic Journal: Macroeconomics*, 2(4):53–87, 2010.

James Levinsohn and Amil Petrin. Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies*, 70(2):317–341, 2003.

Chen Lian and Yueran Ma. Anatomy of corporate borrowing constraints. *The Quarterly Journal of Economics*, 136(1):229–291, 2020.

Enrique G Mendoza and Vivian Z Yue. A general equilibrium model of sovereign default and business cycles. *The Quarterly Journal of Economics*, 127(2):889–946, 2012.

Virgiliu Midrigan and Daniel Yi Xu. Finance and misallocation: Evidence from plant-level data. *The American Economic Review*, 104(2):422–458, 2014.

Benjamin Moll. Productivity losses from financial frictions: Can self-financing undo capital misallocation? *American Economic Review*, 104(10):3186–3221, 2014.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Steven Olley and Ariel Pakes. The dynamics of productivity in the telecomunications equipment industry. *Econometrica*, 64:1263–97, 1996.

Tim Opler, Lee Pinkowitz, Rene Stultz, and Rohan Williamson. The determinants and implications of corporate cash holdings. *Journal of Financial Economics*, 52:3–46, 1999.

Ajay Shenoy. Estimating the production function under input market frictions. *Review of Economics and Statistics*, pages 1–45, 2020.

Lucciano Villacorta. Estimating country heterogeneity in capital-labor substitution using panel data. Technical report, Mimeo, 2018.

Jeffrey M Wooldridge. On estimating firm-level production functions using proxy variables to control for unobservables. *Economics letters*, 104(3):112–114, 2009.