STEG WORKING PAPER

PRODUCTIVITY, INVESTMENT AND WEALTH DYNAMICS UNDER FINANCIAL FRICTIONS: AN EMPIRICAL INVESTIGATION OF THE SELF-FINANCING CHANNEL

Alvaro Aguirre, Matias Tapia,
and Lucciano Villacorta

FEBRUARY 2023
STEG WP054

Productivity, Investment and Wealth Dynamics under Financial Frictions: An Empirical Investigation of the Self-financing Channel*

Alvaro Aguirre
Central Bank of Chile

Matias Tapia
Central Bank of Chile

Lucciano Villacorta
Central Bank of Chile

January 30, 2023

Abstract

We develop a new empirical framework to provide microeconomic evidence on the mechanisms underlying macroeconomic models with financial frictions and assess the self-financing channel. Using administrative panel data, we estimate firm-level productivity and its effect on firms' decisions. Our framework is robust to financial frictions, whereas standard methods used to estimate productivity dynamics are biased. The productivity process is largely non-linear, with larger persistence for more productive firms, while persistence can change significantly in the face of extreme events. We uncover a distribution of investment and wealth accumulation propensities in response to productivity shocks. These propensities are heterogeneous in the stock of wealth and productivity level: (i) investment propensities are larger for high-productivity firms and high-wealth firms, and (ii) wealth accumulation propensities are larger for high-productivity firms with low levels of wealth. We provide evidence of collateral-based and earning-based constraints. Our estimates support the existence of self-financing but show that its impact is limited.

JEL classification: C33, E23, O11, L0
Keywords: Investment propensity, wealth dynamics, self-financing, financial frictions, production function, productivity process, proxy variable approach; panel data, nonlinear.

*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 micro data in honor of Manuel Arellano, STEG Workshop on Firms, Frictions and Spillovers 2022, the BSE Summer Forum Workshop on Financial Shocks, Channels, and Macro Outcomes 2022, SED 2022, LACEA-LAMES 2022 and 2019, the 26th International Panel Data Conference, the 2020 World Congress of the Econometric Society, European Economic Association 2020, Santiago Macroeconomic Workshop 2020, PUC Chile, CEMFI, U. Diego Portales, Central Bank of Chile, and U Chile for their comments, and Diego Huerta and Cristian Valencia for excellent research assistance. This study was developed within the scope of the research agenda conducted by the Central Bank of Chile (CBC) in the economic and financial affairs of its competence. 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 are those of the authors and do not necessarily represent the views of the Central. Bank of Chile or its board members. 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 analyze 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 generate misallocation of 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.[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-level 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. For example, the persistence of firm-level productivity over time is crucial for the strength of self-financing [Moll, 2014]. If productivity is very persistent, productive firms with low levels of wealth will have larger incentives to build up wealth faster and increase their investment. More generally, micro-level evidence on the actual extent of self-financing for different types of firms can, if needed, guide targeted policies that can alleviate financial constraints. This paper attempts to bridge up this gap, bringing together the insights and lessons from the quantitative macro models with financial frictions with what we can learn from evidence using detailed firm-level data on production, investment, and savings decisions.

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, we are the first paper to estimate, using micro-data, the relevant policy functions for investment and wealth accumulation decisions under financial fictions derived from macro models. 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.

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

[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] As emphasized by 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".

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. We believe 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.

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 where the persistence of productivity shocks depends on the previous level of productivity and the magnitude and the sign of new shocks. Typically, persistence is increasing in productivity, with persistence being almost one for the most productive firms and as low as 0.75 for low-productivity firms. This has important implications for the strength of self-financing, as high persistence provides stronger incentives for highly productive firms to accumulate wealth. However, large productivity shocks can change the persistence of the productivity process. For example, extremely negative shocks to an ex-ante, very productive firm can reduce productivity persistence to 0.7, whereas extreme positive shocks to a low-productivity firm reduce persistence to 0.65. This means that large, infrequent shocks not only have a direct effect on impact but can also alter the existing relationship between past and current productivity. *Second*, we find heterogeneous patterns in the response of firm investment and wealth accumulation to productivity shocks. We are the first paper to document a nonlinear relationship between investment and productivity at the firm level, with larger responses in investment to productivity shocks at higher levels of previous productivity, regardless of the level of wealth. This result is consistent with the fact that more productive firms display higher productivity persistence. Also, consistent with models of collateral constraints, the investment reaction to positive productivity shocks is increasing in wealth for all productivity levels. For instance, the investment elasticity to a productivity shock goes from 0.05 to 0.10 for low-productivity firms, from 0.15 to 0.3 for medium-productivity firms, and from 0.4 to 0.6 for high-productivity firms when we move along the wealth distribution. Moreover, our results also lend support to the existence of earning-based constraints ([as in Lian and Ma, 2020, Drechsel, 2022]), as even the investment of productive firms located at the lower end of the wealth distribution strongly react to a positive productivity shock (the elasticity is close to 0.4 for a high-productivity firm at the lower end of the wealth distribution as opposed to 0.05 for a low-productivity firm with the same low wealth). *Third*, we are the first paper to document that the effect of productivity on wealth accumulation is heterogeneous in wealth and productivity in a way that is consistent with the notion of self-financing. In very productive, low-wealth firms, the transmission of persistent income shocks to savings is almost one-to-one. However, this response weakens significantly as we move upwards along the wealth distribution - reaching below 0.5 for highly productive, wealthy firms- or downwards along the productivity distribution - being below 0.2 for wealthy firms with low productivity-. *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 very slow,

as differences in MPK persist for almost four 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 economic models that study the self-financing channel [e.g. Buera and Shin, 2011, Moll, 2014, Midrigan and Xu, 2014]. However, 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 us with 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 wealth and productivity. Therefore, we can characterize the complete distribution of micro-level investments and wealth accumulation propensities in response to productivity shocks. In that respect, 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, Straub, 2019], which exploits panel data to estimate the degrees to which consumption decisions respond to unobserved household income shocks, but applied to a firm setup. 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. We extend this estimation method to allow for financial frictions, as we show that, 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. 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 policy functions. 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 data models with latent variables [Hu and Schennach, 2008, Hu and Shum, 2012, Arellano and Bonhomme, 2017, Arellano et al., 2017], we propose a sequential identification scheme to nonparametrically identify the production function, the productivity process, and the policy functions. From an instrumental variable perspective, both policy functions can be thought of as noisy measures of unobserved productivity. If conditional independence holds, such that the production function and both policy functions are independent conditional on productivity and observed state variables, the wealth accumulation policy provides an external instrument for investment (the proxy variable with noise) in the production function regression. Intuitively, due to self-financing, a positive co-movement between investment and wealth accumulation decisions is related to changes in productivity that can be used to identify the production function. Once the production function parameters are identified, the productivity process is identified from the dynamic dependence structure of the firm net income process, and the policy rules 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.

An important advantage of our approach (compared to a full structural estimation) is econometric transparency in the sense of Andrews et al. [2017], Andrews et al. [2020] and Bonhomme [2020]. First, we formally discuss identification and clearly show how the conditional independence assumption and the Markovian assumption–justified by the economic insights of structural models with financial frictions–enable us to construct dynamic restrictions that are used to identify the nonlinear policy functions, despite the presence of latent productivity. Second, our IV estimator is transparent, as it directly connects our estimates to the relevant moments and variation in the data that "drive" the estimator (see the discussion in Andrews et al. [2020]). 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 as in Catherine et al. [2018].

Our framework also uncovers 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.

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, Song et al., 2011, Buera and Shin, 2013b, Caggese and Cuñat, 2013, Moll, 2014, Midrigan and Xu, 2014], as we

guide our empirical specification by the general implications of these models, i.e., self-financing by incumbents undoes the effect of financial frictions and allows firms to invest closer to the optimal level. As mentioned, our main 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 estimations 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 not only on real variables but also on 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 the sensitivity of firms' investment to cash flows beyond profitability. Typically, profitability is captured by Tobin's Q or other observable characteristics of a firm. A second related area of literature discusses the determinants of firms' cash holding decisions and relates them to firm characteristics such as growth opportunities and risk management.[3] In our framework, the investment and wealth accumulation policy functions are two of our outcomes, and we are able to 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. Our results show that net wealth is a significant determinant of investment and that wealth accumulation decisions are affected by the firm's productivity process.

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 theoretically and empirically the biases that appear when the proxy method is used to estimate the production function under the environment of macro models with collateral constraints. Second, our paper uses the insights and economic mechanisms presented in those 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 in order 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 the identification of the model. Section 5 describes the estimation methods. Section 6 describes the data and presents the main empirical results. Section 7 concludes.

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

2 A Simple Model with Financial Frictions

Our starting point is a stylized structural model akin to the ones used in the macro literature to study and quantify the effects of financial frictions and the power of self-financing [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 used to identify the empirical model, and illustrates the nature of the biases incurred by the proxy variable approach under financial constraints,

Consider a firm maximization problem where a price-taking incumbent firm with initial wealth A_{it}, capital K_{it} and productivity Z_{it} 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}:

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

where Y_{it} is the value added produced by firm i. Investment, which determines the next period's capital, is decided before the firm observes its next period's productivity, while labor is decided contemporaneously with productivity.[4] The function $g(\cdot)$ is assumed to be convex, which given the use of linear preferences, rules out corner solutions.[5] The firm discounts future flows at β, capital depreciates at rate δ, and the firm pays interest rate r for its debt, implicitly defined by $K_{it} - A_{it}$.

As is standard in the literature, the log of productivity z_{it} follows a Markovian process

$$z_{it} = \varphi\left(z_{it-1}\right) + \eta_{it}, \tag{1}$$

where $\varphi\left(z_{it-1}\right) = E\left[z_{it} \mid z_{it-1}\right]$ is a non-parametric function of z_{it-1} and η_{it} is a shock.

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

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

Equation 2 implies that debt is limited by the repayment capacity of the firm through a combination of its productivity z_{it} and its current wealth a_{it}. This simple reduced form captures the idea that financial friction depends on the profitability of the firm and its financial status.

[4] This assumption implies that it takes a full period for new capital to be ordered, delivered, and installed. Some papers assume capital is chosen within the period to reduce the state space considerably [e.g. Midrigan and Xu, 2014].

[5] 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 corner 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.

The first term in 2 is known as *asset-based collateral constraints*, 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 factor in earning-covenants and earning-based lending (see Lian and Ma [2020]). Lian and Ma [2020] and Ivashina et al. [2022] provide substantial evidence that both type of financial constraints are prevalent in developed and developing countries, whereas Aguirre [2017], Brooks and Dovis [2020], Drechsel [2022] and di Giovanni et al. [2022] show that both constraints are quantitatively important.

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}. This wedge implies that MPKs will not equalize across firms, so that the equilibrium allocation of capital, which depends on the current distribution of A_{it} and Z_{it}, is not efficient. Equation (3) generates 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 will be crucial to finance future investments. The FOC, in this case, 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 provides a benefit if the constraint is expected to 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 marginal product of capital net of borrowing costs, the value of the multiplier. Since productivity is persistent, higher productivity today increases the expected marginal product of future capital, generating a positive correlation between productivity and wealth accumulation. As emphasized by Moll [2014] the higher the persistence of productivity the higher the firm's 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 intensity of financial frictions and the strength of self-financing are reflected in the responses of firm investment and wealth accumulation to persistent productivity shocks and how these responses depend on the amount of available collateral.

This simple setup illustrates the **goal of this paper**: to flexibly characterize, using micro-data, 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 the paper we follow the industrial organization literature and recover firm productivity without relying on distributional assumptions by estimating the firm production function. However, the model described above provides insights into the biases that appear when estimating these objects using standard empirical methods that do not account for financial frictions, and how these biases can distort the interpretation of the production function and the underlying productivity process. We illustrate our general 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].[6] These biases problematic, as consistent estimation of the production function and the productivity process is important given the crucial role played by these objects in structural macro models that quantitatively study the self-financing channel.

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.[7] The challenge in the estimation of β_l and β_k is that z_{it} is an unobservable variable for the econometrician which 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 relies on using 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 can eliminate the endogeneity problem and get consistent estimates of β_l and β_k. Intuitively, the OP method interprets observed differences in investment between firms in the data as differences in unobserved productivity between those firms.

However, in the model with financial frictions described above, the investment function arising from (3) does not only depend on productivity and initial capital, but also on net worth and its influence on the firms access to credit. Therefore, under borrowing constraints, equally productive firms with different levels of wealth might have different levels of investment, capital, and output. In consequence, if the implied heterogeneity in output is driven by heterogeneity in capital due to differences in financial frictions across firms, the OP approach will wrongly assign such dispersion to productivity, as it will misinterpret differences in investment as solely coming from differences in productivity. As a result, the OP productivity proxy captures an important part of the effect of capital on output, underestimating the true marginal impact of capital, and generating a downward bias en in the estimated coefficient for capital . Conversely, as long as financial frictions are relatively less severe in the labor market, 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 determining output, overestimating the labor elasticity. Furthermore, these biases in the estimates of factor elasticities can translate into

[6] We provide a detailed explanation of the biases in Appendix A.1.

[7] 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 production function estimation with monopolist competition and heterogeneous markups, see De Loecker [2011a,b] and Bond et al. [2021].

significant differences in 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. As a consequence, OP will deliver biased estimates of the productivity distribution across firms. Our empirical exercises illustrate these biases, and how our methodology can correct them, both with actual firm-level data as well from simulated data using a model with financial frictions.

3 General Empirical Framework

This section discusses the empirical model we take to the data 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).[8]

As our aim is 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 $(\eta_{it} \mid z_{t-1})$ uniform distributed. This quantile model is a direct non-parametric model for the distribution of productivity. Compared to the standard 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, the persistence of productivity after a given innovation can vary across firms that start at different productivity levels. 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 (positive or negative) shocks can change the relationship between current and past productivity and cancel the cumulative effects of past shocks. This modeling approach has been introduced by Arellano et al. [2017] to model persistent income shocks to households. We are the first paper that introduces this model 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 about the stochastic processes underlying both shocks are explained in detail at the end of this section. 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 stylized model discussed in the previous section, but each policy function is augmented by a stochastic shock:

$$i_{it} = h_t\left(z_{it}, k_{it}, a_{it}, v_{it}\right), \tag{9}$$

$$a_{it+1} = g_{t+1}\left(z_{it}, a_{it}, k_{it}, w_{it+1}\right). \tag{10}$$

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 in Section

[8] 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.

2. 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 the model studied in Section 2, shocks to collateral constraints could affect the investment policy function. It may well be the case that firms 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' financial portfolios or fixed assets. If these occur in the interim between the distribution of dividends (when equation 5 is solved) and the moment in which the firm uses wealth as collateral to borrow (when equation 4 is solved), these shocks they will appear as unplanned changes in the valuation of collateral.[9] More generally, the inclusion of stochastic shocks can also bridge the transition from the insights of the stylized model in Section 2 to an empirical model that deals with actual micro-data, and the presence of issues such as measurement error that can emerge from the use and combination 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 (9) 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 (9), to control for the existence of collateral constraints and (ii) the explicit modeling of wealth dynamics in (10) 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 these papers replace z_{it} for value-added.[10]

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. As a result, our main objects of interest are the following average derivative effects with respect to z_{it}:

$$\Phi^h_{it} = \Phi^h\left(a_{it}, k_{it}, z_{it}\right) = E_{v_{it}}\left[\frac{\partial h_t\left(z_{it}, k_{it}, a_{it}, v_{it}\right)}{\partial z}\right] \tag{11}$$

$$\Phi^g_{it+1} = \Phi^g\left(a_{it}, k_{it}, z_{it}\right) = E_{w_{it+1}}\left[\frac{\partial g_{t+1}\left(z_{it}, k_{it}, a_{it}, w_{it+1}\right)}{\partial z}\right] \tag{12}$$

where the expectations are taken with respect to the idiosyncratic shocks in the policy functions. Therefore, Φ^h_{it} and Φ^g_{it+1} measure the average propensities of investment and wealth accumulation in response to productivity shocks. Importantly, these propensities are heterogeneous, and vary with the firms' stock of wealth and productivity level. By characterizing how these propensities vary along the wealth and productivity distributions we provide novel evidence on financial frictions and the self-financing channel. For example, in collateral constraint

[9] The inclusion of v_{it} in the investment policy 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 not an object of interest by itself, as it only serves a role as an auxiliary equation to recover the production function.

[10] 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.

models, Φ_{it}^h is increasing in a_{it}, whereas in earning-based models [Drechsel, 2022] and forward looking constraint models [Buera et al., 2015] Φ_{it}^h it is increasing in z_{it}. Moreover, in models with a self-financing channel as in Moll [2014], Φ_{it+1}^g is always positive and decreasing in a_{it}.

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

$$l_{it} = n_t\left(z_{it}, a_{it}, k_{it}, w_{l,it}\right), \tag{13}$$

where equation (13) 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 a shock that is independent across periods and 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 shocks 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-(i) is a natural assumption in macro models with a self-financing channel as the one presented earlier; Assumption 2-(ii) is a standard assumption both in macro models as well as in the empirical literature that estimates production functions.

4 Identification

In this section, we establish the identification of the nonlinear dynamic panel model presented in the previous section. Identification challenges in our model are more demanding than those of firm dynamics models studied in the proxy variable literature due to the presence of additional shocks in the policy functions. Therefore, it is important to show that the model we aim to estimate can actually be identified from the data. Our model takes the form of nonlinear state-space models. Recently, Hu and Schennach [2008], Hu and Shum [2012], 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}, ε_{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 impose the following high-level conditions that connect with the 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 desity $f_{A|B}\left(a \mid b\right)$.

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.[11] 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. Also, macro models with only asset-based constraints fulfill this condition as long as the financial constraint is a soft constraint where firms can borrow as much as they want, paying a premium in the interest rate that depends on the level of 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].[12] 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

[11] 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].

[12] 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.

in Section 2, where conditional on the same level of financial friction, 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 $(a_{it+}, z_{it}, a_{it}, k_t)$ follows a multivariate normal distribution with zero mean. In that case, the completeness condition will require that $E[a_{it+1} z_{it}] \neq 0$, which is ensured by the self-financing channel. Similarly, 1-(v) is a completeness condition that 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. *(Sequential identification) In a production function model with Markovian Hicks-neutral productivity and financial frictions as in (7)-(13), if Assumption 1, Assumption 2 and condition 1 (i)-(v) hold, then β_k, β_l, $Q_z(z_{t-1}, \eta_{it})$, h_t, g_{t+1} are identified from data on y_{it}, k_{it}, l_{it}, i_{it}, a_{it} for $T \geq 4$.*

The sketch of identification is sequential. First, we identify the production function parameters β_k and β_l. We then establish the identification of the productivity process, and finally, we show the identification of the policy functions h_t and g_t. Below we discuss the sketch of the sequential identification and leave the details for Appendix A.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 following 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:

$$f(y_t, i_t \mid a_{t+1}, X_t) = \int f(y_t \mid z_t, k_t, l_t) f(i_t \mid z_t, X_t) f(z_t \mid a_{t+1}, X_t) dz_t \tag{14}$$

We notice that equation (14) 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 investment and wealth accumulation policy functions are nonlinear in Appendix 2 (see Proposition 1) and to build intuition below we discuss 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 $a_{it+1} = g_z z_{it} + g_a a_{it} + w_{it+1}$.[13] Both policies give us information on z_{it}. Similar to the proxy variables, we can invert the

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

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:

$$y_{it} = \beta_l l_{it} + \beta_k k_{it} + \pi_1 i_{it} + \pi_2 a_{it} + \tilde{\varepsilon}_{it} \tag{15}$$

where $\tilde{\varepsilon}_{it} = \varepsilon_{it} + \pi_4 v_{it}$. In the absence of investment shocks, a simple OLS regression between 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.[14]

The self-financing channel is key for identification. According to the model in Section 2, more productive firms should increase investment 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}. The identification sketch that we develop here provides a direct and simple estimation procedure by doing an IV regression to the proxy method where the external instrument is justified by the theoretical insights of macro models. Hence, a_{it+1} can be used as an instrument for investment in equation (15,) 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 $\{\beta_l, \pi_1, \pi_2\}$.

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). [15]

$$\tilde{y}_{it} = z_{it} + \varepsilon_{it} \tag{16}$$

Given that z_{it} is Markovian and ε_{it} is i.i.d over time, equation (16) 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 (16) 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 $z_{it} = \rho_z z_{it-1} + \eta_{it}$, we can use equation (16) 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

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

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

$\tilde{y}_{it-2}$ 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 A.2) 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 in appendix A2).

$$f\left(a_1, k_1 \mid \tilde{y}^T\right) = E\left[f\left(a_1, k_1 \mid z_1\right) \mid \tilde{y}_i^T = \tilde{y}^T\right] \tag{17}$$

$$f\left(a_{t+1} \mid a^t, k^t, \tilde{y}^T\right) = E\left[f\left(a_{t+1} \mid z_t, a_t, k_t\right) \mid a_i^t = a^t, k_i^t = k^t, \tilde{y}_i^T = \tilde{y}^T\right] \tag{18}$$

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

where the expectation in (17) 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 (18) and (19) 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 (17) 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 (17) 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 (17). Similarly, equations (18) and (19) 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 (18) and (19). 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 policy functions. 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 (16) 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 parameters of 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 empirical model presented in Section 3. First, we consider a parsimonious model where at least one of the policies is a quasi-linear function in productivity and separable in productivity and the policy

shock. 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 panel data 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 a polynomial of degree one in z_{it} and separable in z_{it} and the policy shock. The other policy function as well as the distribution of the shocks are left unrestricted. For example, consider the following investment policy function:

$$i_{it} = h\left(z_{it}, k_{it}, a_{it}, v_{it}\right) = h_1\left(k_{it}, a_{it}\right) + h_2\left(k_{it}, a_{it}\right) z_{it} + v_{it}, \tag{20}$$

It is important to notice that model (20) 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 (20) in terms of productivity:

$$z_{it} = \pi_1\left(k_{it}, a_{it}\right) + \pi_2\left(k_{it}, a_{it}\right) i_{it} + \omega_{it} \tag{21}$$

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 (21) in the production function:

$$y_{it} = \beta_l l_{it} + \phi\left(k_{it}, a_{it}\right) + \pi_2\left(k_{it}, a_{it}\right) i_{it} + \omega_{it+1} + \varepsilon_{it}, \tag{22}$$

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 (22) does not provide a consistent estimator of β_l since $E\left(\omega_{it} \mid i_{it}\right) \neq 0$. However, given *Assumption 1*, a_{it+1} can be used as an instrument for i_{it} in equation (22). Therefore, we propose the following two-stage procedure:

First Stage: Estimate (22) with an IV estimator using $\pi_2\left(k_{it}, a_{it}\right) a_{it+1}$ as the instrument for $\pi_2\left(k_{it}, a_{it}\right) i_{it}$. The IV regression delivers a consistent estimator of β_l, $\phi\left(k_{it}, a_{it}\right)$ and $\pi_2\left(k_{it}, a_{it}\right) a_{it+1}$. For instance, in the linear case where $g_2\left(k_{it}, a_{it}\right) = 1$, a_{it+1} will be the instrument for i_t.

Second Stage: Combining equation (21) with the Markovian model for the productivity process $z_{it} = \rho_z z_{it-1} + \eta_{it}$:

$$z_{it} = \rho_z \pi_1\left(k_{it-1}, a_{it-1}\right) + \rho_z \pi_2\left(k_{it-1}, a_{it-1}\right) i_{it-1} + \rho_z \omega_{it-1} + \eta_{it}, \tag{23}$$

Replacing equation (23) into the production function:

$$y_{it} - \beta_l l_{it} = \beta_k k_{it} + \rho_z \pi_1\left(k_{it-1}, a_{it-1}\right) + \rho_z \pi_2\left(k_{it-1}, a_{it-1}\right) i_{it-1} + \rho_z \omega_{it-1} + \eta_{it} + \varepsilon_{it}, \tag{24}$$

using *assumption 1* we can define the following moment condition from equation (24)

$$E\left(\omega_{it-1} + \eta_{it} + \varepsilon_{it} \mid k_{it}, k_{it-1}, a_{it-1}, a_t\right) = 0, \tag{25}$$

The moment condition in (25) 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:

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

where τ represent 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 is 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.

(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 production function framework to estimate the nonlinear model with latent variables. See details in Appendix A.3.

6 Data and Empirical Results

6.1 Data

Our database comes from administrative records generated by Chile's tax collection agency (*Servicio de Impuestos Internos* - SII). The records covers all firms that operate in the formal sector and all formal wage employment in Chile. Each firm in this administrative dataset is assigned a unique identifier by SII, so they can be tracked across time while at the same time preserving anonymity and the confidentiality of the data. We use information contained in income tax form F22, which is submitted annually by firms. The data set contains information on *firms* (as opposed to *plants*) of all sizes and sectors, although we focus on firms operating in the manufacturing sector. Firms are defined as productive units that generate revenue, utilize production factors and operate under a unique tax ID that allows us to track them across time. Data is available on an annual frequency.

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 individual workers that were employed on the firm, and therefore allows us to calculate a measure of annual employment adjusted by the number of months per worker.

Crucially, form F22 also 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[16]. This allows us to combine the information on the production side traditionally used in the literature on production functions and TFP estimates with information on the firm's self-reported wealth and its evolution across time.

To clean up the raw data and have a dataset that is consistent with our empirical strategy, 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 5 workers. Third, we build a measure of annual investment by using the annual change in the capital stock and assuming a 10% depreciation rate[17] 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 over whether firms are publicly traded, the relatively small coverage of the Chilean stock market (768 firms across all sectors) implies that a very large share 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

[16] 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).

[17] 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.

provide detailed information on the production side of firms 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 information on the production side is not unique to our dataset, as similar data is available, with long and detailed information 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 an 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. 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.

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 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 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 Diggs and Kaboski [2022] [18].

Finally, a relevant reference point for the dataset used in this paper is ENIA, the manufacturing sector survey for Chilean firms that has been widely used in the literature (see, for example, Gandhi et al. [2020], among many others). Similarly to this dataset, ENIA has rich information on production, investment and employment, but is silent regarding the firm's financial position, so it cannot be used to implement our framework. Interestingly, the OP estimates of the production function parameters using our administrative set reported in the next section are similar to those that can be obtained using similar methodologies with ENIA. This provides a form of external validation to this dataset in the sense that it is associated to estimates for Chile that are quantitatively consistent with those obtained in the large literature that has used ENIA.

6.2 Empirical Results

We now use the data presented in the previous section to implement the empirical methodology discussed in Section 5. Following our previous discussion, we begin by estimating firm-level production functions, correctly accounting for the presence of financial frictions, and showing

[18] "Perhaps the biggest obstacle in researching financial frictions in developing countries is data availability. Ideally, data would consist of information on firm ability (potential) 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, both privately and publicly owned, and formal and informal firms." (Diggs and Kaboski, 2022)

	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

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

Our production function estimates, robust to financial friction, differ significantly from those obtained with the proxy variable approach. The differences align with the predictions from the stylized model described in Section 2. The same holds true for the underlying productivity process. We show that the productivity process of firms is highly non-linear, as the persistence of productivity shocks depends on the previous level of productivity and the features of new productivity shocks. Crucially, the persistence is heterogeneous across firms with higher values for ex-ante high productive firms. However, extremely large shocks can change the persistence of the productivity process. Therefore, an unusually large negative productivity shock on a very productive firm can permanently alter the relationship between past and future productivity. These findings are in stark contrast with the standard linear AR(1) productivity process typically assumed in the literature.

Regarding the characterization of the policy functions, we find a large degree of heterogeneity in the sensitivity of firm investment and wealth accumulation to productivity shocks. We present novel evidence on the nonlinear relationship between investment and productivity at the firm level, with larger responses in investment to productivity shocks in firms with higher levels of previous productivity, at all levels of wealth. Our results are consistent with the notion of collateral constraints, with a more considerable sensitivity to productivity in firms with larger wealth, but also with the existence of earning-based constraints, as productive firms with low levels of wealth are still able to invest more in the face of positive productivity shocks. We also show novel evidence supporting the existence of self-financing at the firm level, with the savings propensity to productivity being very large in low-wealth, productive firms. However, the impact of self-financing appears to be limited, as convergence in the marginal product of capital between constrained and unconstrained firms is slow.

6.2.1 Production Functions

To highlight the importance of considering financial frictions in the estimation of the firm production function and the firm productive process, we start by comparing the results of 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 provides our main benchmark to previous literature. To have an alternative benchmark we also compare our results with LP -the proxy variable approach of Levinsohn and Petrin [2003]-, which uses intermediate inputs as an auxiliary equation to recover productivity.

As discussed in Section 2, we expect OP to underestimate the capital elasticity, and to overestimate the labor elasticity, as it incorrectly interprets differences in value added due to financial constraints as generated by differences in productivity and not in capital. This bias in estimated productivity makes the co-movement of output and labor stronger than expected, generating a larger estimated elasticity of labor. By a similar argument, we expect the same type of bias in other methodologies relying on a proxy variable approach, such as LP.

Table 2 presents the results of the full estimation of the production function parameters (β_l, β_k) using the four methodologies. There are significant differences across estimators, with a general pattern that is consistent with the presence of financial constraints and with the theoretical predictions derived earlier. Controlling for wealth in the policy functions allows us to 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, we can disentangle productivity shocks from transitory shocks that can temporarily affect investment and saving decisions. The estimate of β_l is 0.67 for OP, and, as expected, decreases significantly for the estimates that are robust to financial constraints and allow for shocks to the policy equations, to 0.44 in Proxy-IV and 0.46 in SEM.

Conversely, the opposite pattern holds for the elasticity of 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 the LP estimators, which suggest that financial frictions are also present in the demand for intermediate goods as in Mendoza and Yue [2012] and Bigio and La'o [2020].[19]

These differences in the estimated 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 (for instance, Buera and Shin [2013a] use 0.79 while Restuccia and Rogerson [2008] and Midrigan and Xu [2014] use 0.85). This lower span of control relative to OP implies a larger entrepreneurial income share that can be retained by firms, which allows for a faster accumulation of wealth to overcome financial constraints.

To complement our results, we simulate data from an extended version of the stylized model presented in section 2 to confirm the biases of the proxy variable approach and the robustness of our proposed estimators. In line with the empirical estimates with the Chilean data, we set $\beta_k = 0.43$ and $\beta_l = 0.44$ in the calibrated model. See Appendix A.4 for model and calibration details.

Table 3 presents the estimates for simulated data. As expected, OP delivers biased esti-

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

mates, whereas Proxy-IV and SEM recover the true underlying parameters.[20] Therefore, data generated from a quantitative model, which explicitly includes financial frictions and the theoretical mechanisms described in Section 2, provides validation to our insights regarding the biases of traditional methodologies in the presence of financial constraints, as well as validation to our novel 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.

	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 Appendix A.4

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

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. There are relevant differences in the dispersion of the estimated productivity distributions across both methodologies. In OP, the standard deviation of productivity is 0.18, significantly lower than 0.40 under SEM (see Table 4). After controlling for financial frictions, we can see that extreme productivity shocks are more likely to occur (relative to a productivity estimator that does not consider financial frictions). Underestimating the productivity dispersion could have aggregate implications as this parameter affect the conclusions of quantitative exercises. In quantitative models, a lower dispersion implies less productivity fluctuation over time, reducing the effects of financial frictions in distorting allocations. 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 productivity level of the firm. 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 differentials across firms is once again consistent with the presence of financial frictions: as their actual investment is relatively low, 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 the methods that are robust to frictions.

6.2.3 Productivity Process: Persistence

As discussed earlier, the persistence of productivity is a key object for the self-financing channel, as it directly relates to the incentives for wealth accumulation. For instance, Moll [2014] shows that low persistence in productivity leads to large effects of financial frictions on aggregate TFP, as the self-financing channel is less powerful. This is the result of weaker incentives for wealth accumulation when positive productivity shocks are not expected to last for long.

To highlight the importance of controlling for financial frictions when estimating productivity persistence with micro-data, we first show the results of the estimation of a linear AR(1) model for productivity using the proxy variable approach and our estimator robust to financial frictions. Table 4 presents results for productivity persistence when we fit a linear model. The first row displays the autocorrelation of the estimated productivity ρ_z.We can see that the estimated autocorrelation is considerably lower under OP. 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 potentially 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, implying that the assumption of linearity used in the previous literature might be at odds with the data. To disentangle the role played by past productivity shocks and new productivity shocks on the nonlinear persistence, we estimate two different models. The left panel of Figure 2 depicts the estimated persistence of productivity for different levels of initial productivity

(horizontal axis) from a model that is non-linear in past productivity but separable in new shocks $z_{it} = \varphi(z_{it-1}) + \eta_{it}$. This model allows the persistence to be heterogeneous across firms but does not allow new shocks to change the current persistence. Interestingly, 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 display a smaller persistence (around 0.67), whereas ex-ante very productive firms display a very high persistence (close to one). Therefore, highly productive will very likely remain at the upper end of the productivity distribution in the future, while low-productivity firms have a shot at becoming more productive in the future. This novel result has important implications for the quantitative macro literature that study the role of financial friction. According to models as in Moll [2014] and Buera et al. [2015], high persistence in productivity reinforces the incentives for self-financing. The fact that persistence is the highest for highly productive firms bodes well for the notion of self-financing, as it suggests that these firms have both the ability and the incentives to build up collateral in order to grow out of their financial constraints and 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. 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 different values of 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 depending on the size and sign of the shock. For the most common types of shocks-events 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 0.7, whereas a high-productive firm that experiences a very positive shock displays a persistence close to one. However, persistence can change abruptly in the face of extreme events, consistent with similar results for household income shocks (Arellano et al. [2017]). 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 to firms at the bottom of the productivity distribution that face 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, permanently altering the trajectory of productivity. Therefore, in the aftermath of an unusually large shock, the incentives to self-finance can change drastically.

For example, in the wake of a very adverse productivity shock, a previously highly productive firm might be much less willing to accumulate wealth towards the future.

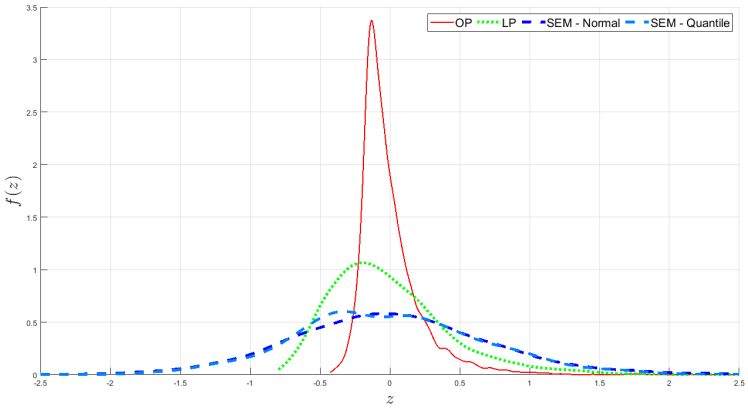

FIGURE 1: Estimated distribution of productivities

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.

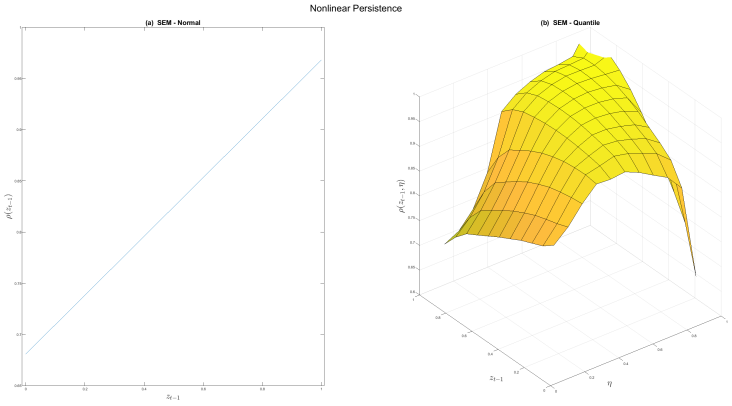

FIGURE 2: Estimated persistence of productivity

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 normal errors along the distribution of initial 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 initial value of productivity.

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-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 nonparametrically the investment policy function of a model with financial frictions 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 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 response of investment 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.[21] For a fixed capital level (a given firm size), evaluating the propensity for different wealth-to-capital ratios $\frac{A}{K}$ allows us to analyze how the investment response to productivity shocks depends on the financial situation of the firm. Note that firms with lower values of $\frac{A}{K}$ are, by definition, highly-leveraged firms.

Figure 3 shows that the estimated investment response to productivity is very heterogeneous,

[21] Confidence intervals are presented in Appendix A.5.

with values ranging from 0.05 to 0.6. Propensities are the smallest in low-productivity firms with low levels of wealth. This is consistent with the idea that these firms are less able to adjust investment in response to a positive and persistent productivity shock as they might be collateral constrained and can not rely too much on current and future earnings. Therefore, this provides evidence for the existence of financial frictions in the form of collateral constraints.

However, investment propensities increase as we move along both wealth and productivity distribution.

In general, the sensitivity of investment to productivity shocks increases with z. This is, for a given level of wealth and capital, investment responses to productivity shocks are larger for ex-ante, more productive firms. This 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, as is the case in the models with earning-based constraints as in Lian and Ma [2020], Drechsel [2022], and di Giovanni et al. [2022] or forward-looking constraints as in Aguirre [2017], DeMarzo and Fishman [2007] and Brooks and Dovis [2020], in which firms can use their future cash-flows as collateral. In those models, more productive firms are able to take more debt for a given net wealth level and expand investment more. 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). Even if these firms are collateral constrained, their investment can strongly react to a positive and persistent productivity shock, as opposed to the response of low productivity firms with low levels of $\frac{A}{K}$. 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. Moreover, the larger investment propensity for more productive firms is also consistent with the characteristics of the estimated non-linear productivity process described in the previous subsection, and the fact that persistence increases with the level of productivity.

Another important message from figure 3 is that investment propensities are increasing in wealth for all values of productivity. For instance, as we move along the wealth distribution, the propensity of low-productivity firms doubles from 0.05 to 0.1. For the high-productivity firms, propensities increase from 0.4 to 0.6, the largest propensity increment (in levels) when we move from the bottom to the top of the wealth distribution. This result might reflect that, even with earning-based constraints, productive firms with low levels of wealth (net worth) are still more financially constrained (in terms of the distance to their optimal size), and therefore are the ones that benefit the most from an additional unit of wealth.

An interesting pattern that emerges from Figure 3 is that the positive relationship between the firm's investment propensity and wealth varies with the firm's productivity. For high-productivity firms, the investment propensity converges to its maximum value of 0.6 for values of $\frac{A}{K}$ around 0.4, suggesting that those firms 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 that those firms still face relevant constraints.

Comparing panels (a) and (b) of figure 3 we can notice that the propensity is decreasing in k, with a more pronounced pattern for highly-productive firms. 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 level of capital.

To have a taste of how propensities behave using the actual combinations of state variables that we see in the data, we compute the propensity of each of the firms in our sample, and we plot it against the wealth-to-capital ratio $\frac{A}{K}$ in figure 4 panels (a)-(c). To analyze heterogeneity across different productivity levels, 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, indicating that the investment of those firms does not react much to productivity shocks. However, the propensity increases up to 0.3 as we move along the distribution of $\frac{A}{K}$. This positive relationship between the investment propensity and $\frac{A}{K}$ is present across all three productivity groups. Panels (b) and (c) show that the propensities for median- and high- productivity firms start at 0.25 and 0.45, respectively. These propensities are 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 crucial for these firms, as propensities increase for firms with higher levels of $\frac{A}{K}$. For median- and high-productivity firms, the propensity increases up to 0.4 and 0.6 (on average) and stabilizes around these numbers. The positive relation between investment propensities and $\frac{A}{K}$ for high-productivity firms suggests that earnings are not sufficient to self-finance all the investment, and a combination of earnings and wealth are essential for all firms with low levels of wealth as in di Giovanni et al. [2022]. 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.

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) $\hat{\Phi}^{g}_{t+1}(a, k, z)$ 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 levels of wealth. Similarly, for a given combination of wealth and productivity, in most cases, 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]. As predicted by theoretical models with the self-financing channel, the propensity decreases as we move along the wealth distribution (up to half of the value) 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 (the propensity is between 0.4-0.5) when wealth is low. 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 investment propensity increases with wealth (see figure 4-(c)), so these firms also 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. As emphasized in di Giovanni et al. [2022], even with earnings-based constraints, more productive firms are more financially constrained (in terms of the distance to their optimal size) for a given level of wealth (net worth) and are the ones that benefit the most from an additional unit of wealth. In the benchmark parametrization of di Giovanni et al. [2022], a high-productivity firm has better access to credit than a low-productivity firm, for the same level of wealth, through the earning-based constraint. However, the high-productivity firm is still "more constrained" since it is further away from its optimal capital level.

Moreover, in models 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.

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

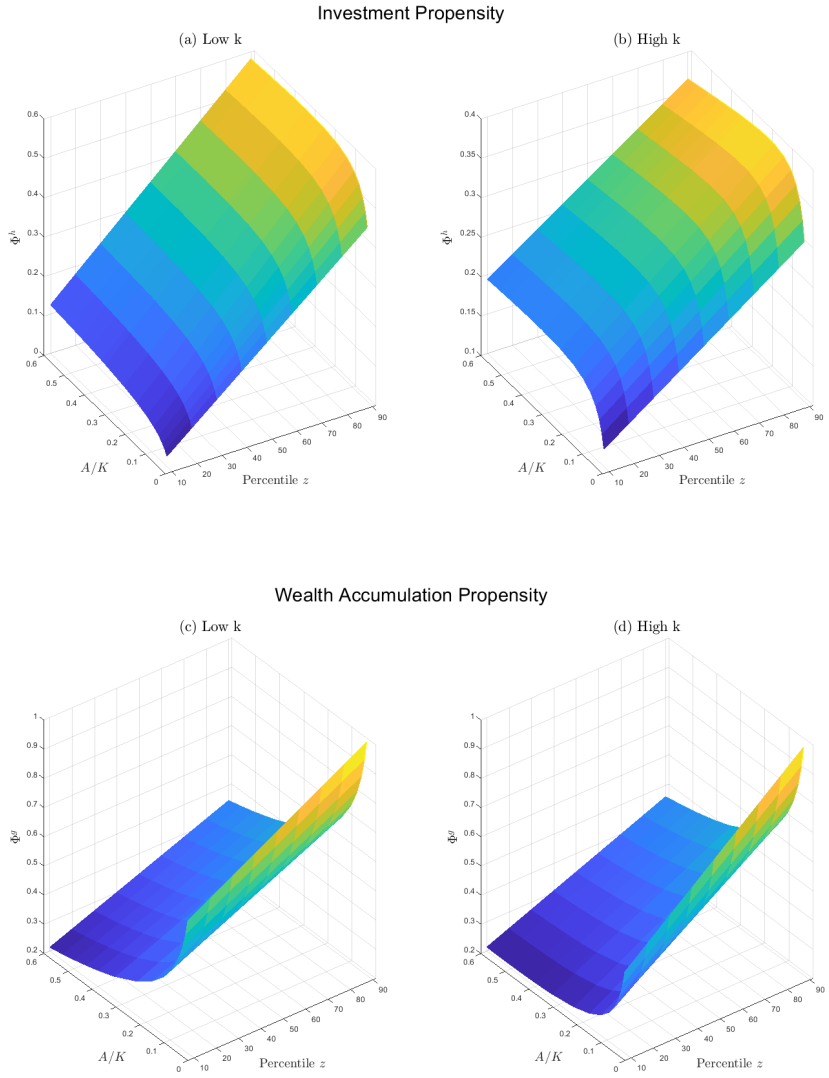

FIGURE 3: Nonlinear model: Investment and Wealth accumulation propensities to productivity

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.

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

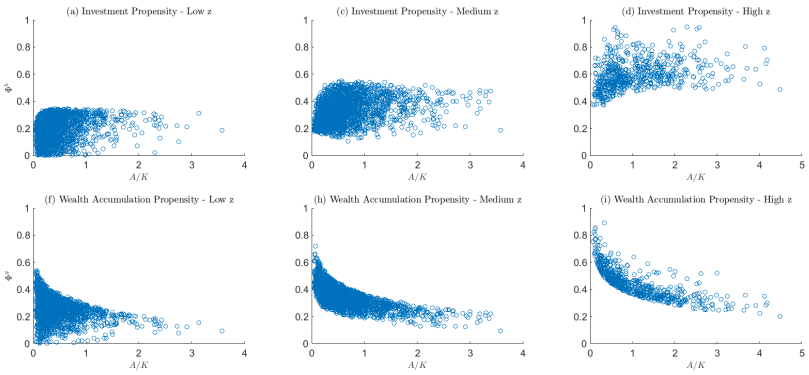

FIGURE 4: Investment and Wealth accumulation propensities in response to productivity

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.

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 across time 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 constant 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 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, among firms in the 90th percentile of the productivity distribution, convergence in the marginal product of capital between firms in the 10th and 90th wealth percentiles takes more than 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.

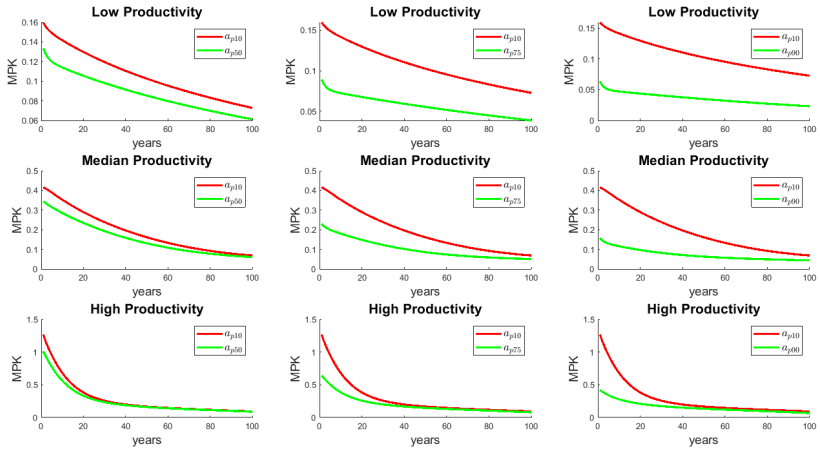

FIGURE 5: Convergence in the marginal product of capital across firms

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.

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 argue that standard approaches to recover productivity process from production function estimations fail under the presence of financial frictions that limit the firm's ability to hire inputs, as the auxiliary equations used to characterize input decisions do not hold. For instance, in the case of the OP estimator, the auxiliary investment equation does not account for wealth, a relevant variable for capital decisions in macro models with financial constraints such as Moll [2014] and Buera and Shin [2013b]. We argue that this renders a considerable bias in the estimation of the parameters of the firm's production function and, therefore, in the estimation of the characteristics of the productivity process. As an alternative, we extend the OP approach to account for financial frictions, introducing wealth and unobservable firm-specific shocks in the investment demand function. This flexible framework allows us 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 not accounting for financial friction biases the estimates of the production function and underestimates the dispersion in productivity. Additionally, we 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.

We use our setup to provide a detailed analysis of the firm´s policy functions, with a particular interest in understanding the mechanics of the self-financing channel. We show that, consistent with theoretical predictions in the presence of financial frictions, the marginal effect of productivity on investment is increasing in wealth and decreasing in capital. We also find a positive and significant marginal effect of productivity on wealth accumulation, stronger for more constrained firms, which provides support to the existence of an active self-financing channel. We also use our estimated empirical model to measure the power of self-financing on reducing misallocation by studying the convergence of MPKs of two firms with the same productivity but with different levels of financial frictions. We show that the MPKs of these firms converge over time, although the convergence is not fast and takes time. For instance, when we compare firms at the 10th percentile with firms at the 90th percentile of the wealth distribution, the MPK of poor firms is around three times the MPK of wealthy firms at the initial period, and it takes more than 40 years to see convergence in their MPKs. Still, half of the initial gap in their MPKs disappears after ten years.

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. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, 132(4): 1553–1592, 2017.

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

Manuel Arellano. Uncertainty, persistence, and heterogeneity: A panel data perspective. *Journal of the European Economic Association*, 12(5):1127–1153, 2014.

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

Manuel Arellano and Stéphane Bonhomme. Nonlinear panel data methods for dynamic heterogeneous agent models. *Annual Review of Economics*, 9:471–496, 2017.

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.

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

Stéphane Bonhomme. Discussion of "transparency in structural research" by isaiah andrews, matthew gentzkow, and jesse shapiro. *Journal of Business & Economic Statistics*, 38(4): 723–725, 2020.

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.

Sylvain Catherine, Thomas Chaney, Zongbo Huang, David Alexandre Sraer, and David Thesmar. Quantifying reduced-form evidence on collateral constraints. *forthcoming Journal of Finance*, 2018.

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, 2011a.

Jan De Loecker. Recovering markups from production data. *International Journal of Industrial Organization*, 29(3):350–355, 2011b.

Peter M DeMarzo and Michael J Fishman. Optimal long-term financial contracting. *The Review of Financial Studies*, 20(6):2079–2128, 2007.

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.

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 and Matthew Shum. Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics*, 171(1):32–44, 2012.

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.

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.

Søren Feodor Nielsen et al. The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000.

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.

Diego Restuccia and Richard Rogerson. Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic dynamics*, 11(4):707–720, 2008.

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

Zheng Song, Kjetil Storesletten, and Fabrizio Zilibotti. Growing like china. *American economic review*, 101(1):196–233, 2011.

Ludwig Straub. Consumption, savings, and the distribution of permanent income. *Unpublished manuscript, Harvard University*, 2019.

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

Appendix A.1: The bias in the OP estimator under financial frictions

Here we provide a more formal assessment of the types of biases emerging under OP in the context of financial frictions, and, by extension, on other methodologies relying on the proxy variable approach. As mentioned, Olley and Pakes [1996] propose a proxy variable approach to address the endogeneity problem that arises when estimating the parameters β_l and β_k from a value-added production function in logs, using data on value added y_{it}, capital k_{it} and labor l_{it}:

$$y_{it} = \beta_l l_{it} + \beta_k k_{it} + z_{it} + \varepsilon_{it}, \tag{26}$$

where ε_{it} is measurement error in value added. The main challenge in the estimation of β_l and β_k is that z_{it} is an unobservable variable for the econometrician which is potentially correlated with the observable 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 relies on using the investment policy function as an auxiliary equation to obtain information on the unobserved productivity z_{it}. For example, in the absence of constraints, we can see from the investment policy function (4) that: $i_{it} = h\left(z_{it}, k_{it}\right)$. Under the assumptions that z_{it} is the only unobserved variable for the econometrician in h (known as the scalar unobserved assumption) and that h is monotonic in z_{it}, we can invert the policy function to recover productivity as $z_{it} = h^{-1}\left(i_{it}, k_{it}\right)$ and construct valid moment conditions. For instance, we can rewrite (26) as:

$$y_{it} = \beta_l l_{it} + \beta_k k_{it} + h^{-1}\left(i_{it}, k_{it}\right) + \varepsilon_{it}. \tag{27}$$

Since ε_{it} is assumed to be uncorrelated with the inputs, OP propose to approximate $h^{-1}\left(i_{it}, k_{it}\right)$ with a high-order polynomial on investment and capital and run an OLS regression of y_{it} on l_{it}, k_{it}, and the non-linear, time-dependent polynomial $h^{-1}(i_{it}, k_{it})$ to estimate β_l and β_k. However, the OLS regression identifies β_L, but cannot separate β_k from the linear part of $h^{-1}(i_{it}, k_{it})$. Thus, in a second step, OP exploits the Markovian productivity process to estimate β_k by regressing the following model:

$$\hat{\pi}_t\left(i_{it}, k_{it}\right) = \beta_k k_{it} + \rho\hat{\pi}_{t-1}\left(i_{it-1}, k_{it-1}\right) - \rho\beta_k k_{it-1} + \eta_{it} + \hat{\varepsilon}_{it} \tag{28}$$

where $\hat{\pi}_t\left(i_{it}, k_{it}\right)$ denotes the estimated fraction of output explained by investment and capital in the first step, $\pi_t(i_{it}, k_{it}) = \beta_k k_{it} + h^{-1}(i_{it}, k_{it})$ [see e.g. Ackerberg et al., 2015] for details.

As discussed, observed differences in investment between firms in the data are intepreted as productivity differentials under OP. Hence, by controlling for investment in the production function, OP can eliminate the endogeneity problem and get consistent estimates of β_l and β_k. However, under borrowing constraints, differences in investment between firms might not only reflect differences in productivity, but also differences in borrowing capacity.

In the model with financial frictions described in Section 2, the investment function arising from (3) depends not only on productivity and initial capital, but also on net-worth, through its direct influence on the strength of financial frictions. When we invert the investment policy function in (4) we obtain $z_t = h^{-1}(i_{it}, k_{it}, a_{it})$, with $h_i^{-1} > 0$, $h_k^{-1} > 0$ and $h_a^{-1} \leq 0$. Therefore, for a given level of investment, firms facing more severe constraints due to low levels of net-worth are more productive. The intuition is direct: For a given productivity level, an unconstrained

firm will always invest more than a constrained firm. Therefore, for a given level of investment, it must be that the unconstrained firm is less productive. Replacing z_{it} in the production function we have:

$$y_t = \beta_l l_{it} + \beta_k k_{it} + h^{-1}(i_{it}, k_{it}, a_{it}) + \varepsilon_t \tag{29}$$

Hence, when implementing OP's first step, the term that captures the severity of the constraint due to net-worth would go to the error term of the OP regression in equation (27). Thus, if firms operate under borrowing constraints, the OP regression will render biased estimates of β_l and β_k due to the correlation of the regressors with the omitted variable a_{it}. Given that the OP estimation proceeds by two steps, we can analyze the biases separately. Let's focus first on the estimation of β_l. To see the sign of the correlation between l_{it} and a_{it} replace the expression for z_{it} obtained after inverting (4) in the FOC for labor (**??**):

$$l_{it} = c_l + \frac{1}{1-\beta_l}\left(\beta_k k_{it} + w + h^{-1}(i_{it}, k_{it}, a_{it})\right) \tag{30}$$

Therefore, after controlling for k_{it} and i_{it}, the correlation between l_{it} and the OP residual is positive.[22][23] Because OP cannot control for a fraction of productivity, which goes into the residual term when applying OP to (29), and as labor is increasing in productivity, the coefficient is biased upwards and $\hat{\beta}_l^{OP} > \beta_l$. To see the intuition suppose there are two firms with different productivities but that have the same level of capital and investment due to differences in collateral. OP will tend to equalize estimated productivity between the two, despite differences in output. The productive firm, that is more financially constrained, will choose to hire more workers, since frictions do not directly affect the labor market.[24] Since the OP estimator equalizes productivity between the two firms (given that they have the same investment), it will assign all the difference in output to differences in the amount of labor, leading to an overestimation of β_l.

In the case of the capital elasticity the relevant regression is the one implemented in the second stage (equation (28)). In the OP estimation, the function $\hat{\pi}_{t-1}()$ does not include a_{t-1} and this part of the function goes to the regression's error term. Given that $h_a \geq 0$ in equation (4) and that k_{it} is increasing in i_{it-1}, there is a positive correlation between the stock of capital used in production, k_{it}, and a_{it-1} - the level of collateral at the moment the investment decision is taken-. Therefore, $\tilde{h}_a < 0$ implies a negative correlation between the OP residual in equation (28) and k_{it}, leading to a downward bias: $\hat{\beta}_k^{OP} < \beta_k$. Intuitively, financial constraints generate differences in investment and capital for equally productive firms. The OP framework interprets the observed differences in investment as differences in unobserved productivity, and assigns part of the observed differences in output, which are due to capital, to variations in the productivity proxy, implying a lower estimated marginal effect of capital.

[22] For example, if $h^{-1}(i_{it}, k_{it}, a_{it})) = \tilde{h}_i i_{it} + \tilde{h}_k k_{it} + \tilde{h}_a a_{it}$ were linear, then the sign of the biases in $\hat{\beta}_l^{OP}$ and $\hat{\beta}_k^{OP}$ will depend on $\tilde{h}_a E\left[l_{it} a_{it} \mid i_{it}, k_{it}\right] > 0$ and $\tilde{h}_a E\left[k_{it} a_{it-1} \mid \hat{\pi}_{t-1}, k_{it-1}\right] < 0$.

[23] Note that l_{it} depends only on the constants c_l and w, and on state variables, so it is linearly dependent with the rest of the regressors in the production function regression [see Ackerberg et al., 2015]. In our empirical model we allow for the existence of an additional determinant of labor that can capture firm-specific iid shock in wages.

[24] Other models consider that financial constraints can affect the labor input as well. However, we should still expect an upward bias on β_l when the effect of frictions in the labor input are less severe. In our empirical model we will allow the labor input to also depend on the collateral constraint.

Appendix A.2: Proof of Identification

Here we discuss theorem 1 and show that β_k, β_l, $\varphi(z_{it-1})$, h_t, g_{t+1} are identified from data on y_{it}, k_{it}, l_{it}, i_{it}, a_{it} for $T \geq 4$ in a sequential way. First, we establish identification of the parameters of the production function. Second once β_k and β_l are identified, we show that the joint and marginal distributions of the productivity process are identify from the time series dependence structure of the net income process. Finally, once the conditional distribution of the productivity process given the firm net income process is identified, we show that h_t, g_{t+1} are identified.

Step 1: Production function Using the conditional independence assumption in *assumption 1* we can write the following conditional distribution of the observed variables $f(y_{it}, i_{it} \mid a_{it+1}, X_{it})$ in terms of some pieces of the model:

$$f(y_{it}, i_{it} \mid a_{it+1}, X_{it}) = \int f(y_{it} \mid z_{it}, i_{it}, a_{it+1}, X_{it}) f(i_{it} \mid z_{it}, a_{it+1}, X_{it}) f(z_{it} \mid a_{it+1}, X_{it}) dz_{it}, \tag{31}$$

where $f(y_{it} \mid z_{it}, k_{it}, l_{it})$ is the conditional distribution of the 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. Thus, we have that $f(y_{it} \mid z_{it}, i_{it}, a_{it+1}, X_{it}) = f(y_{it} \mid z_{it}, k_{it}, l_{it})$ and $f(i_{it} \mid z_{it}, a_{it+1}, X_{it}) = f(i_{it} \mid z_{it}, X_{it})$, and we can re-write (31) as

$$f(y_{it}, i_{it} \mid a_{it+1}, X_{it}) = \int f(y_{it} \mid z_{it}, k_{it}, l_{it}) f(i_{it} \mid z_{it}, X_{it}) f(z_{it} \mid a_{it+1}, X_{it}) dz_{it} \tag{32}$$

Now, the identification challenge is to recover the latent conditional density of the production function $f(y_{it} \mid z_{it}, k_{it}, l_{it})$ given the observed conditional density $f(y_{it}, i_{it} \mid a_{it+1}, X_{it})$. We notice that given assumption 1 and the structure of our dynamic model, our setup can be framed into the setup studied in Hu and Schennach [2008] and Hu et al. [2020]. Hence, Theorem 1 of Hu and Schennach [2008] can be applied to our setting to show that $f(y_{it} \mid z_{it}, k_{it}, l_{it})$ is identified from the data. Once we identify $f(y_{it} \mid z_{it}, k_{it}, l_{it})$ we can construct $E[y_{it} \mid z_{it} = 0, k_{it}, l_{it}] = \beta_l l_{it} + \beta_k k_{it}$ and identify β_k, β_l with a regression between $E[y_{it} \mid z_{it} = 0, k_{it}, l_{it}]$ and (l_{it}, k_{it}) as in theorem 1 in Hu et al. [2020].

Discussion: An important difference of our framework from Hu et al. [2020] is that our model with financial frictions provides a 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). Hence, we do not have to use the policy rule in $t+1$ to avoid collinearity between inputs and therefore k_{t+1} is not part of the covariates in X_t. This allow us to have a standard law of motions for capital as in Olley and Pakes [1996] and Ackerberg et al. [2015] without the need of an unobserved component affecting the law of motion of capital. The latter is particularly important in applied work because most of the cases the researcher do not have data on both capital and investment separately and use the perpetual inventory method to recover the capital series from investment or vice-versa.

We then have the following result, which is a direct application of theorem 1 in Hu and Schennach [2008] and theorem 1 in Hu et al. [2020].

Proposition 1. *Under the conditional independence assumption in assumption 1, the high-level conditions in condition (1) (i)-(iv), and the assumption that the ε_{it} has mean zero, β_l and β_k are identified from the observed density $f(y_{it}, i_{it} \mid a_{it+1}, X_{it})$*

To show how theorem 1 of Hu and Schennach [2008] can be applied to our setup, we will follow their paper and define the integral operators and show that it admits an eigenvalue-eigenvector decomposition that can be learned from data. Then, to build intuition and remark the importance of the wealth accumulation equation, we will make a connection with the IV setup discussed in the linear model. Lets define $L_{y;I|a,X}$ as the integral operator such that $L_{y;I|a,X} = \int f(y_t, i_t \mid a_{t+1}, X_t)\, p(a_{t+1} \mid X_t)\, da$ and $D_{I;z|X}$ is a "diagonal" matrix operator mapping the function $g(z \mid X)$ to the function to the function $f(i_t \mid z_t, X_t)\, g(z \mid X)$ for a given value of investment i. Analogously, $L_{y|z,k,l}$ and $L_{a|z,X}$ are the integral operators associated with the conditional densities $f(y_t \mid z_t, k_t, l_t)$ and $f(a_{t+1} \mid z_t, X_t)$, respectively. Equation (32) can be expressed in terms of integral operators:

$$L_{y;I|a,X} = L_{y|z,k,l} D_{I;z|X} L_{z|a,X} \tag{33}$$

Integrating both sides of (32) with respect to I:

$$L_{y|a,X} = L_{y|z,k,l} L_{z|a,X} \tag{34}$$

From (34), we can see that the identification of $L_{y|z,k,l} = L_{y|a,X} L_{z|a,X}^{-1}$, our object of interest, has the form of an IV regression where a_{it} is the instrument for the endogenous variable z_{it} after controlling for covariates in X_{it}. This type of IV approach is unfeasible because z_{it} is unobservable. However, replacing (34) in (33) we get:

$$L_{y;I|a,X} L_{y|a,X}^{-1} = L_{y|z,k,l} D_{I;z|X} L_{y|z,k,l}^{-1} \tag{35}$$

Note that the observed quantity $L_{y;I|a,X} L_{y|a,X}^{-1}$ in (35) admits an eigenvector-eigenvalue decomposition $L_{y|z,k,l} D_{I;z|X} L_{y|z,k,l}^{-1}$. Therefore, $L_{y|z,k,l}$ is identify as the eigenvector of $L_{y;I|a,X} L_{y|a,X}^{-1}$ of (35). If $L_{y|z,k,l}$ is identify, then $f(y_t \mid z_t, k_t, l_t)$ is identify.

Rank Condition (Injectivity) To identify $L_{y|z,k,l}$ from (35), the inverse of $L_{y|a,X}$ has to exist. Looking at (34) we can show that $L_{y|a,X}$ has an inverse if $L_{y|z,k,l}$ and $L_{a|z,X}$ are invertible. Given the linearity and additivity of the Cobb Douglas production function in logs, the assumption that the characteristic function of ε_{it} has no zeros on the real line will ensure injectivity (and invertibility) of $L_{y|z,k,l}$. The operator $L_{a|z,X}$ is injective (and invertible) if there is sufficient variation in the densities $f(a_{t+1} \mid z_t, a_t, k_t)$ for different values of z_{it}. The condition for $f(a_{t+1} \mid z_t, a_t, k_t)$ requires an statistical dependence between wealth accumulation a_{it+1} and productivity z_{it} conditioned on the observed state variables. This requirement can be met by the self-financing channel in equation (10) which implies a positive relationship between productivity and wealth accumulation for all constrained and non-constrained firms. In the IV terminology, the later is a relevance condition, that ensures that a_{it+1} is valid instrument for z_{it}, similar to the condition discussed in the linear case. Note that the expression $L_{y;I|a,X} L_{y|a,X}^{-1}$ in (35) looks like and IV regression using i_{it} as the proxy measure with error of z_{it} and a_{it+1} as the instrument for the proxy measure once we control for X_{it}.

Step 2: Productivity Process Given that our production function is Cobb-Douglas with Hicks neutral productivity, the net income process (after netting out the firm production function from the endogenous inputs) in (16) is linear and additive in the two unobserved components. The linearity and the stochastic assumptions on z_{it} and ε_{it} allow us to frame our model into the Nonlinear Markov model studied in Arellano et al. [2017]. Hence the identification of the productivity process follows the same arguments in Appendix A.1 and the supplemental material S.4 in Arellano et al. [2017].

Proposition 2. ***Identification of the Productivity Process***. *In a Cobb-Douglas production function with Markovian Hicks neutral productivity as in equations (??)-(8), if assumption (1) and condition (1)(v) hold and β_l and β_k are previously identified, then the joint distribution of $(\varepsilon_{i2,\cdots},\varepsilon_{iT-1})$ and the joint distribution of $(z_{i2,\cdots},z_{iT-1})$ are identified from i.i.d observations of $(\tilde{y}_{i1,\cdots},\tilde{y}_{iT})$ where $\tilde{y}_{it}$ is the net-income process for $T \geq 3$. With $T \geq 4$ the Markov probability $f_{z_t|z_{t-1}}(z_{it} \mid z_{it-1}))$ and $\phi = E[z_{it} \mid z_{it-1}]$ are identified.*

To provide some intuition on how identification works in our production function model we follow the discussion in Arellano [2014] and Arellano et al. [2017] and applied to our firm net income process. Following Arellano [2014] and Arellano et al. [2017], we first discuss the non-parametric identification of the distribution of ε_{it} for all t. Then using the linear structure of equation (16), by deconvolution, we can identify the distribution of z_{it}.

Given assumption 1 (i) and (ii) we can write the following nonlinear IV equation:

$$\tilde{y}_{it} = \psi(\tilde{y}_{it-1}) + \zeta_{it} \tag{36}$$

$$\tilde{y}_{it-1}\tilde{y}_{it} = \phi(\tilde{y}_{it-1}) + \upsilon_{it} \tag{37}$$

where $E[\zeta_{it} \mid \tilde{y}_{it-2}] = 0$ and $E[\upsilon_{it} \mid \tilde{y}_{it-2}] = 0$, and $\psi(.)$ and $\phi(.)$ are the solutions of an IV regression where $\tilde{y}_{it-2}$ is the instrument of $\tilde{y}_{it-1}$ in (36) and (37): $E[\tilde{y}_{it} - \psi(\tilde{y}_{it-1}) \mid \tilde{y}_{it-2}] = 0$ and $E[\tilde{y}_{it-1}\tilde{y}_{it} - \phi(\tilde{y}_{it-1}) \mid \tilde{y}_{it-2}] = 0$. The solutions $\psi(.)$ and $\phi(.)$ exist and are unique if both the conditional distributions of $\tilde{y}_{it} \mid \tilde{y}_{it-1}$ and $\tilde{y}_{it-1} \mid \tilde{y}_{it}$ are complete. This is a nonlinear relevance assumption that is ensured by the markovian condition of z_{it}. The distribution of $\tilde{y}_{it} \mid \tilde{y}_{it-1}$ is complete if $E[\phi(\tilde{y}_{it}) \mid \tilde{y}_{it-1}] = 0$ implies that $\phi(\tilde{y}_{it}) = 0$ for all ϕ in some space of functions (Newey and Powell 2003).

Identification of $\psi(.)$ and $\phi(.)$ relies on the autocorrelation structure in the data $(\tilde{y}_{i1,\cdots},\tilde{y}_{iT})$. Note that both $\psi(.)$ and $\phi(.)$ are data objects that can be estimated with data on $\{\tilde{y}_{it-2}, \tilde{y}_{it-1}, \tilde{y}_{it}\}$.

Given assumption 1 (parts (i) and (ii)), $\{\tilde{y}_{it-2}, \tilde{y}_{it-1}, \tilde{y}_{it}\}$ are independent given z_{it-1}. Provided that the conditional distribution of z_{it-1} given $\tilde{y}_{it-2}$ is complete we have:

$$E(\tilde{y}_{it} \mid z_{it-1}) = E(\psi(\tilde{y}_{it-1}) \mid z_{it-1}), \tag{38}$$

$$z_{it-1}E(\tilde{y}_{it} \mid z_{it-1}) = E(\phi(\tilde{y}_{it-1}) \mid z_{it-1}). \tag{39}$$

Equation (38) uses the condition that $E(\psi(\tilde{y}_{it-1}) \mid z_{it-1}, \tilde{y}_{it-2}) = E(\psi(\tilde{y}_{it-1}) \mid z_{it-1})$ and $E(\tilde{y}_{it} \mid z_{it-1}, \tilde{y}_{it-2}) = E(\tilde{y}_{it} \mid z_{it-1})$, while equation (39) uses also the condition that $E(\varepsilon_{it-1} \mid z_{it-1}) = 0$ and $E(\phi(\tilde{y}_{it-1}) \mid z_{it-1}, \tilde{y}_{it-2}) = E(\phi(\tilde{y}_{it-1}) \mid z_{it-1})$.

Since $\psi(.)$ and $\phi(.)$ are identified from (36) and (37) and data on $\{\tilde{y}_{it-2}, \tilde{y}_{it-1}, \tilde{y}_{it}\}$, we can use equation (38) and (39) to identify the distribution of ε_{it-1} for a fixed value of z:

$$E_{\varepsilon_{it-1}}[z\psi(z + \varepsilon_{it-1})] = E_{\varepsilon_{it-1}}[\phi(z + \varepsilon_{it-1})] \tag{40}$$

By deconvolution we can recover the density of ε_{it-1} from (40). Using the same argument we can recover the density of ε_{it} using $\{\tilde{y}_{it-1}, \tilde{y}_{it}, \tilde{y}_{it+1}\}$, for all $t = \{2, \ldots T-1\}$. By the separability of $\tilde{y}_{it} = z_{it} + \varepsilon_{it}$, once we identify the distribution of $(\varepsilon_{i2,\cdots}, \varepsilon_{iT-1})$, we can identify the distribution of $(z_{i2,\cdots}, z_{iT-1})$ given the observed data on $(\tilde{y}_{i2,\cdots}, \tilde{y}_{iT-1})$, assuming that the characteristic functions of ε_{it-1} do not vanish on the real line. Note that we need a panel with $T \geq 4$ for identifying the Markovian process of productivity . With $T \geq 4$ we can identify the join distribution of (z_{i2}, z_{i3}) which in turn identify the conditional distribution of z_{i3} given z_{i2}. If we assume that the productivity process is stationary we have identified the conditional distribution of z_{it} given z_{it-1} for all t.

Step 3: Policy Functions Once we have identified β_k, β_l and $f(z_1 \mid \tilde{y})$ we can identify $f(a_1, k_1 \mid z_1)$ and $f(a_{t+1} \mid z_t, a_t, k_t)$ and $f(k_{t+1} \mid z_t, a_t, k_t)$ for all $t > 1$ in a sequential way starting with period 1 in a similar way as in Arellano et al. [2017].

Proposition 3. ***Identification of the Policy Functions***. *In a Cobb-Douglas production function with Markovian Hicks neutral productivity as in equations (??)-(8), if assumption (1), (2) and condition (1)(v) hold and $f(z_1 \mid \tilde{y})$ is previously identified, then $f(a_1, k_1 \mid z_1)$, $f(a_{t+1} \mid z_t, a_t, k_t)$ and $f(k_{t+1} \mid z_t, a_t, k_t)$ are identified for all $t > 1$.*

Period 1

$$f(a_1, k_1 \mid \tilde{y}) = \int f(a_1, k_1 \mid z_1, \tilde{y}) f(z_1 \mid \tilde{y}) \, dz_1, \tag{41}$$

by *assumption 1*, $f(a_1, k_1 \mid z_1, \tilde{y}) = f(a_1, k_1 \mid z_1)$ equation (41) can be expressed as:

$$f(a_1, k_1 \mid \tilde{y}) = \int f(a_1, k_1 \mid z_1) f(z_1 \mid \tilde{y}) \, dz_1. \tag{42}$$

Equation (42) can be rewritten as the following moment restriction:

$$f(a_1, k_1 \mid \tilde{y}) = E\left[f(a_1, k_1 \mid z_1) \mid \tilde{y}_i = \tilde{y}\right] \tag{43}$$

where the expectation is taken with respect to the density of z_{i1} given $\tilde{y}_i$ and for a fixed values of a_1 and k_1. Provided that the distribution of $(z_{i1} \mid \tilde{y}_i)$, which is identified from the production function structure is complete, the unknown density $f(a_1, k_1 \mid z_1)$ is identified from (43). The density $f(a_1, k_1, z_1 \mid \tilde{y}) = f(a_1, k_1 \mid z_1) f(z_1 \mid \tilde{y})$ is also identified.

Using Bayesian rule, we can identify the following density:

$$f(z_1 \mid a_1, k_1, \tilde{y}) = \frac{f(a_1, k_1, z_1 \mid \tilde{y})}{f(a_1, k_1 \mid \tilde{y})}$$

Period 2 Like the analysis in period 1, we can use *assumption 1* to express $f(a_2 \mid a_1, k_1, \tilde{y})$ as:

$$f(a_2 \mid a_1, k_1, \tilde{y}) = \int f(a_2 \mid z_1, a_1, k_1) f(z_1 \mid a_1, k_1, \tilde{y}) \, dz_1 \tag{44}$$

where $f(a_2 \mid a_1, k_1, \tilde{y}) = f(a_2 \mid z_1, a_1, k_1)$. Equation (44) can be rewritten in terms of the following moment restriction:

$$f(a_2 \mid a_1, k_1, \tilde{y}) = E\left[f(a_2 \mid z_1, a_1, k_1) \mid a_{i1} = a_1, k_{i1} = k_1, y_i = y\right] \tag{45}$$

Equation (45) provides identification for $f(a_2 \mid z_1, a_1, k_1)$ as long as $f(z_1 \mid a_1, k_1, \tilde{y})$, which is identified in period 1, is complete in $\tilde{y}_i$). Note that $f(a_2, z_1 \mid a_1, k_1, \tilde{y})$ is also identified. Similarly, $f(k_2 \mid z_1, a_1, k_1)$ (and consequently $f(k_2, z_1 \mid a_1, k_1, \tilde{y})$) is identified from

$$f(k_2 \mid a_1, k_1, \tilde{y}) = E\left[f(k_2 \mid z_1, a_1, k_1) \mid a_{i1} = a_1, k_{i1} = k_1, y_i = y\right] \tag{46}$$

Given *assumption* 1 $f(a_2, k_2 \mid z_1, a_1, k_1) = f(k_2 \mid z_1, a_1, k_1) f(a_2 \mid z_1, a_1, k_1)$. Using Bayesian rule and *assumption* 1 we recover $f(z_1 \mid a_2, k_2, a_1, k_1)$ from:

$$f(a_2, k_2 \mid z_1, a_1, k_1) = \frac{f(z_1 \mid a_2, k_2, a_1, k_1) f(a_2, k_2 \mid a_1, k_1)}{f(z_1 \mid a_1, k_1)}$$

Given that $f(z_1 \mid a_2, k_2, a_1, k_1)$ is identified from above, $f(z_2 \mid z_1)$ is identified from the net-income process, and given *assumption* 1 we can identified: $f(z_2, z_1 \mid a_2, k_2, a_1, k_1) = f(z_1 \mid a_2, k_2, a_1, k_1) f(z_2 \mid z_1)$, which in turns allow us to identified $f(z_2 \mid a_2, k_2, a_1, k_1, \tilde{y})$ using Bayesian rule and given *assumption* 1:

$$f(z_2 \mid a_2, k_2, a_1, k_1, \tilde{y}) = \int \frac{f(\tilde{y} \mid z_2, z_1) f(z_2, z_1 \mid a_2, k_2, a_1, k_1)}{f(\tilde{y} \mid a_2, k_2, a_1, k_1)} dz_1$$

Period 3 Using *assumption* 1 and *assumption* 2 we have:

$$f(a_3 \mid a_2, k_2, a_1, k_1, \tilde{y}) = \int f(a_3 \mid z_2, a_2, k_2) f(z_2 \mid a_2, k_2, a_1, k_1, \tilde{y}) dz_1 \tag{47}$$

Provided that $f(z_2 \mid a_2, k_2, a_1, k_1, \tilde{y})$, which is identified from above, is complete in $(a_{i1}, k_{i1}, \tilde{y})$, $f(a_3 \mid z_2, a_2, k_2)$ is identified from 47. Analogously, $f(k_3 \mid z_2, a_2, k_2)$ is identified from:

$$f(k_3 \mid a_2, k_2, a_1, k_1, \tilde{y}) = \int f(k_3 \mid z_2, a_2, k_2) f(z_2 \mid a_2, k_2, a_1, k_1, \tilde{y}) dz_1$$

Given assumption *assumption* 2 $f(a_{t+1} \mid z_t, a_t, k_t)$ and $f(k_{t+1} \mid z_t, a_t, k_t)$ are identified provided that for all $t > 1$, the distribution of $(z_{it} \mid a_i^t, k_i^t, \tilde{y}_i)$ is complete in $\left(a_i^{t-1}, k_i^{t-1}, \tilde{y}_i\right)$.

Appendix A.3: Stochastic EM Estimation Algorithm (SEM)

We adapt a stochastic EM algorithm to our production function framework to estimate the nonlinear model with latent variables in Section 5.2. Let $X_i^T = \left(y_i^T, k_i^T, l_i^T, a_i^T, \right)$ and z_i^T the history of observables and productivity for firm i, respectively. Given *assumption 1*, the full model in Section 5.2 implies the following integrated moment restrictions:

$$E\left(\int \begin{bmatrix} \sum_{t=2}^{T}\left(a_{it+1} - \sum_{k=1}^{K}\alpha_k^g \phi_k^g\left(z_{it}, k_{it}, a_{it}, \delta_t^g\right)\right)^2 \\ \sum_{t=1}^{T}\left(i_{it} - \sum_{k=1}^{K}\alpha_k^h \phi_k^h\left(z_{it}, k_{it}, a_{it}, \delta_t^h\right)\right)^2 \\ \sum_{t=1}^{T}\left(l_{it} - \sum_{k=1}^{K}\alpha_k^n \phi_k^n\left(z_{it}, k_{it}, a_{it}\right)\right)^2 \\ \sum_{t=1}^{T}\left(y_{it} - \beta_l l_{it} - \beta_k k_{it} - z_{it}\right)^2 \\ \sum_{t=1}^{T}\left(z_{it} - \sum_{k=1}^{K}\alpha_k^{\varphi} \phi_k^{\varphi}\left(z_{it-1}\right)\right)^2 \\ \left(a_{i1} - \sum_{k=1}^{K}\alpha_k^{g1} \phi_k^{g}\left(z_{i1}\right)\right)^2 \end{bmatrix} f\left(z_i^T \mid X_i^T, \theta\right) dz\right) \tag{48}$$

where $f\left(z_i^T \mid X_i^T, \theta\right)$ is the posterior density of the vector z_i^T given the data. The vector $\theta = \left[\theta^y, \theta^h, \theta^g, \theta^{g1}, \theta^n, \theta^{\varphi}\right]$ contains all the parameters of the model in (**??**), $\theta^y = [\beta_k, \beta_l, \sigma_\epsilon]$, $\theta^h = \left[\alpha_1^h \ldots \alpha_K^h, \sigma_v\right]$, $\theta^g = [\alpha_1^g \ldots \alpha_K^g, \sigma_w]$, $\theta^{\varphi} = [\alpha_1^{\varphi} \ldots \alpha_K^{\varphi}, \sigma_\eta]$. Note that (48) are the integrated version of the unfeasible OLS regressions of the equations in (**??**). The OLS are unfeasible because we do not observe z_{it}.

The stochastic EM algorithm possesses computational advantages with respect to a maximum likelihood estimation of the model in Section 5.2, given that each policy function depends on a considerable number of parameters. Therefore, rather than maximize the likelihood with respect to a lot of parameters, our stochastic EM estimator iterates between simulating draws from the posterior distribution of latent productivity given the data $f\left(z_i^T \mid X_i^T, \theta\right)$ and OLS estimations of the parameters in θ.[25]

The two following steps describe our procedure. Starting with a parameter vector θ^0, we iterate the following two steps on $s = 0, 1, 2, \ldots$ until convergence of the θ^s process to a stationary distribution:

1. *Stochastic E-step:* For each firm i, draw $\left\{z_{i1}^{(m)} \ldots z_T^{(m)}\right\}$ M realizations of z_i^T from $f\left(z_i^T \mid X_i^T, \theta\right)$. Using *assumptions 1 and 2* we can express the posterior distribution of z_{it} as a

[25] For instance, if we specify our nonlinear functions as third-order polynomials, the model in Section 5.2 would contain more than 200 parameters to be estimated. If in addition we want to estimate policy functions that include firm fixed effects that maximum likelihood estimation would be computationally infeasible.

function of the likelihoods of the equations in (**??**).

$$
\begin{aligned}
f\left(z_i^T \mid X_i^T, \theta\right) = & \prod_{t=1}^{T} f\left(y_{it} \mid k_{it}, l_{it}, z_{it}, \theta^y\right) \times \\
& \prod_{t=1}^{T} f\left(i_{it} \mid k_{it}, z_{it}, a_{it}, \theta^h\right) f\left(l_{it} \mid k_{it}, z_{it}, a_{it}, \theta^n\right) \times \\
& \prod_{t=2}^{T} f\left(a_{it} \mid z_{it}, k_{it}, a_{it}, \theta^g\right) f\left(a_{i1} \mid z_{i1}, \theta^{g1}\right) \times \\
& \prod_{t=1}^{T} f\left(z_{it} \mid z_{it-1}, \theta^{\varphi}\right) f\left(z_{i1}\right)
\end{aligned}
$$

where $f\left(y_{it} \mid k_{it}, l_{it}, z_{it}, \theta^y\right)$ is the likelihood of the production function, $f\left(i_{it} \mid k_{it}, z_{it}, a_{it}, \theta^h\right)$ is the likelihood of the investment policy rule, $f\left(a_{it+1} \mid z_{it}, k_{it}, a_{it}, \theta^g\right)$ is the likelihood of the wealth policy rule and $f\left(z_{it} \mid z_{it-1}, \theta^{\varphi}\right)$ is the likelihood of the productivity process. To simulate $f\left(z_i^T \mid X_i^T, \theta\right)$, we use a random-walk Metropolis-Hastings sampler, targeting an acceptance rate of 0.3.

2. *M-step:* compute the integrated-OLS estimator of the parameters:

$$
\begin{cases}
\sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \left(y_{it} - \beta_l l_{it} - \beta_k k_{it} - z_{it}^{(m)}\right)^2 \\
\sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \left(i_{it} - \sum_{k=1}^{K} \alpha_k^h \phi_k^h \left(z_{it}^{(m)}, k_{it}, a_{it}, \delta_t^h\right)\right)^2 \\
\sum_{i=1}^{N} \sum_{t=2}^{T} \sum_{m=1}^{M} \left(a_{it+1} - \sum_{k=1}^{K} \alpha_k^g \phi_k^g \left(z_{it}^{(m)}, k_{it}, a_{it}, \delta_t^g\right)\right)^2 \\
\sum_{i=1}^{N} \sum_{t=2}^{T} \sum_{m=1}^{M} \left(z_{it}^{(m)} - \sum_{k=1}^{K} \alpha_k^{\varphi} \phi_k^{\varphi} \left(z_{it-1}^{(m)}\right)\right)^2 \\
\sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \left(l_{it} - \sum_{k=1}^{K} \alpha_k^n \phi_k^n \left(z_{it}^{(m)}, k_{it}, a_{it}\right)\right)^2
\end{cases}
\tag{49}
$$

In practice, we stop the iterative procedure after S=500 iterations and check the convergence of the estimates. In each iteration of the chain, we simulate 100 draws from step 1 (i.e., M=100). We start the algorithm from different initial values (OP, LP, or Proxy-IV) and get similar results. The statistical properties of a similar stochastic algorithm have been studied in Nielsen et al. [2000] in a likelihood context and in Arellano and Bonhomme [2016] in a GMM context where the M-step consists of quantile-based regressions. Arellano and Bonhomme [2016] show that the estimates of the stochastic EM algorithm for parametric models (where R does not grow with the sample size) are asymptotically normally distributed as M and N tend to infinity (for fixed R) with an asymptotic variance that is the asymptotic variance of the method-of-moments estimator of the integrated moment restrictions. Our M-step, which consists of a set of OLS regressions, can be framed in the GMM framework studied in Arellano and Bonhomme [2016]. Therefore, θ has the following distribution as N and M go to infinity:

$$
\sqrt{N\left(\hat{\theta} - \theta\right)} \xrightarrow{d} N\left(0, \Sigma\right)
$$

where Σ is the asymptotic variance of the GMM estimator of (48).

Appendix A.4: Estimations using Simulated Data

We use an extended version of the stylized model presented in Section 2 to generate data that is consistent with the theoretical framework that explicitly accounts for collateral constraints. We use this data to provides a validation of our proposed empirical specification.

The spirit of the model and its theoretical implications are very similar to those of the model presented in Section 2, although we generalize it in two dimensions. First, we no longer impose linearity in preferences and assume a CRRA utility function with risk aversion coefficient σ.[26] Second, we introduce adjustment costs to capital. We choose a standard quadratic function with a parameter η determining its size.[27] Note that the introduction of adjustment costs implies that capital is a state variable, as in our empirical estimations.

We assume a specific functional form for the general collateral constraint described in Section 2:

$$\kappa(A_{it}, Z_{it}) = (\lambda + \lambda_z(z_{it} - \bar{z}))A_{it}$$

where λ and λ_z are constants, z_{it} is the log of Z_{it} and $\bar{z}$ its mean, and we impose $\lambda+\lambda_z(\min(z_{it})-\bar{z}) \geq 1$. Thus, for a given level of collateral, the capital to assets ratio is strictly increasing in productivity.

In line with the estimates in Section 6.2, we set $\beta_k = 0.43$ and $\beta_l = 0.44$ in the calibrated model. This implies a span of control parameter of 0.87. In the case of the productivity process, we impose a linear Markov process, $z_{t+1} = \rho z_t + \mu_t$ and, consistent with our estimations, set $\rho = 0.82$ and $\sigma_\mu = 0.42$. We calibrate three key parameters to match certain moments of the sample. These parameters define the strength of the collateral constraint (λ and λ_z) and determine the relevance of adjustment costs η. The moments we use to calibrate them are the mean capital-to-output ratio, which is 1.69, the net assets-to-output ratio, which is 0.89, and the correlation between productivity and the net assets-to-capital ratio is 0.3. For the rest of the parameters, we use standard values: discount factor $\beta = 0.8$, risk aversion coefficient $\sigma = 0.2$, depreciation rate $\delta = 0.1$, and interest rate $r = 4\%$.

We use the calibrated model to generate simulated data and use that data to replicate the empirical estimations of the previous section.[28]

Model simulations can also be used to explore how the biases of the production function estimates vary with the intensity of financial frictions, i.e., with different values of the parameters governing the collateral constraint, λ y λ_z. For instance, when λ decreases from 2.5 -the value found in the calibrated version of the model- to 2, the OP bias in β_l grows from 0.07 to 0.11, and the OP bias in β_k goes from -0.04 to -0.06. When we make collateral constraints more severe through a change in λ_z, the effects on OP are similar. When λ_z goes from 0.5 -the value found in the calibration- to 0, the biases for β_l y β_k increase to 0.17 and -0.07, respectively.

[26] We remove the convex function $g(\cdot)$ included in Section 2, as it is no longer needed to have an interior solution.

[27] Specifically we use $\eta(I_{it}/K_{it})^2 K_{it}$

[28] Following Ackerberg et al. [2015] we introduce iid shocks to wages. This generates extra variability on labor that is not due to variation in the state variables, allowing us to identify β_l in the first stage.

Appendix A.5: Confidence Intervals

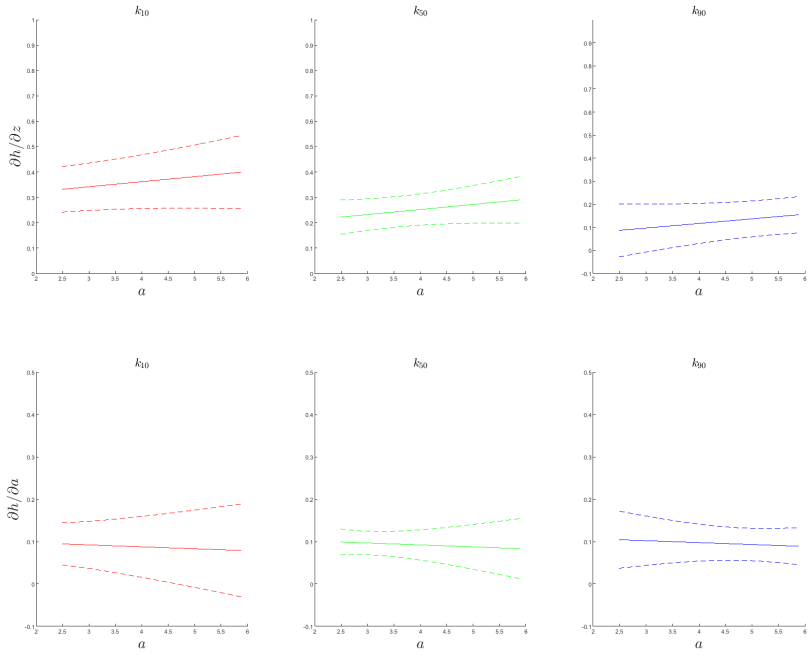

FIGURE 6: Confidence Intervals: Marginal effect of productivity and wealth on investment

Notes: The top panel exhibits the estimated derivative effect of productivity in the investment policy function and its 95% confidence intervals using the proxy-IV approach. The figure displays how the effect changes along different values of the stock of wealth and is evaluated at three different levels of stock of capital (10th, 50th, and 90th percentile of the capital distribution). The bottom panel of the figure exhibits the estimated derivative effect of the stock of wealth (previous wealth) in the investment policy function and its 95% confidence intervals using the proxy-IV approach. The figure displays how the effect changes along different values of the stock of wealth and is evaluated at three different levels of stock of capital (10th, 50th, and 90th percentile of the capital distribution).

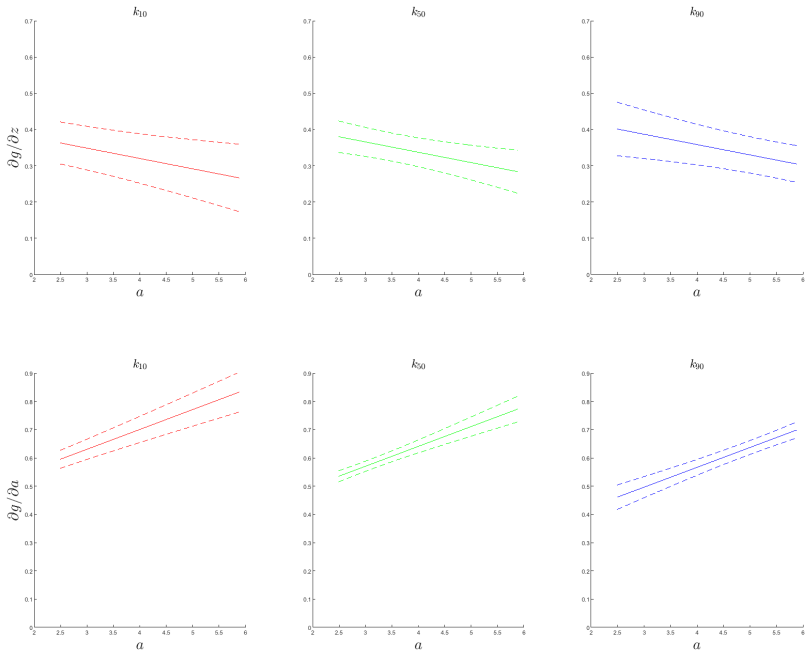

FIGURE 7: Confidence Intervals: Marginal effect of productivity and wealth on wealth accumulation

Notes: The top panel exhibits the estimated derivative effect of productivity in the wealth policy function and its 95% confidence intervals using the proxy-IV approach. The figure displays how the effect changes along different values of the stock of wealth and is evaluated at three different level of stock of capital (10th, 50th and 90th percentile of the capital distribution). The bottom panel of the figure exhibits the estimated derivative effect of the stock of wealth (previous wealth) in the wealth policy function and its 95% confidence intervals using the proxy-IV approach. The figure displays how the effect changes along different values of the stock of wealth and is evaluated at three different level of stock of capital (10th, 50th and 90th percentile of the capital distribution).

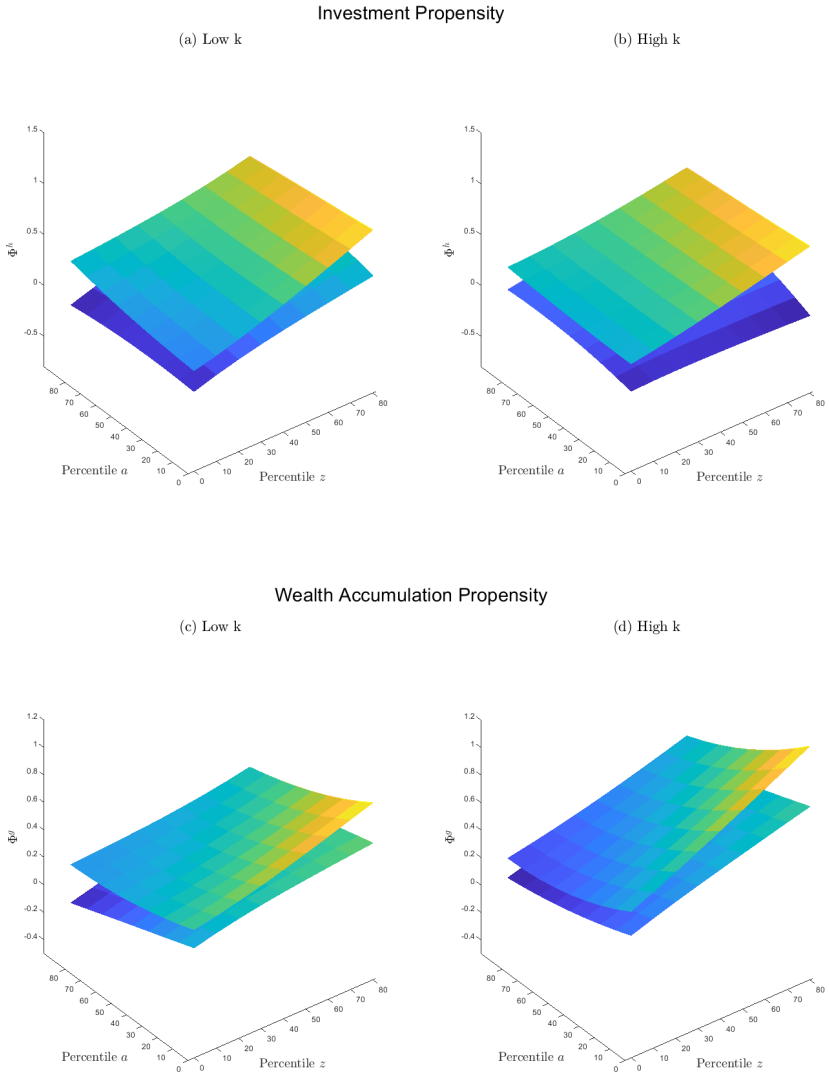

FIGURE 8: Confidence Intervals: Investment and Wealth Accumulation propensities

Notes: The figure exhibits the 95% confidence intervals of the estimated derivative effect of productivity in the investment and wealth policy functions using the SEM method.