Equilibrium Value and Profitability Premiums

Hengjie Ai, Jun E. Li and Jincheng Tong *

August 4, 2022

Abstract

Standard production-based asset pricing models cannot simultaneously explain the value premium and the gross profitability premium. Empirically, we show that value and profitability sorted portfolios differ in the persistence of productivity. We develop a general equilibrium model where firm-level productivity has a two-factor structure with different persistence. We demonstrate that with capital adjustment costs and variable capital utilization, our model can simultaneously account for both the gross profitability premium and the value premium.

JEL Code: E2, E3, G12

*Hengjie Ai (hengjie.ai@wisc.edu) is affiliated with the University of Wisconsin-Madison, Jun E. Li (junli@saif.sjtu.edu.cn) is affiliated with Shanghai Advanced Institute of Finance, Shanghai Jiao Tong University, and Jincheng Tong (jincheng.tong@rotman.utoronto.ca) is at the University of Toronto. We thank Winston Dou, Lukas Schmid, and participants at WFA and MFA for helpful comments.

1 Introduction

Historically, firms with higher profitability ratios, defined as the ratio of gross profit to total assets, earn higher average returns than those with lower profitability ratios. This is often referred to as the gross profitability premium (Novy-Marx (2013)). At the same time, stocks with higher ratios of the book value of assets relative to their market value, or value firms, earn a higher average return than those with higher market-to-book ratios (growth firms). The difference in the average return of value and growth firms is called the value premium.

The coexistence of the value and profitability premiums presents a challenge to the literature on Q-theory and production-based asset pricing models, for example, Zhang (2005). In these models, productivity is typically specified as an AR(1) process and is the only source of exogenous variation that determines the cross-sectional heterogeneity in firms' market-to-book and profitability ratios. High productivity firms are high profitability ratio firms but are simultaneously also growth firms: they have a higher market value of assets relative to their book value and, therefore, higher market-to-book ratios. As a result, traditional production-based asset pricing models cannot separate the value factor and the profitability factor. By construction, a profitability premium must imply a value discount. That is, these models can generate either a value premium or a profitability premium, but not both. In this paper, we develop a general equilibrium model where Q-theory holds and firm-level productivity has a two-factor structure that differs in persistence. We show that our model can not only distinguish the profitability factor from the value factor but also account for the coexistence of the value and profitability premiums.

To distinguish profitability from value empirically, we first provide empirical evidence for the two-factor structure in firm productivity. We show that consistent with the standard production-based asset pricing literature, both high profitability firms and growth firms have higher investment rates, higher cash flow growth rates, and faster sales growth rates at the time of portfolio formation compared to low profitability and value firms, respectively. However, several differences between the profitability and value sorted portfolios suggest that market-to-book sorting differentiates firms in terms of the permanent component of productivity shocks, while portfolios sorted on profitability differ in terms of the transitory component of their productivity.

First, the differences in investment rate, cash flow growth rate, and sales growth rate are significantly higher for market-to-book sorted portfolios than profitability sorted portfolios. For example, the median investment rate is about 7% for firms in the value portfolio, and is more than twice as large, about 15% for the growth portfolio. In contrast, the median investment rates of low and high profitability sorted portfolios are 8% and 12%, respectively.

Second, the differences in investment rates, cash flow growth rates, and sales growth rates are much more persistent for market-to-book sorted portfolios than profitability sorted portfolios. Empirically, we show that the spreads in cash flow and sales growth rates for market-to-book sorted portfolios remain significant five years after portfolio formation, while the difference in the above measures vanishes one year after portfolio formation for gross profitability sorted portfolios.

In addition, we formally estimate a two-factor model of firm productivity and decompose it into a permanent and a transitory component. We demonstrate that market-to-book sorted portfolios differ more significantly in terms of the permanent component of productivity, while gross profitability sorted portfolios have a higher spread in the transitory component of productivity and a much smaller spread in the permanent component.

Motivated by the above empirical evidence, we develop a general equilibrium model where firm productivity has a permanent and a transitory component. The transitory component has a large standard deviation in one-period innovations but a very small persistence. In contrast, the permanent component has a small conditional standard deviation but follows a random walk. This specification allows our model to distinguish the profitability factor from the value factor. A positive shock to the transitory component of productivity raises the current-period profitability of the firm but, because of its lack of persistence, has a quantitatively small impact on the market-to-book ratio. A positive shock to the permanent component of productivity, on the other hand, has a small impact on the current-period profitability. The impact is small because the size of the permanent shock is typically small, but it significantly affects the market value of the firm, as the shock is expected to last into the future. As a result, from a quantitative perspective, consistent with the data, the market-to-book sorting mainly differentiates firms along the permanent component of productivity shocks, while gross profitability sorting selects firms based primarily on the transitory component of their productivity.

We incorporate two sources of aggregate shocks into our model: total factor productivity shocks and shocks to the marginal cost of capital utilization. Our model generates a value premium as in Zhang (2005). In our model, firms that experience low permanent components of the productivity shocks are value firms. The fixed operating cost and capital adjustment cost provide a form of operating leverage, making these firms riskier relative to growth firms, where the adjustment cost is a smaller fraction of firm cash flow.

The key for our model to account for the profitability premium is variable capital utilization: firms can choose a higher rate of capital utilization at the expense of faster depreciation. In the presence of capital utilization costs, high profitability firms and growth firms respond very differently. High profitability firms, knowing that shocks are transitory, respond by increasing the rate of capital utilization and not investment. In contrast, growth firms are those who experience positive shocks in the permanent component of productivity. Anticipating that the impact of these shocks will persist into the future, they respond by increasing investment and lowering capital utilization. In our model, the cost of capital utilization is a form of operating leverage and makes high profitability firms and value firms riskier.

In summary, value and profitability sorting are determined by the permanent and transitory components of idiosyncratic productivity shocks in our model. At the aggregate level, due to firms' optimal choices of investment and capital utilization, value firms are endogenously more exposed to aggregate productivity shocks, while high profitability firms are more sensitive to shocks to the marginal cost of capital utilization.

Our model is set in a general equilibrium framework, while standard production-based asset pricing models of the cross-section of expected returns typically assume an exogenous pricing kernel. In a general equilibrium setup, the cross-sectional distribution of firm types becomes a relevant state variable that determines the properties of the pricing kernel. Because endogenous investment decisions depend not only on productivity shocks but also on the property of the pricing kernel, partial equilibrium models cannot account for this equilibrium feedback mechanism.

The above setting allows us to identify a general equilibrium feedback mechanism where the composition of firms affects the volatility of the stochastic discount factor and predicts returns in equilibrium. In the cross-section, value firms and high profitability firms have higher expected returns than growth firms and low profitability firms, respectively. In our model, this mechanism also affects the expected returns in aggregate time series. We provide empirical evidence which is consistent this prediction of our model. We show that the relative market value of value versus growth firms and the relative market value of high versus low profitability firms can positively predict future returns.

We calibrate our model to match aggregate moments and examine its ability to explain the cross-section. Our model closely matches standard macroeconomic quantity dynamics. It also delivers realistic financial moments, including a high equity premium and a low and smooth risk-free interest rate. We report four quantitative successes of our model. First, the model generates a realistic heterogeneity in firm investment policies for value/growth and high/low profitability sorted portfolios. In our model, the heterogeneity in the permanent component of firm productivity leads to a significant spread in market-to-book ratios. As a result, as in the data, market-to-book sorted portfolios exhibit a large dispersion in investment rates, which depends primarily on the permanent component of productivity. In contrast, profitability sorted portfolios differ mostly in the transitory component of productivity and do not have a significant difference in investment rates.

Second, variable capital utilization provides a resolution for the profitability premium. Utilizing capital more intensively raises a firm's current profit, leading to faster capital depreciation. Under the optimal policy, firms with higher transitory productivity shocks tend to choose higher utilization rates than firms that experience adverse transitory shocks. Faster capital depreciation due to higher utilization increases the exposure of profitable firms to the aggregate utilization cost shock. As a result, high profitability firms are riskier and require a higher expected return in equilibrium. We show that this utilization channel gives rise to a significant profitability premium in our calibration.

Third, our model retains the success of standard Q-theory-based models in terms of generating a realistic value premium. In our model, the adjustment cost constitutes a form of operating leverage: value firms are typically firms with low permanent productivity. They disinvest and are heavily affected by adjustment costs. Firms with a high permanent component of productivity have a higher market-to-book ratio. Their investment becomes a hedge against aggregate productivity shocks and makes their equity claims less risky.

Fourth, we show that consistent with empirical evidence, double sorting on both the market-to-book ratio and the profitability ratio makes both the value premium and the profitability premium more pronounced. In our model, the true state variables that predict firm-level returns are the permanent and transitory components of productivity. Both market-to-book ratios and profitability ratios are noisy measures of these fundamental state variables. Double sorted portfolios provide a more accurate separation of firm types and exhibit a stronger pattern in expected returns.

Literature review

This paper builds on the large literature on Q-theory-based asset pricing models and the cross-section of equity returns. Zhang (2005) develops a Q-theory-based asset pricing model to explain the value premium. Belo, Lin, and Bazdresch (2014) and İmrohoroğlu and Tüzel (2014) use a similar approach to study the impact of other firm characteristics, such as labor hiring and firm-level productivity, on the cross-section of equity returns. Donangelo, Gourio, Kehrig, and Palacios (2019) examine the relation between firm-level labor share and the cross-section of equity returns. Herskovic, Kind, and Kung (2018) study the low-frequency comovement between value and size premiums using a standard Q-theory-based asset pricing model. Kuehn, Simutin, and Wang (2017) explore the connection between firms' exposures to labor market conditions and stock returns. Kuehn and Schmid (2014) extend the Q-theory asset pricing model to study the pricing of corporate bonds. Eisfeldt and Papanikolaou (2013) incorporate organizational capital into the Q-theory asset pricing paradigm and investigate its impact on firms' stock returns. As pointed out by Novy-Marx (2013), in the above Q-theory-based models, the capital adjustment costs serve as operating leverage and generate a value premium. However, because high productivity firms are simultaneously high market-to-book and high profitability ratio firms, these models would also imply a negative profitability premium. In addition, all of the above models are partial equilibrium models that take the pricing kernel as given, while our model is a general equilibrium one, where the pricing kernel and firms' risk exposure are both endogenously determined in equilibrium.[1]

Several recent papers also study the coexistence of value and profitability premiums. Dou et al. (2021) develop a model with dynamic strategic competition to explain the profitability premium. Both Dou et al. (2020) and Kogan et al. (2020) develop equilibrium models where the value premium and the gross profitability premium coexist. Dou et al. (2020) focus on how competitiveness affects the heterogeneity of returns across industries but do not explain the within-industry value and profitability premium. Kogan et al. (2020) emphasize the importance of variable input as an operating hedge. The economic mechanism in our model is quite different from both of the above papers. In addition, our general equilibrium setup allows us to study the feedback mechanism between the cross-section distribution of firms and the properties of the stochastic discount factor.

Our paper also builds on the previous literature that develops tractable models to study the cross-section of expected returns in general equilibrium. Gomes et al. (2003) develop an analytical tractable framework where the cross-section of firm characteristics has a stationary distribution to study the cross-section of expected returns. Ai and Kiku (2013) develop a model with balanced growth and long-run risk and use the technique from the firm dynamics literature (Luttmer (2007)) to solve the aggregate problem. Several works in the production based asset pricing literature use the Krusell and Smith (1998) method to numerically solve their models.[2] Our aggregation method is different from all of the above. We take advantage of the homogeneity in firms' decision problems and summarize the cross-sectional distribution as a low-dimensional distribution.[3] We exploit the analytical expression for the law of motion of this distribution to solve the model using the local projection method. Our methodology could potentially be applied to a larger class of models.

Our paper is also related to the literature that emphasizes the importance of variable

[1]A different strand of production-based asset pricing models explore the implications of investment-specific shocks (IST), for example Papanikolaou (2011), Kogan and Papanikolaou (2013), Kogan and Papanikolaou (2014), and Li (2018). Kogan and Papanikolaou (2012) provide an excellent survey of this literature.

[2]Recent works that approximate firm distribution using distributional moments include Favilukis and Lin (2015), Chen (2018), Bai et al. (2019), and Favilukis et al. (2020).

[3]Other papers that use the same technique to reduce the dimensionality of the firm distribution in general equilibrium setups include Ai and Bhandari (2021), and Tong and Ying (2020).

capital utilization in understanding business cycle dynamics and asset market valuations. Greenwood et al. (1988) and Jaimovich and Rebelo (2009) show that variable utilization is an important ingredient accounting for the positive correlation between consumption and investment. Garlappi and Song (2017) show that variable utilization plays an important role in understanding the pricing of investment-specific shocks. Grigoris and Segal (2020) find that low capital utilization is associated with high equity returns, and such a pattern is consistent with a standard Q-theory model with variable capital utilization.

The rest of the paper is organized as follows. We first present the empirical evidence of a two-factor structure of firm level productivity in Section 2. Section 3 develops a general equilibrium model that embeds a two-factor structure into firm productivity shocks. We study the asset pricing implications of the model in Section 4 and present our quantitative results in Section 5. Section 6 concludes.

2 Empirical evidence

In this section, we replicate the profitability premium and its relationship with the value premium documented by Novy-Marx (2013) and highlight the challenge that they pose for standard production-based asset pricing models. We also present empirical evidence for a two-factor structure of firm-level productivity, which motivates our theoretical development. Detailed data construction can be found in Appendix D.1.

2.1 Value and profitability premiums

Following Novy-Marx (2013), we define gross profitability as revenue minus cost of goods sold divided by total assets. We construct the market-to-book ratio following Fama and French (2015) and use standard procedures to sort firms into five portfolios ranked by their profitability and market-to-book ratios, respectively. Following the convention in this literature, we call the portfolio with the lowest market-to-book ratio the value portfolio and the one with the highest market-to-book ratio the growth portfolio.

In Table 1, we report the median of firm characteristics, including market-to-book ratio, the gross profitability, and total factor productivity for market-to-book sorted portfolios (panel A) and those for profitability sorted portfolios (panel B). To compute firm-level productivity, we assume a Cobb-Douglas production function and use the semi-parametric method of Olley and Pakes (1996) and De Loecker et al. (2020). Table 1 reports the median of the log TFP for all firms in the same portfolio. Details of the estimation procedure and construction of variables can be found in Appendix A and D.1.

We make several observations. First, market-to-book ratio and gross profitability are highly positively correlated. The portfolios in panel A are sorted on market-to-book ratios, which increase monotonically from value to growth by construction. At the same time, the gross profitability is also monotonically increasing from value to growth. Similarly, the profitability sorted portfolios confirm the same pattern. As shown in panel B of Table 1, the profitability ratio increases from low to high by construction, while the market-to-book ratios increase monotonically with the profitability ratio.

Table 1: Firm characteristics

This table compares the median firm characteristics of portfolios sorted by market-to-book (MB) and gross profitability (GP/A). Panel A and B report the firm characteristics of portfolios sorted by MB and GP/A, respectively. $\ln TFP$ is the log of firm-level productivity. We also report the annualized excess returns of value-weighted portfolios $E[R^e]$. The sample period ranges from 1963 to 2020, at an annual frequency.

Panel A: Market-to-book sorted portfolios

	Value	2	3	4	Growth	Growth-Value
MB	0.588	1.083	1.655	2.683	6.200	5.612
GP/A	0.181	0.219	0.278	0.328	0.336	0.155
$\ln TFP$	-3.990	-3.713	-3.524	-3.319	-3.092	0.898
$E[R^e](\%)$	10.115	7.707	7.485	7.019	6.856	-3.259

Panel B: Gross profitability sorted portfolios

	Low	2	3	4	High	High-Low
MB	1.461	1.205	1.478	1.785	2.251	0.790
GP/A	0.001	0.136	0.265	0.411	0.670	0.669
$\ln TFP$	-3.999	-3.938	-3.721	-3.410	-3.210	0.789
$E[R^e](\%)$	4.700	6.333	5.443	7.275	8.018	3.319

Second, despite the robust positive correlation between the market-to-book ratio and gross profitability, portfolios formed on these two characteristics show opposite patterns in returns. In panel A of Table 1, the portfolio returns are monotonically decreasing from the value to the growth portfolio. This is the well-documented value premium. In contrast, as documented by Novy-Marx (2013), sorting on the gross profitability produces a monotonic increasing pattern of average returns from low to high gross profitability, as shown in panel B.

Third, firm-level productivity is monotonically increasing in market-to-book ratio and gross profitability. We report the median of firm-level productivity for market-to-book and gross profitability sorted portfolios in the third row of panel A and B, respectively.

This comparison highlights the challenge for production-based asset pricing models. The increasing pattern of estimated TFP with respect to the market-to-book ratio is certainly consistent with the large literature that builds Q-theory-based asset pricing models where firm-level heterogeneity is driven by Markov productivity shocks, for example, Zhang (2005), Belo et al. (2014), İmrohoroğlu and Tüzel (2014), Favilukis and Lin (2015), Li (2018), and Gomes and Schmid (2021). However, because the firm-level productivity shock is typically the only source of exogenous variation for firm characteristics, these models would unambiguously imply a negative profitability premium.

The solution we propose for the above puzzle is that firm-level productivity has a two-factor structure: a permanent component and a transitory component. Market-to-book ratio sorting primarily selects firms based on the permanent component of productivity shocks, while the profitability factor loads mainly on the transitory component. In the rest of the paper, we first provide empirical evidence for the two-factor structure of productivity, then develop a model in which the two-factor structure translates into differences in expected returns for market-to-book and profitability sorted portfolios in a general equilibrium setup.

2.2 The two-factor structure of productivity shocks

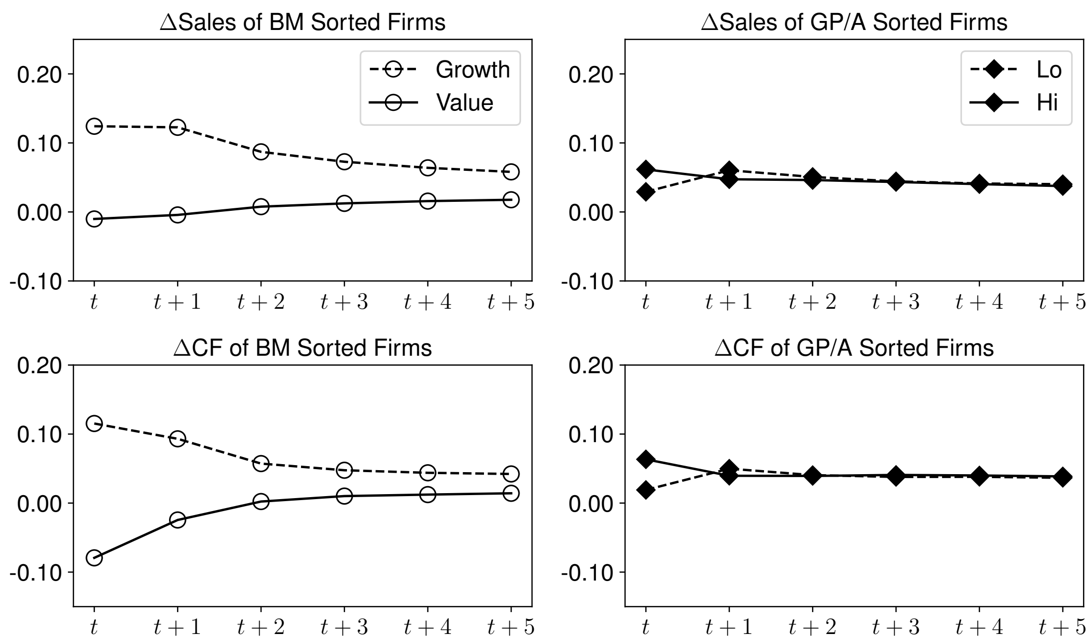
Cash flow and sales growth rates To motivate a two-factor structure of firm-level productivity shocks, we first present illustrative evidence for the persistence of the cash flow growth rates for market-to-book and profitability sorted portfolios. In order to assess the persistence of the productivity shocks, we keep track of firms and compute their cash flow and sales growth rates up to five years after portfolio formation for the five market-to-book ratio sorted portfolios and those for the five gross profitability sorted portfolios. In Figure 1, we plot the median sales growth rates (top panels) and the median cash flow growth rates (bottom panels) for the growth and value portfolios (left panels), and do the same for the high and low gross profitability portfolios (right panels) both at and after portfolio formation.

We see in the left panels of Figure 1 that firms in the growth portfolio have significantly higher sales and cash flow growth rates at the portfolio formation year (time t), and the difference persists into the future. The difference in the sales and cash flow growth rates for growth and value portfolios remains significant after five years of portfolio formation.

High gross profitability firms also have higher sales and cash flow growth rates at the portfolio formation year, as shown in the right panels of Figure 1. However, in sharp contrast with the market-to-book ratio sorted portfolios, the difference in the cash flow and sales growth rates for high and low profitability sorted portfolios completely disappears after one year.

Figure 1: Firm characteristics around portfolio formation

This figure reports the median of firm characteristics after portfolio formation year t. The left and right panels report the firm characteristics around the portfolio formation year of BM and GP/A sorted portfolios, respectively. The definition of market-to-book and profitability follows Fama and French (2015) and Novy-Marx (2013), respectively. The sample starts in 1963 and ends in 2020, at an annual frequency.



The above evidence is consistent with the hypothesis that the productivity difference that separates value and growth portfolios is very persistent. In contrast, the productivity difference for high and low profitability sorted portfolios is rather transitory.

Investment Next, we show that consistent with our two-factor structure hypothesis, both the growth and high gross profitability portfolios exhibit higher investment rates than the value and low gross profitability portfolios, respectively. However, the difference in the investment rate is significantly higher for the market-to-book ratio sorted portfolios than for the gross profitability sorted portfolios.

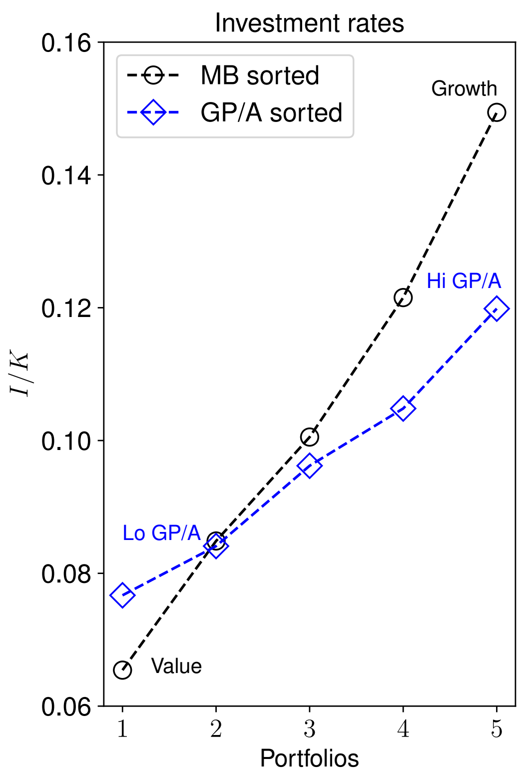
In Figure 2, we plot the investment rate (the ratio of investment to total capital stock) for the five market-to-book sorted portfolios (circles) and the same for the five profitability sorted portfolios (diamonds).[4] Clearly, the investment rate is monotonically increasing from the low profitability portfolio to the high profitability portfolio. The market-to-book ratio

[4]We do not plot a confidence interval for the point estimates, because the large number of firms in each portfolio means the standard error for the point estimates is close to zero.

sorted portfolios exhibit the same pattern, the investment rate increases from the value to the growth portfolio.

Figure 2: Investment rates

This figure presents the median investment rates across market-to-book (MB) and gross profitability (GP/A) sorted portfolios. The horizontal axis denotes the MB and GP/A sorted portfolios. The sample starts in 1963 and ends in 2020, at an annual frequency.



However, the spread in the investment rate is significantly higher for the market-to-book sorted portfolios than for the profitability sorted portfolios. Firms in the growth portfolio have investment rates that are 9% higher than value firms. By comparison, the investment rate for the high profitability portfolio is 3% lower than that of the growth portfolio. The ranking of investment rates between the two portfolios switches as we go to the value and low profitability portfolio on the left. As gross profitability decreases, the investment rate does too, but the increase in investment rate is much faster for the market-to-book sorted portfolios. Firms in the value portfolio invest at a much lower rate than firms in the low profitability portfolio. The above evidence is consistent with a recent paper, Byun et al. (2019), who document that firm investment responds strongly to permanent shocks but weakly to temporary shocks.

The above evidence also supports the hypothesis that the market-to-book sorted portfolios differ in the persistent component of firm-specific productivity, while profitability sorted portfolios differ in the transitory component. Because investment decisions are forward looking, book-to-market sorting produces heterogeneity in the persistent component of productivity shocks and leads to large differences in investment rates. In contrast, gross profitability sorting mainly identifies the transitory component of productivity, which has a much smaller impact on investment rates by comparison because the initial productivity shock is expected to diminish quickly over time. We confirm the above intuition in the structural model we develop in Section 3.

Productivity decomposition This section provides a formal statistical decomposition of the productivity shocks at the firm level to validate our hypothesis of the two-factor structure of productivity shocks. We demonstrate that the market-to-book sorted portfolios have a substantially higher spread in terms of the permanent component of productivity shocks. In contrast, profitability sorted portfolios differ more significantly in the transitory component.

We assume that firm-level productivity has a permanent component, denoted by $X_{j,t}$, and a transitory component, denoted by $Z_{j,t}$. The log of firm-level total factor productivity can be written as $\ln TFP_{j,t} = \ln X_{j,t} + \ln Z_{j,t}$, with

$$\ln X_{j,t} = \mu_j + \ln X_{j,t-1} + \sigma_{X,j}\varepsilon_{j,t}, \tag{1}$$

$$\ln Z_{j,t} = \rho_{Z,j} \ln Z_{j,t-1} + \sigma_{Z,j}\eta_{j,t}, \tag{2}$$

where $\varepsilon_{j,t}$ and $\eta_{j,t}$ are i.i.d. and follow the standard normal distribution. That is, the permanent component is modeled as a random walk, and the transitory component is modeled as an AR(1) process. We follow the procedure of Harvey (1985) to construct the permanent and transitory components for productivity shocks for each firm. In Appendix A, we provide a detailed procedure of constructing and decomposing firm-level productivity and a statistical test for the nonstationarity of the productivity process.

We report our decomposition results for market-to-book sorted portfolios (panel A) and those for profitability sorted portfolios (panel B) in Table 2. We make three observations. First, all three measures of productivity, $\ln TFP$, its permanent component, $\ln X$, and the transitory component, $\ln Z$, are increasing from value to growth and from low to high profitability portfolios. This pattern is consistent with the intuition that growth firms have a higher productivity than value firms and high gross profitability firms have higher productivity than low profitability ones. Second, the market-to-book sorted portfolios exhibit a substantial difference in the permanent component of productivity. The spread in $\ln X$ from value to growth is 0.77, while the same is significantly smaller, 0.62 ,for gross profitability

Table 2: Permanent and transitory component

This table shows the permanent and transitory components of firm-level productivity. We decompose the firm-level productivity into a permanent and a transitory component, using the unobserved components method. In the tables $\ln X$ denotes the trend component, and $\ln Z$ represents the transitory component of the productivity. The sample starts in 1963 and ends in 2020, at an annual frequency.

Panel A: Market-to-book sorted portfolios

	Value	2	3	4	Growth	Growth-Value
MB	0.588	1.083	1.655	2.683	6.200	5.612
$\ln TFP$	-3.990	-3.713	-3.524	-3.319	-3.092	0.898
$\ln X$	-3.828	-3.618	-3.458	-3.273	-3.054	0.774
$\ln Z$	-0.024	-0.008	0.000	0.003	0.009	0.033

Panel B: Gross profitability sorted portfolios

	Low	2	3	4	High	High-Low
GP/A	0.001	0.136	0.265	0.411	0.670	0.669
$\ln TFP$	-3.999	-3.938	-3.721	-3.410	-3.210	0.789
$\ln X$	-3.807	-3.749	-3.608	-3.358	-3.183	0.624
$\ln Z$	-0.042	-0.013	-0.005	0.000	0.006	0.048

sorted portfolios. Third, compared to market-to-book sorted portfolios, the profitability sorted portfolios have a higher spread in terms of the transitory component of productivity $\ln Z$. The above evidence is consistent with the hypothesis that market-to-book sorting mainly differentiates firms in terms of the permanent component of their productivity. The profitability factor loads primarily on the transitory component of firm-level productivity.[5]

3 Model setup

Motivated by the empirical evidence in the last section, we build a general equilibrium model where firm-level productivity has a two-factor structure, which includes a transitory component and a permanent component. Our model provides a unified framework to study the joint dynamics of macroeconomic quantities and asset prices. Our purpose is to demonstrate that the model provides a quantitative explanation for the coexistence of the

[5]We also perform the decomposition of firm-level productivity using HP filter, the results also suggest that the market-to-book sorted firms mainly differ with respect to the permanent component, and the profitability sorted firms mainly differ with respect to the cycle component.

value and profitability factors and the coexistence of the value and profitability premiums.

Households Time is infinite and discrete. We assume that the representative household has a recursive preference with risk aversion γ and an intertemporal elasticity of substitution (IES) ψ, as in Epstein and Zin (1989) :

$$\mathbf{U}_t = \left\{ (1-\beta)\mathbf{C}_t^{1-\frac{1}{\psi}} + \beta \left(E_t \left[\mathbf{U}_{t+1}^{1-\gamma} \right] \right)^{\frac{1-\frac{1}{\psi}}{1-\gamma}} \right\}^{\frac{1}{1-\frac{1}{\psi}}}, \tag{3}$$

where $\mathbf{U}_t$ is time-t utility and $\mathbf{C}_t$ is time-t consumption. The household receives both labor and capital income. Households trade equities and a risk-free bond. In every period t, the household purchases B_t units of the risk-free bond, which pays a risk-free interest rate $R_{f,t}$ in the next period. The household can purchase $\omega_{j,t}$ shares of the equity of firm j. We denote $V_{j,t}$ as the cum-dividend equity price of firm j at time t and $D_{j,t}$ as the dividend payment of firm j at time t. We use W_t for the real wage at time t and $L_{j,t}$ for the household's labor supply to firm j. The household budget constraint at time t can be written as

$$\mathbf{C}_t + \int \omega_{j,t} \left(V_{j,t} - D_{j,t} \right) dj + B_t = W_t \int L_{j,t} dj + R_{f,t-1} B_{t-1} + \int \omega_{j,t-1} V_{j,t} dj, \tag{4}$$

where the integration is with respect to all firms in the economy. Hereafter, our convention is to use boldface for aggregate quantities and regular font for firm-level quantities.

In the recursive equilibrium we construct, $\mathbf{C}_t$ and $\mathbf{U}_t$ are functions of a vector of state variables, which we denote as S_t. We postpone the specification of S_t and its law of motion until we introduce the definition of equilibrium later in the section. Under recursive preference, the stochastic discount factor that prices one unit of consumption goods paid at period $t+1$ (at state S_{t+1}) into period t (at state S_t)$-$consumption numeraire can be constructed from the household's intertemporal marginal rate of substitution:

$$M_{t+1}(S_{t+1}|S_t) = \beta \left(\frac{\mathbf{C}(S_{t+1})}{\mathbf{C}(S_t)} \right)^{-\frac{1}{\psi}} \left(\frac{\mathbf{U}(S_{t+1})^{1-\gamma}}{E_t \left[\mathbf{U}(S_{t+1})^{1-\gamma} \right]} \right)^{\frac{1/\psi-\gamma}{1-\gamma}}. \tag{5}$$

To save notation, we will denote $M_{t+1}(S_{t+1}|S_t)$ as M_{t+1} below.

Production function At time t, firm j produces output $Y_{j,t}$ from capital $K_{j,t}$ and labor $L_{j,t}$ using a Cobb-Douglas production function of the following form:

$$Y_{j,t} = A_t^{1-\alpha\nu} \left[\left(X_{j,t} \cdot Z_{j,t} \right)^{1-v} \left(u_{j,t} K_{j,t} \right)^{\nu} \right]^{\alpha} L_{j,t}^{1-\alpha}, \tag{6}$$

where $1-\alpha$ is labor share and ν is the span of control parameter, as in Atkeson and Kehoe (2005). To allow for variable capital utilization, we denote $u_{j,t}$ as the utilization rate of physical capital. In the above equation, A_t is the aggregate productivity, and $X_{j,t} \cdot Z_{j,t}$ is firm-specific productivity. To model a two-factor structure of productivity shock, we let $X_{j,t}$ be the permanent component and $Z_{j,t}$ be the transitory component. The law of motions of $X_{j,t}$ and $Z_{j,t}$ follow equation (1) and (2) respectively. For parsimony, in the model, we assume that the parameters for these stochastic processes μ_j, $\sigma_{X,j}$, $\rho_{Z,j}$, and $\sigma_{Z,j}$ are common across all firms and denote them as μ, σ_X, ρ_Z, and σ_Z, respectively. In addition, we assume that the aggregate productivity follows:

$$\ln A_{t+1} - \ln A_t = \mu_A + \zeta_t,$$

where $\zeta_t = \rho_\zeta \zeta_{t-1} + \varepsilon_t^\zeta$ is an AR(1) process and ε_t^ζ is a normally distributed shock, i.i.d. over time.

Variable capital utilization As argued by Novy-Marx (2013), in the data, value and profitability factors have weak and negative correlations and are likely proxies for different sources of aggregate risk. To account for this fact, we introduce a second aggregate shock, shocks to capital depreciation. Let $I_{j,t}$ denote firm j's investment at time t. The law of motion of firm j's capital accumulation follows

$$K_{j,t+1} = \big(1 - \delta(u_{j,t}, \theta_{t+1})\big) K_{j,t} + I_{j,t}, \tag{7}$$

where $\delta(u_{j,t}, \theta_{t+1})$ is the capital depreciation rate that depends on the last-period utilization rate $u_{j,t}$ as well as an aggregate shock, θ_{t+1}, which is modeled as a Markov process.[6] We assume a constant elasticity capital depreciation function as in the variable capital utilization literature, such as Greenwood et al. (1988), Jaimovich and Rebelo (2009), Garlappi and Song (2017), and Grigoris and Segal (2020):

$$\delta\left(u_{j,t}, \theta_{t+1}\right) = \delta_k + \theta_{t+1} \frac{u_{j,t}^{1+\lambda} - 1}{1+\lambda}, \tag{8}$$

where $\lambda > 0$ is the curvature parameter for the depreciation rate function, δ_k is the level parameter that determines the depreciation rate at the deterministic steady state, and θ_{t+1} is an aggregate shock that affects the depreciate rate of all firms. In the above equation, the $\delta(u, \theta)$ function is increasing and convex with respect to u: a more intensive usage

[6]Other papers that emphasize the importance of capital depreciation shocks include Gertler and Kiyotaki (2010), Gertler and Karadi (2011), Kozlowski et al. (2018), and in continuous-time setups, He and Krishnamurthy (2013) and Brunnermeier and Sannikov (2014).

of capital this period accelerates its depreciation and lowers capital accumulation in the following period. Firms choose capital utilization rates to trade off current-period production versus depreciation in the next period.

It is clear from equation (8) that firms with high utilization rates are more sensitive to depreciation rate shocks. As we will show later in this section, while the value factor loads heavily on the aggregate productivity shock, the profitability factor mainly reflects firms' exposure with respect to depreciation shocks. The coexistence of productivity shocks A_t and depreciation shocks θ_t allows our model to distinguish the value factor from the profitability factor and jointly explain the value and profitability premiums.

Firms' problem We focus on the recursive equilibrium where profit maximization can be written as a dynamic programming problem. An individual firm is identified by a vector of firm-specific state variables, (K, X, Z). Let $\mathbb{K}\times\mathbb{X}\times\mathbb{Z}$ be the state space of firm characteristics, and let Γ be a density on $\mathbb{K}\times\mathbb{X}\times\mathbb{Z}$ that describes the cross-sectional distribution of firms. Let $S = (A, \theta, \Gamma)$ be the vector of aggregate state variables, including the current-period productivity, depreciation shocks, and the distribution of firm-level state variables. In the recursive equilibrium, S_t summarizes aggregate states and determines equilibrium prices through market clearing conditions. At time t, taking the law of motion of aggregate state variables as given, firm j chooses the dividend payout, $D_{j,t}$, investment, $I_{j,t}$, and capital utilization rate, $u_{j,t}$, to maximize its cum-dividend value:

$$V(K_{j,t}, Z_{j,t}, X_{j,t}; S_t) = \max_{I_{j,t},u_{j,t},D_{j,t}} D_{j,t} + E\left[M_{t+1}V(K_{j,t+1}, X_{j,t+1}, Z_{j,t+1}; S_{t+1})|X_{j,t}, Z_{j,t}, S_t\right], \tag{9}$$

subject to the following budget constraint;

$$D_{j,t} + I_{j,t} + H(I_{j,t}, K_{j,t}) \leq \Pi(K_{j,t}, Z_{j,t}, X_{j,t}; S_t \| u_{j,t}), \tag{10}$$

and capital accumulation equation (7). Investment entails a standard convex adjustment cost that takes a quadratic form, $H(I_{j,t}, K_{j,t}) = \frac{h}{2}(\frac{I_{j,t}}{K_{j,t}} - \delta)^2 K_{j,t}$

In equation (9), Firm j's profit at time t, $\Pi(K_{j,t}, Z_{j,t}, X_{j,t}; S_t \| u_{j,t})$, is given by

$$\Pi\left(K_{j,t}, Z_{j,t}, X_{j,t}; S_t \| u_{j,t}\right) = \max_{L_{j,t}} A_t^{1-\alpha\nu}\left[(X_{j,t}Z_{j,t})^{1-\nu}(u_{j,t}K_{j,t})^{\nu}\right]^{\alpha} L_{j,t}^{1-\alpha} - W_t L_{j,t}.\text{[7]} \tag{11}$$

It is convenient to simplify equation (11) by maximizing out labor and using the labor market

[7] We use the notation $(\cdot\|u_{j,t})$ to emphasize that profit at time t depends not only on time-t state variables, but also on the choice variable $u_{j,t}$.

clearing condition $\int L_{j,t} dj = 1$ to obtain the expression for optimal labor input at firm j:

$$L_{j,t} = \frac{(X_{j,t}Z_{j,t})^{1-\nu}(u_{j,t}K_{j,t})^{\nu}}{\int (X_{i,t}Z_{i,t})^{1-\nu}(u_{i,t}K_{i,t})^{\nu} di}. \tag{12}$$

Taking advantage of the constant income share property of the Cobb-Douglas production function, we can write firm j's output as

$$Y(K_{j,t}, Z_{j,t}, X_{j,t}; S_t \| u_{j,t}) = A_t \left[A_t^{-\nu}(X_{j,t}Z_{j,t})^{1-\nu}(u_{j,t}K_{j,t})^{\nu}\right] \left[\int A_t^{-\nu}(X_{i,t}Z_{i,t})^{1-\nu}(u_{i,t}K_{i,t})^{\nu} di\right]^{\alpha-1}, \tag{13}$$

and the operating profit function as

$$\Pi(K_{j,t}, Z_{j,t}, X_{j,t}; S_t \| u_{j,t}) = \alpha Y(K_{j,t}, Z_{j,t}, X_{j,t}; S_t \| u_{j,t}). \tag{14}$$

Let $u(K, Z, X; S)$ be the policy function for capital utilization and define $\Psi(S)$ as

$$\Psi(S) = A^{-\nu} \left[\int (XZ)^{1-\nu}(u\,(K, Z, X; S)\,K)^{\nu} d\Gamma\,(K, Z, X)\right]. \tag{15}$$

We can conveniently write the profit function as

$$\Pi(K, Z, X; S \| u) = \alpha(AXZ)^{1-\nu}(uK)^{\nu}\Psi(S)^{\alpha-1}. \tag{16}$$

Using this notation, it is straightforward to integrate across all firms and compute aggregate output as

$$\mathbf{Y}_t = \int Y_{j,t} dj = A_t \Psi(S_t)^{\alpha}. \tag{17}$$

Intuitively, the aggregate production function is Cobb-Douglas, and $\Psi(S_t)$ is the productivity-adjusted capital stock of the economy.[8] In Appendix C, we develop an expression for $\Psi(S_t)$ that depends on the cross-sectional joint distribution of productivity and capital.

Entry, exit, and the dynamics of firm distribution At the end of each period t, after dividend payout and investment, incumbent firms receive a random death shock with probability κ_D and exit the economy. Upon receiving the death shock, the firm exits the economy, and its capital stock evaporates.

At the same time, a measure κ_D fraction of new firms come into existence with initial

[8]Intuitively, our model has capital misallocation (Hsieh and Klenow (2009)) due to the presence of adjustment cost. $\Psi(S_t)$ accounts for capital misallocation. In the absense of capital misallocation, $\Psi(S_t)$ is proportional to the aggregate capital stock of the economy.

productivity $\left(\bar{Z}, \bar{X}\right)$. The assumption on the entrance measure κ_D is a normalization that ensures that the total measure of firms in the economy is 1 at all times. Newborn firms start to produce one period after they enter the economy. As a result, let j be a new entrant firm at time t, the representative agent optimally choose the amount of initial capital $\bar{K}_{j,t}$ to solve the following optimization problem:

$$\bar{V}(S_t) = \max_{\bar{K}_{j,t}} -\bar{K}_{j,t} - \bar{H}(\bar{K}_{j,t}) + (1-\kappa_D)E\left[M_{t+1}V(\bar{K}_{j,t}, Z_{j,t+1}, X_{j,t+1}; S_{t+1})|\bar{X}, \bar{Z}, S_t\right]. \quad (18)$$

Let $\bar{K}(S_t)$ be the policy function of the above optimization problem. Here, setting up a new firm requires a convex entry cost $\bar{H}(\bar{K}_{j,t})$. With a probability of $(1-\kappa_D)$, new entrants survive to the next period. Conditioning on survival, new entrants carry $\bar{K}_{j,t} = \bar{K}(S_t)$ amount of capital to the next period and become an incumbent firm to produce output.

The entry and exit dynamics described above allow us to derive a law of motion for the density of the distribution of firm-level state variables, Γ_t. At time t, firm type is described by a vector of firm-specific state variables (K_t, X_t, Z_t). Let $I(K, Z, X; S)$ and $u(K, Z, X; S)$ be the policy function of incumbent firms. The law of motion of K_t is given by:

$$K_{t+1} = \left[1 - \delta\left(u(K_t, X_t, Z_t; S_t), \theta_{t+1}\right)\right] K_t + I\left(K_t, Z_t, X_t; S_t\right). \quad (19)$$

In addition, consistent with our empirical specifications (1) and (2), we write the law of motion of the exogenous productivity shocks X_t and Z_t as

$$\begin{aligned} \ln X_{t+1} &= \mu + \ln X_t + \sigma_X \varepsilon_{t+1}, \\ \ln Z_{t+1} &= \rho_Z \ln Z_t + \sigma_Z \eta_{t+1}. \end{aligned} \quad (20)$$

Given the current aggregate state S_t and the realization of next period shock θ_{t+1}, for any K_{t+1}, let $\mathbb{I}_{K_{t+1}}(K_t, X_t, Z_t|S_t, \theta_{t+1})$ be the indicator function on the state space $\mathbb{K} \times \mathbb{X} \times \mathbb{Z}$ that takes a value of 1 if (K_t, X_t, Z_t) is such that equation (19) holds. Let $\bar{\mathbb{I}}_{K_{t+1}}(S_t)$ be the indicator function that takes a value of 1 if $\bar{K}(S_t) = K_{t+1}$. The law of motion of Γ can be written as follows. For any $(K_{t+1}, X_{t+1}, Z_{t+1}) \in \mathbb{K} \times \mathbb{X} \times \mathbb{Z}$

$$\begin{aligned} &\Gamma_{t+1}(K', X', Z') \\ =&(1-\kappa_D)\int \mathbb{I}_{K'}(K, X, Z|S_t, \theta_{t+1})\phi\left(\frac{\ln X' - \mu - \ln X}{\sigma_x}\right) \cdot \phi\left(\frac{\ln Z' - \rho_z \ln Z}{\sigma_z}\right) \cdot \Gamma_t(K, X, Z)dKdXdZ \\ &+ \kappa_D \bar{\mathbb{I}}_{K'}(S_t)\phi\left(\frac{\ln X' - \mu - \ln \bar{X}}{\sigma_x}\right)\phi\left(\frac{\ln Z' - \rho_z \ln \bar{Z}}{\sigma_z}\right), \end{aligned} \quad (21)$$

where ϕ is the density of the standard normal distribution. The above law of motion has an intuitive interpretation. $\Gamma_{t+1}(K', X', Z')$ is the density of firms at location (K', X', Z') in period $t+1$. They can come from either incumbent firms or new entrants. The first line in (21) accounts for the incumbent firms: if an incumbent firm is such that equation (19) holds with $K_{t+1} = K'$, the indicator function is 1, $\Gamma_t(K, X, Z)$ is the density of such firms in period t, and $\phi\left(\frac{\ln X' - \mu - \ln X}{\sigma_x}\right) \cdot \phi\left(\frac{\ln Z' - \rho_z \ln Z}{\sigma_z}\right)$ is the probability for such firms to reach (X', Z'). The second line which accounts for new entrants can be interpreted similarly.

Recursive competitive equilibrium A recursive competitive equilibrium consists of:

- A value function for the representative household's utility, $\mathbf{U}(S)$, and the associated consumption policy function, $\mathbf{C}(S)$,
- A value function for incumbent firms, $V(K, Z, X; S)$, and the associated policy functions, $I(K, Z, X; S)$ and $u(K, Z, X; S)$,
- A value function for new firms, $\bar{V}(S)$, and the associated policy function, $\bar{K}(S)$,
- A stochastic discount factor, $M(S'|S)$, and
- A law of motion of the distribution of firms, Γ,

such that the following conditions are met:

1. given the equilibrium stochastic discount factor and the law of motion of Γ, the value function $V(K, Z, X; S)$ and the policy functions, $I(K, Z, X; S)$ and $u(K, Z, X; S)$, solve the profit maximization problem for incumbent firms, (9),
2. the value function $\bar{V}(S)$ and the policy function $\bar{K}(S)$ solve the profit maximization problem for new entrants, (18),
3. the equilibrium SDF is consistent with household consumption, (5),
4. the law of motion of Γ is consistent with firms' optimal policies, (21), and
5. goods market clears, that is, $\forall S$,

$$\begin{aligned}\mathbf{C}(S) + \int \Big(I(K, Z, X; S) + H(I(K, Z, X; S), K)\Big)\Gamma(K, Z, X)dKdZdX + \bar{K}(S) + \bar{H}(\bar{K}(S)) \\ = \int A^{1-\alpha\nu}\Big[u(K, Z, X; S)^{\nu} K^{\nu}(ZX)^{1-\nu}\Big]^{\alpha}\Gamma(K, Z, X)dKdZdX,\end{aligned} \tag{22}$$

and,

6. labor market clears.[9]

4 Asset pricing implications

We provide a summary of the main asset pricing implications of our model, which serves as a road map for the rest of the paper.

A1 The market-to-book ratio and the gross profitability are both increasing functions of X and Z. However, quantitatively, firms' market-to-book ratio is primarily determined by the permanent component, X, and the transitory component Z mainly determines the gross profitability.

A2 While firm-level cash flow responds positively to both X and Z, the impulse responses with respect to X are much more persistent than those with respect to Z.

A3 The firm investment rate is increasing in X, and the capital utilization rate is increasing in Z.

A4 The equity value of high-X firms is less sensitive to productivity shocks A_t than the equity value of low-X firms. The equity value of high-Z firms is more exposed to capital depreciation shocks θ_t than that of low-Z firms.

Taken together, the above asset pricing implications provide an explanation for the co-existence of the value and profitability premiums and the empirical evidence we document in Section 2. In the rest of this section, we use equilibrium conditions to explain the intuition for the above results. In Section 5, we calibrate our model and assess the quantitative importance of the proposed economic mechanisms.

Market-to-book ratio and investment In this section, we use firms' optimality condition with respect to investment to provide an explanation for implications A1 and A3 above. To simplify our analysis, we first note that under the assumption that the growth rates of aggregate productivity shocks are i.i.d., the firm's value function and policy functions satisfy a homogeneity property. To understand the homogeneity property, note that equation (16) implies that if we define $k_{j,t} = \frac{K_{j,t}}{A_t X_{j,t}}$, then the output function can be written as $Y(K, Z, X; S\|u) = A(XZ)^{1-\nu}(uk)^{\nu}\Psi(S)^{\alpha-1}$ and is homogeneous of degree 1 with respect to AX.

[9] To simplify equilibrium conditions, we have imposed labor market clearing to derive the optimal labor allocation equation (12).

In Appendix C, we use this homogeneity property to further show that the recursive equilibrium defined in the last section can be constructed using pricing and policy functions defined on a lower dimensional space. Under this construction, the cross-sectional distribution of firm types can be summarized by a measure defined on the space of (k, Z), which we denote as a summary measure m. In addition, under this construction, $\Psi(S)$ defined in equation (15) is only a function of m, which, with a bit of an abuse of notation, we will denote as $\Psi(m)$. This allows us to write the output function as $Y(K, Z, X; S\|u) = AXy(Z, k, m||u)$, where $y(Z, k, m||u) = Z^{1-\nu}(uk)^{\nu}\Psi(m)^{\alpha-1}$ is the productivity-normalized value of output. The value function and policy functions satisfy a similar homogeneity property, which we summarize in the following proposition.

Proposition 1. *(Homogeneity)*

The value function and policy functions for the firm maximization problem satisfy the following homogeneity property:

$$\begin{aligned} V\left(K_{j,t}, Z_{j,t}, X_{j,t}; S_t\right) &= A_t X_{j,t} v\left(k_{j,t}, Z_{j,t}; m_t\right), \\ I\left(K_{j,t}, Z_{j,t}, X_{j,t}; S_t\right) &= A_t X_{j,t} i\left(k_{j,t}, Z_{j,t}; m_t\right), \\ D\left(K_{j,t}, Z_{j,t}, X_{j,t}; S_t\right) &= A_t X_{j,t} d\left(k_{j,t}, Z_{j,t}; m_t\right), \end{aligned} \tag{23}$$

for some normalized value functions and policy functions, v, i, and d defined on a lower dimensional space. The definition and the the law of motion of the summary measure m are given in Appendix C.

Proof. See Appendix B. □

The above proposition implies that in the cross-section, firm type can be described by two state variables, $k_{j,t}$ and $Z_{j,t}$. As will soon be clear, quantitatively, the market-to-book ratio is primarily determined by the normalized capital $k_{j,t}$, and the gross profitability depends mainly on the state variable $Z_{j,t}$. Firms with high k have accumulated a high level of capital stock relative to the permanent component of their productivity and are high book-to-market ratio firms. Firms with high transitory component of productivity Z have a high profitability ratio.

The homogeneity property summarized in the above proposition is also important for numerically solving the model, as it reduces the dimensionality of the value and policy functions. It also allows us to summarize the cross-section distribution of firms with a lower dimensional measure that can be computed efficiently using local approximations.

Because the adjustment cost function $H(I, K)$ is constant returns to scale, we write $H(I, K) = h\left(\frac{I}{K}\right) K$. To save notation, we denote $h_K\left(\frac{I}{K}\right) = \frac{\partial}{\partial K} h\left(\frac{I}{K}\right) K = h\left(\frac{I}{K}\right) - h'\left(\frac{I}{K}\right)\frac{I}{K}$

as the partial derivative of the adjustment cost function with respect to capital. In addition, we denote $M_{t,t+s} = \Pi_{\tau=1}^{s} M_{t+\tau}$ as the s-period-ahead pricing kernel at date t. We also define $\hat{\beta}_{t,t+s} = (1-\kappa_D)^s \prod_{\tau=1}^{s-1} \Big(1 - \delta(u_{t+\tau}, \theta_{t+\tau+1})\Big)$ as the effective depreciation rate of capital from period t to period $t+s$. The following proposition is the standard Q-theory relationship between the investment rate and the discounted value of the marginal product of capital.

Proposition 2. *(Q-theory of investment)*

The optimal investment-to-capital ratio satisfies

$$\left[1 + h'\Big(\frac{I_t}{K_t}\Big)\right] = \mathbb{E}_t\Big[\sum_{s=1}^{\infty} M_{t,t+s}\hat{\beta}_{t,t+s}\alpha\nu \frac{Y_{t+s}}{K_{t+s}}\Big] - \mathbb{E}_t\Big[\sum_{s=1}^{\infty} M_{t,t+s}\hat{\beta}_{t,t+s} h_K\Big(\frac{I_{t+s}}{K_{t+s}}\Big)\Big], \tag{24}$$

where to save notation, we suppress the firm subscript j.

The equilibrium market-to-book ratio for firm j at time t satisfies

$$\frac{V_t - D_t}{K_{t+1}} = \left[1 + h'\Big(\frac{I_t}{K_t}\Big)\right] + (1-\nu)\mathbb{E}_t\Big[\sum_{s=1}^{\infty}(1-\kappa_D)^s M_{t,t+s}\frac{\Pi_{t+s}}{K_{t+1}}\Big]. \tag{25}$$

Proof. See Appendix B. □

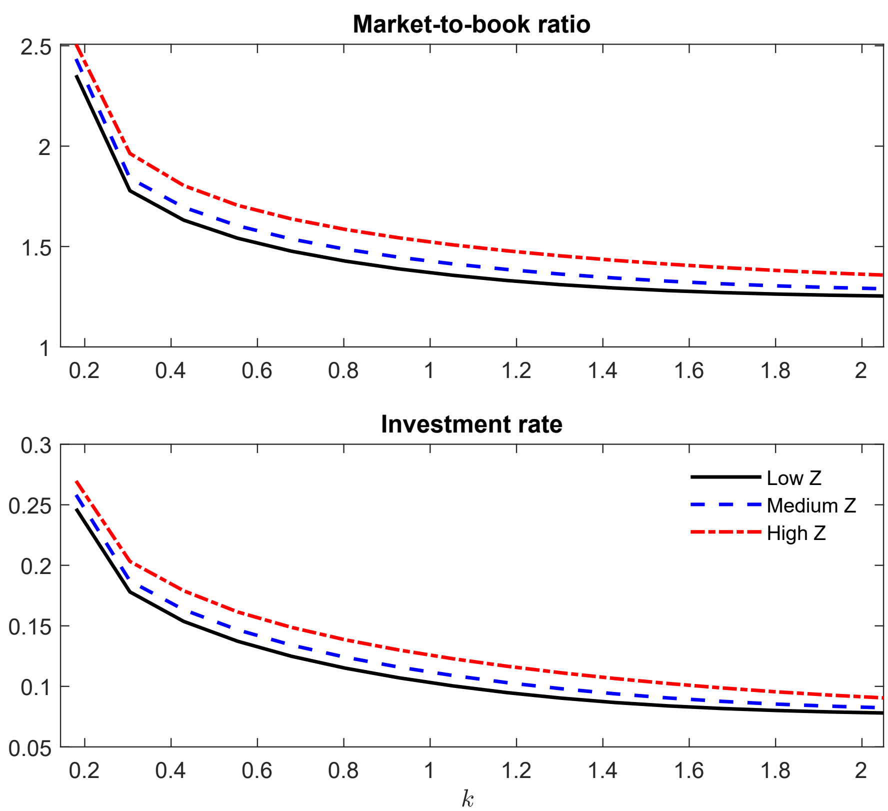
Equation (24) is the first-order optimality condition with respect to investment. The left-hand side is the marginal cost of investment. The right-hand side is the marginal benefit, which equals the present value of the future marginal product of capital plus the savings in adjustment cost due to an additional unit of investment. Note that in our model, the production function is Cobb-Douglas with decreasing returns to scale. As a result, factor income is a constant share of output, with labor share $1-\alpha$, capital share $\alpha\nu$, and profit share $\alpha(1-\nu)$. As a result, in equation (24), the marginal product of capital is proportional to its average product, with a factor of $\alpha\nu$, which is the share of capital income in total output.

Equation (25) relates the average Q or the market-to-book value of a firm to the present value of its future profit. In our model, V_t is the cum-dividend value of the firm at time t. The left-hand side of equation (25) is the ex-dividend value of the firm divided by its total capital stock, K_{t+1}, or equivalently, the market value of the firm divided by its book value. It is well known that with constant returns to scale, average Q should equal marginal Q. In our model, the production technology is decreasing returns to scale, average Q in equation (25) exceeds marginal Q in equation (24), as average Q includes profit share, but marginal Q does not. It is clear from equation (25) that the average Q equals marginal Q plus the share of profit accrued to one unit of capital, where the share of profit in total revenue is $1-\nu$. Quantitatively, however, the difference between marginal Q and average Q in our model is not very large. As ν is close to one, the profit share is fairly small in our calibration.

As we remark in asset pricing implication A1, quantitatively, the firm-level market-to-book ratio is mainly affected by the permanent component of productivity shock X and not the transitory component Z. A positive shock to both X and Z raises future productivity. Because of adjustment cost, increases in investment do not completely offset the impact of productivity shocks on the marginal product of capital. As a result, both marginal Q and average Q rise. However, the impact of X is long-lasting, while that of Z is transitory. To illustrate the difference between X and Z on firm investment and the market-to-book ratio, we plot the market-to-book ratio and the investment rate ($\frac{I}{K}$) as functions of the state variable k for three values of Z in Figure 3.

Figure 3: Investment and market-to-book ratio

This figure plots firms' optimal policy for the investment rate and market-to-book ratio, as functions of normalized firm-level capital k and transitory productivity Z, at the deterministic steady state. We define the normalized firm-level capital as $k = \frac{K}{AX}$. The solid black, dashed blue, and dotted red lines denote firms with low, medium, and high levels of transitory productivity Z, respectively.



In our formulation, $k = \frac{K}{AX}$. Hence, the level of permanent productivity shock X is inversely related to the state variable k. It is clear from Figure 3 that the investment rate and the market-to-book ratio are increasing in both X and Z. Note that the differences in the investment rate and the market-to-book ratio are much more pronounced across k than those for different values of Z.

In our model, individual firm characteristics are summarized by two state variables, k and Z. Because the impact of Z on the market-to-book ratio is relatively small, sorting firms on the market-to-book ratio separates them primarily along the k dimension. Intuitively, high-k firms have a large capital stock but a relatively low level of the permanent component of productivity. They are on the right of the state space in Figure 3 and have a significantly lower investment rate than firms with a high market-to-book ratio, or growth firms. This feature of our model is, therefore, consistent with the evidence in Figure 2, where market-to-book sorted portfolios differ significantly in terms of their investment rate. In contrast, sorting on profitability ratio primarily separates firms along the Z dimension. Because in our model, the transitory shocks Z are larger in magnitude than the permanent shocks X.

To illustrate asset pricing implications A2 and A3, we plot the impulse response functions of a median firm's normalized operating profit or cash flow, $y(Z, k, m)$, with respect to a one-standard-deviation shock to X (solid black line) and that with respect to a one-standard-deviation shock to Z (dotted red line), respectively, in the top panel of Figure 4. In the bottom panel, we plot the impulse response functions of the same firm's investment rate with respect to X and Z shocks. The horizontal axis is the number of years after the initial shock, and the vertical axis is the percentage deviation from the steady state. All parameters are chosen the same as those in the calibration we present in Table 3 of Section 5.1.[10]

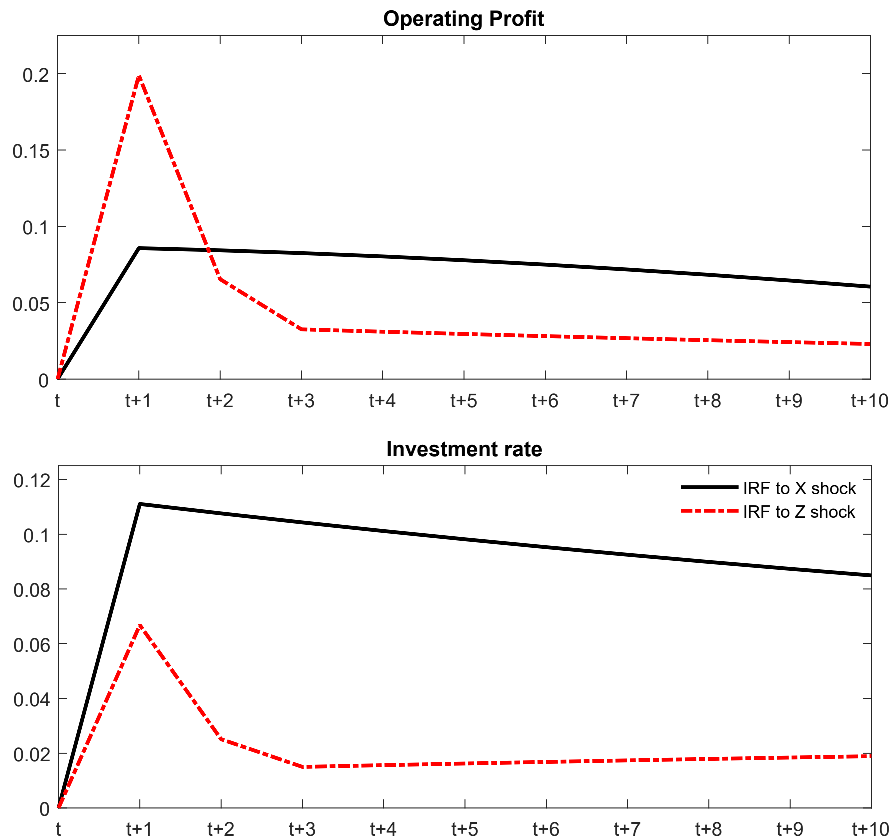
Asset pricing implication A2 states that the impulse response of firm cash flow with respect to the permanent shock X is long-lasting, while that with respect to Z is transitory. This is due to the difference in the persistence in these shocks, which is highlighted in the top panel of Figure 4. As in the data, the transitory shock Z has a larger conditional volatility than X but a lower persistence. As a result, the firm operating profit responds strongly to shocks to Z contemporaneously. However, due to the lack of persistence, this effect diminishes quickly over time. The effect of the permanent productivity shock, although smaller upon impact, is much more persistent. This pattern of our model is thus consistent with the empirical evidence in Figure 1 on the pattern of cash flow growth rates for market-to-book and profitability sorted portfolios.

The bottom panel of 4 illustrates the intuition for investment in asset pricing implication A3. Clearly, both shocks raise the level of the investment rate. Despite the higher conditional volatility in Z, however, the impulse response of the investment rate with respect to X is much more prominent and persistent. Investment decisions are forward-looking. Because

[10]Note that the impulse response function depends on the firm-level state variable (Z, X). In Figure 4, we choose both X and Z to be their medians.

Figure 4: Impulse response functions for transitory and permanent shocks

This figure compares the impulse response of the firm's investment rate and operating profit to positive shocks to transitory productivity Z and permanent productivity X while shutting down aggregate uncertainty. The size of both shocks is assumed to be one standard deviation above their steady-state values, respectively. Both shocks impact the firm at time $t+1$. The x-axis denotes simulation time, and the y-axis denotes the log deviation from the steady state.

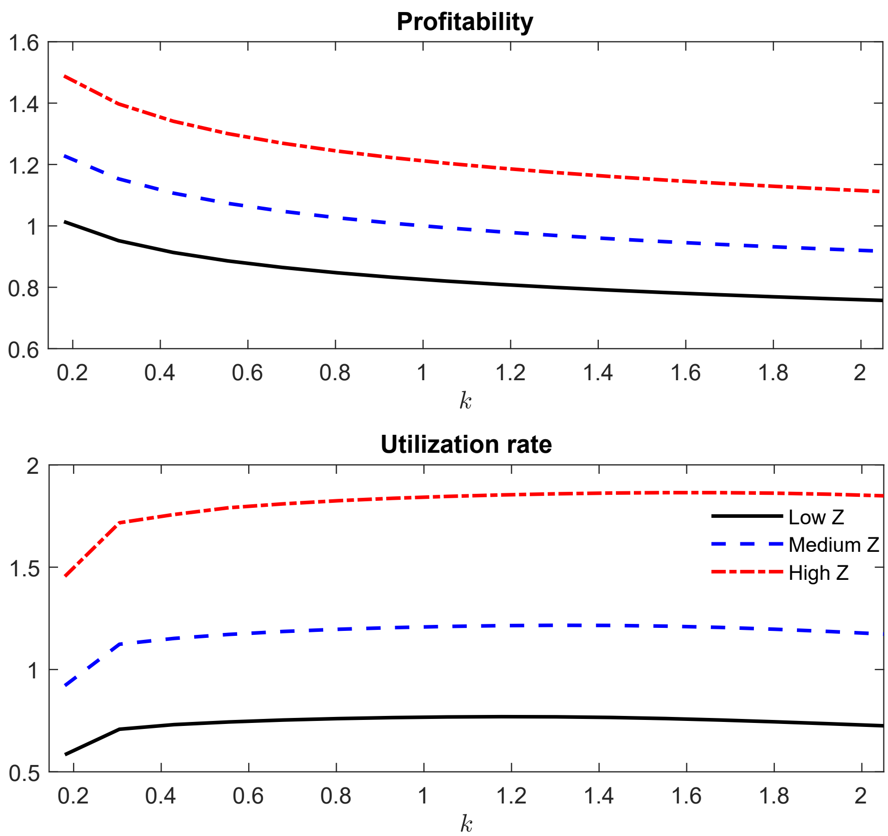


X shocks are permanent, they are expected to affect the marginal product of capital for a long period. As a result, investment has a strong contemporaneous response. In addition, due to the presence of adjustment costs, the impact of the permanent productivity shock on investment persists into the future. In contrast, despite the large shock to the transitory component Z, investment has only a moderate contemporaneous response, and the effect quickly diminishes after one year.

Capital utilization and profitability In this section, we discuss the optimality condition with respect to capital utilization and the intuition for related asset pricing implications. A firm's profitability ratio, that is, gross profit divided by capital, is given by $\frac{\Pi(K,Z,X;S)}{K} = \alpha\frac{y(Z,k,m)}{k} = Z^{1-\nu}u^{\nu}k^{\nu-1}\Psi(m)^{\alpha-1}$. Clearly, the profitability ratio in our model is an increasing function of Z. It is also increasing in X, because $k = \frac{K}{AX}$ is inversely related to X. Quantitatively, however, because the conditional volatility of Z in our model is much larger than the conditional volatility of X, heterogeneity in gross profitability in the cross-section is primarily driven by differences in the transitory component of productivity Z.

Figure 5: Utilization and profitability

This picture plots firms' optimal policy for capital utilization and the gross profitability, as functions of normalized firm size k and transitory productivity Z, at the deterministic steady state. We define the normalized firm-level capital as $k = \frac{K}{AX}$. The dashed, dotted, and solid lines denote firms with low, medium, and high transitory productivity Z.



To illustrate the dependence of the profitability ratio on state variables, in Figure 5, we plot the gross profitability ratio (top panel) and the policy function of capital utilization rate (bottom panel) as functions of normalized capital stock, k, for three levels of productivity Z. As shown in the top panel of the figure, the difference in the gross profitability ratio is quite substantial across different levels of Z, but almost flat across different levels of k. These patterns imply that qualitatively, profitability is an increasing function of both X and Z in our model. However, quantitatively, profitability is primarily driven by state variable Z and not k. This is in sharp contrast with the pattern of the market-to-book ratio in Figure 3, where the heterogeneity mainly comes from k.

The top panels of Figures 3 and 5 together confirm asset pricing implication A1. That is, sorting on the market-to-book ratio identifies the heterogeneity in the permanent component of productivity X, and sorting on the gross profitability identifies the transitory component of productivity Z.

Asset pricing implication 3 states that while investment responds strongly to the permanent component of productivity shock X and weakly to the transitory component Z, capital utilization increases sharply with Z but responds weakly and even negatively to X. To understand this point, we first state a proposition about the optimality condition for capital utilization.

Proposition 3. *The optimal utilization policy is given by*

$$u_t^{1+\lambda} = \frac{\nu \Pi_t}{(1-\kappa_D)E[M_{t+1}\theta_{t+1}\tilde{q}_{t+1}]}, \tag{26}$$

where $\tilde{q}_{t+1}$ is defined recursively in equation (B8) in Appendix B.

In our model, a higher utilization rate increases current-period output at the expense of faster capital depreciation. Firms choose the capital utilization rate optimally to balance the cost and benefit. The numerator of equation (26) is the current period capital income (which is a constant fraction of output) and therefore captures the marginal benefit of utilization. The denominator is the marginal benefit of saving an additional unit of capital: it is the present value of the marginal benefit of investment in the next period. The variable $\tilde{q}_{t+1}$ is the marginal value of capital at the beginning of period $t+1$ before depreciation (See equation equation (B8)). The term $E[M_{t+1}\theta_{t+1}\tilde{q}_{t+1}]$ includes θ_{t+1}, because the cost of capital utilization depends on the depreciation rate in the next period, which is a function of θ_{t+1}.

Proposition 3 is intuitive. The capital utilization rate is increasing in the current-period profit but decreasing in the present value of the marginal benefit of investment. An increase in Z raises both the current-period output (the numerator) and the marginal benefit of investment (the denominator). As we explained above, the impact of Z on the market-to-book ratio is minimal due to the lack of persistence. However, its impact on the current-period output is substantial due to its large magnitude. In the bottom panel of Figure 5, we plot the optimal utilization rate as functions of k for different levels of Z. It clearly demonstrates that the optimal utilization rate is strongly increasing in Z but is almost flat in X. Taken together, the bottom panels of Figures 3 and 5 confirm asset pricing implication A3, that is, investment rates increase strongly in the permanent component of productivity X. At the same time, capital utilization is much more sensitive to the transitory component of productivity shock Z.

Value and profitability premiums The two-factor productivity structure allows us to distinguish value/growth from profitability sorted firms. In our model, sorting on the market-to-book ratio distinguishes firms in terms of the permanent component of productivity and identifies the state variable k. Sorting on profitability distinguishes firms in terms of the transitory component of productivity and identifies the state variable Z. Therefore, the above mechanisms distinguish the value factor and profitability factor as characteristics that capture different aspects of firm-level fundamentals. In this section, we discuss asset pricing implication A4 of our model by focusing on the relationship between k and Z and expected returns, and the coexistence of the value and profitability premiums.

First, risk exposure to capital depreciation shocks is increasing in Z, generating a profitability premium in our model. From equation (7), we see that capital depreciation is a negative cash flow to firms and constitutes a form of operating leverage. As we explained earlier, high-Z firms have higher capital utilization and, therefore, a faster capital depreciation rate. A positive depreciation shock to θ is a negative shock to the economy as it depletes the total capital stock. These shocks have a more significant impact on high-Z firms, as they have larger capital depreciation. Low-Z firms, by contrast, are much less affected by depreciation shocks due to their low capital utilization and depreciation rate. As a result, the profitability premium in our model arises because the profitability ratio is increasing in Z.

Note that the state variable k has very little impact on the optimal capital utilization rate, as shown in the bottom panel of Figure 5. This is because an increase in the permanent component of productivity X has two offsetting effects. It increases the current-period productivity and encourages capital utilization on the one hand and raises expected future productivity, and creates an incentive to save on capital utilization and lower depreciation on the other hand. As a result, firms in different book-to-market ratio sorted portfolios have little difference in the capital utilization rate and little difference in their exposure to capital depreciation shocks.

Next, risk exposure to aggregate productivity shocks is increasing in k, generating a value premium. Recall that $k = \frac{K}{AX}$. As a result, in our model, high-k firms have a low level of permanent productivity shock, X, and are low market-to-book ratio firms. The mechanism for a higher risk exposure with respect to aggregate productivity shock A_t is similar to that in Zhang (2005): the presence of a convex adjustment cost provides a mechanism of operating leverage. The capital adjustment cost constitutes a higher fraction of the cash flow for firms with low levels of permanent productivity shock, therefore, higher operating leverage. As a result, low permanent productivity firms are more exposed to aggregate TFP shocks and require a higher expected return in equilibrium. We now turn to the quantitative implications of our model.

The above operating leverage channel is also present for firms with different levels of Z. As a result, low-Z firms, which are low profitability firms, also have higher operating leverage with respect to aggregate TFP shocks for the same reason. The combination of adjustment cost of investment and TFP shocks dampens the profitability premium. This is the main challenge raised by Novy-Marx (2013) for Q-theory-based asset pricing models. However, quantitatively, this mechanism is not strong enough to completely offset the profitability premium. As shown in Figure 3, the dispersion in investment rates is much higher along the k dimension than the Z dimension. Because adjustment cost is a function of investment rate, the operating leverage coming from the investment channel is more pronounced across book-to-market sorted portfolios than across profitability sorted portfolios. In addition, as we demonstrate in the next section, this offsetting mechanism also allows our model to capture the empirical fact that double sorted portfolios have a more significant profitability premium than single sorted portfolios.

5 Quantitative results

In this section, we calibrate our model and evaluate its ability to jointly account for the cross-section of investment dynamics and expected returns. The general equilibrium setup allows us to calibrate the model parameters to match aggregate moments and evaluate its ability to explain the patterns of the cross-section of investment dynamics and the coexistence of the value and profitability premium, as we documented in Section 2.

In our model, the distribution of firms summarized by the measure m is an aggregate state variable that affects firm investment decisions and stock market valuation. The traditional approach for solving general equilibrium models with heterogeneous firms is Krusell and Smith (1998), where the distribution is summarized by a small set of moments, and the law of the motion of the moments is simulated. We take a different approach to provide a more accurate approximation for the law of motion of this distribution. In Appendix C, we show that the law of motion of the summary measure m can be characterized analytically. This summary measure allows us to discretize this distribution using a high-dimensional vector. More importantly, it allows us to use a local approximation method to solve the model numerically. By approximating the distribution using a high-dimensional vector, our approach can more accurately account for the general equilibrium impact of time-varying distribution. This is important for several quantitative implications of our model.[11]

[11]We use the Dynare to solve our model with third-order approximation.

5.1 Calibration

We calibrate the parameters of our model to match aggregate macroeconomic quantity dynamics whenever possible. Macroeconomic variables are obtained from the Federal Reserve Bank of St. Louis. The sample period is from 1963 to 2020, at an annual frequency. Real output is GDP divided by the GDP deflator. Consumption is personal consumption expenditures divided by CPI. Real investment is defined as the private nonresidential investment divided by the investment goods deflator. Real aggregate capital is the private nonresidential fixed assets deflated by the investment goods deflator.

Table 3: Calibration

Time discount rate	β	0.975
Relative risk aversion	γ	16
IES	ψ	1.5
Capital share in production	α	0.3
Span of control parameter	ν	0.85
Depreciation rate constant	δ_k	0.1
Curvature of depreciation rate	λ	0.255
Capital adj. cost paramter	η	2.5
Mean productivity growth rate	μ_A	0.015
Persistence of long-run risks	ρ_ζ	0.988
Volatility of long-run risks	σ_ζ	0.002
Mean depreciation shock	μ_θ	0.23
Persistence of depreciation shock	ρ_θ	0.11
Volatility of depreciation shock	σ_θ	0.75
Mean of idio. permanent productivity shock	μ_X	-0.02
Volatility of idio. permanent productivity shock	σ_X	0.2
Persistent of idio. transitory productivity shock	ρ_Z	0.15
Volatility of idio. transitory productivity shock	σ_Z	0.8
Exit probability	κ_D	0.1
Initial productivity	X_0	1

We calibrate our model at an annual frequency and present the parameters in Table 3. The first group is related to preference. We choose a risk aversion of $\gamma = 16$ and an intertemporal elasticity of $\psi = 1.5$ in line with the long-run risk literature. We set the time discount factor $\beta = 0.975$ so that together with a growth rate of 1.5% per year at the steady state, our model produces a risk-free interest rate of 2.50% per year, consistent with the data.

The second group of parameters are production technology parameters. We set the capital share parameter α to 0.3 to match the average capital share in the U.S. economy. The span of control parameter ν is set to 0.85, which is broadly consistent with its estimates in the literature, for example, Atkeson and Kehoe (2005). We set the average depreciation rate of capital, $\delta_k = 10\%$ per year following the RBC literature (Kydland and Prescott (1982)). We set the curvature of the variable capital utilization function $\lambda = 0.255$ so that together with the calibrated stochastic process of θ_t, our model produces the volatility of capital utilization rate of about 5% per year, which is consistent with the volatility of capacity utilization rate in the data. We set the adjustment cost parameter $\eta = 2.5$ to match the ratio of the volatility of aggregate investment to the volatility of output. Our value of investment adjustment cost parameter is consistent with the estimations in Cooper and Haltiwanger (2006) and David and Venkateswaran (2019)

The third group of parameters are related to aggregate shocks. We follow Croce (2014) and assume that the aggregate productivity A_t follows the stochastic process:

$$\begin{aligned} \ln A_{t+1} - \ln A_t &= \mu_A + \zeta_t, \\ \zeta_t &= \rho_\zeta \zeta_{t-1} + \varepsilon_t^\zeta, \end{aligned}$$

where ε_t^ζ follow normal distributions with zero mean and standard deviation σ_ζ. We set the growth rate $\mu_A = 1.5\%$ to match the average growth rate of the U.S. economy in our sample. As in Croce (2014), we set the persistence of ζ_t, $\rho_\zeta = 0.988$ and the standard deviation $\sigma_\zeta = 0.2\%$.

We assume the aggregate depreciation shock follows an autoregressive process,

$$\ln \theta_t - \ln \mu_\theta = \rho_\theta (\ln \theta_{t-1} - \ln \mu_\theta) + \varepsilon_t^\theta,$$

where ρ_θ is the persistence, μ_θ is the steady state level of depreciation shock. To calibrate these parameter values, we first impute the time-series of aggregate capital depreciation and capital utilization from the data. In our model, the law of motion of aggregate capital stock is given by:

$$\int K_{j,t+1} dj = \int K_{j,t} dj - \int \delta(u_{j,t}, \theta_{t+1}) K_{j,t} dj + \int I_{j,t} dj. \tag{27}$$

Given the time-series of aggregate capital stock and aggregate investment, the above equation allows us back out the time-series of aggregate capital depreciation,

$$\frac{\int \delta(u_{j,t}, \theta_{t+1}) K_{j,t} dj}{\int K_{j,t} dj} = \frac{\int K_{j,t} dj + \int I_{j,t} dj - \int K_{j,t+1} dj}{\int K_{j,t} dj}. \tag{28}$$

We use the above accounting identity in the data to impute aggregate capital depreciate rates.

We set its persistence $\rho_\theta = 0.11$ and volatility $\sigma_\theta = 0.75$, so that the model-simulated aggregate depreciation rates computed from Equation (28) can match the volatility and autocorrelation of the empirical counterpart. The volatility and autocorrelation of the imputed aggregate capital depreciation rate are 2.2% and 10%, respectively. We provide details of this imputation procedure in Appendix D.2. The steady-state level of depreciation shock μ_θ is set to match the volatility of the aggregate utilization rate in the data, 5%.

The last group of parameters are those for firm-level productivity shocks. We set the standard deviation of permanent shock $\sigma_X = 20\%$ and the standard deviation of transitory shock $\sigma_Z = 80\%$ based on estimates reported in Gourio (2008). We normalize the mean of the permanent component so that the average firm's growth rate in the cross-section is the same as that of the aggregate economy. We calibrate the exogenous firm death rate κ_D at 10% per year to match the average exit rate in the data. The new entrant firms' initial productivity is normalized, and we set it so that the summary measure m integrates into one.

5.2 Aggregate moments

We now turn to the quantitative performance of the model at the aggregate level. We solve and simulate our model at an annual frequency and compute model-implied annual moments. We show that our model is broadly consistent with the key empirical features of macroeconomic quantities and asset prices. Table 4 reports the model-simulated moments of macroeconomic quantities and asset returns and compares them to their counterparts in the data.

In terms of aggregate moments, our model inherits the success of real business cycle models with respect to the quantity side of the economy. The investment-to-output ratio is 22%, which is close to the value of 17% in the data. The volatility of output growth is about 2.35% per annum both in our model and in the data. Our model also delivers a smoother aggregate consumption process (the volatility of consumption growth is 0.94 times the volatility of aggregate output growth) and a more volatile aggregate investment process (investment growth volatility is more than 4 times that of output growth) than the aggregate output, as in the data.

Turning attention to the asset pricing moments (bottom panel), we see that our model produces a low and smooth risk-free rate, with a mean of 2.5% and a standard deviation of 1.18%. We follow Croce (2014) and report equity returns in our production economy by

Table 4: Aggregate moments

This table presents moments from the model simulation and the data, at an annual frequency. The top panel reports the basic statistics of macroeconomic quantities, and the bottom panel reports the moments for asset prices.

	Data	Model
$\sigma(\Delta y)$	2.14	2.35
$\sigma(\Delta c)/\sigma(\Delta y)$	0.94	0.95
$\sigma(\Delta i)/\sigma(\Delta y)$	3.02	4.22
$AC1(\Delta y)$	0.33	0.32
$corr(\Delta y, \Delta c)$	0.88	0.88
$corr(\Delta y, \Delta i)$	0.71	0.64
$E[R^M - R^f]$	7.37	4.48
$\sigma(R^M - R^f)$	16.59	7.46
$E[R^f]$	1.28	2.50
$\sigma(R^f)$	2.22	1.18

assuming a financial leverage of 2. Our model produces an equity premium of 4.48% and a standard deviation of the market return of 7.46% per year. Overall, the asset pricing side of the economy is comparable to standard models with long-run productivity risks.

5.3 Value and profitability premiums

This section presents our model's implications on the cross-section of expected returns as well as key firm characteristics for the value and profitability portfolios that we document in Section 2. We simulate 3000 firms each year, which is roughly the same sample size as CRSP. Using our simulated data, we construct five market-to-book ratios sorted portfolios, and five gross profitability sorted portfolios, as we did in the data.[12]

In Table 5, we report the statistics for market-to-book sorted portfolios in panel A, and those for the profitability sorted portfolios in panel B. We make several observations. First, our model generates a significant value premium as well as a profitability premium. The expected return is monotonically increasing from value to growth, and the spread between the value and growth portfolios is 6.85% per year. Similarly, the expected return is increasing in the gross profitability sorted portfolios in panel B of the table, and the spread between the

[12]The quantitative results presented in this section use numerical solutions in Dynare with second-order approximation, as we need a relatively large number of grid points for the summary measure m to provide an accurate solution for firm-level policy functions.

high and low profitability portfolios is 2.27% per year. This pattern of the expected return is consistent with the value and profitability premiums we report in Table 1.

Table 5: Cross-sectional moments

This table presents the moments of model-simulated data. Panel A and B report the moments of the market-to-book ratio (MB) and the gross profitability (GP/A) sorted portfolios, respectively. We report the median of firm characteristics, including MB, GP/A, and investment rate (I/K). We also report the permanent component $\ln X$ and transitory component $\ln Z$ of firm-level productivity. The annual excess returns of value-weighted portfolios $E[R^e]$ are reported in percentages. We assume a financial leverage of two for model-simulated excess returns. All variables are at an annual frequency.

Panel A: Market-to-book sorted portfolios

	Value	2	3	4	Growth	Growth-Value
MB	1.32	1.41	1.58	1.79	2.50	1.18
GP/A	0.15	0.17	0.18	0.18	0.19	0.04
I/K	0.07	0.09	0.12	0.16	0.24	0.17
$\ln TFP$	-0.10	-0.08	0.02	0.15	0.77	0.87
$\ln X$	-0.15	-0.09	0.03	0.16	0.33	0.48
$\ln Z$	0.01	-0.04	-0.03	-0.06	0.23	0.23
$E[R^e](\%)$	6.12	4.85	3.51	2.83	-0.73	-6.85

Panel B: Gross profitability sorted portfolios

	Low	2	3	4	High	High-Low
MB	1.41	1.63	1.51	1.65	1.65	0.24
GP/A	0.11	0.15	0.18	0.20	0.26	0.11
I/K	0.09	0.13	0.11	0.13	0.13	0.04
$\ln TFP$	-0.89	-0.31	0.09	0.68	1.11	2.00
$\ln X$	0.17	0.18	0.09	0.17	0.07	-0.10
$\ln Z$	-1.04	-0.40	0.00	0.51	1.07	2.11
$E[R^e](\%)$	3.49	4.06	4.23	5.37	5.76	2.27

Second, as in the data, the market-to-book ratios and the gross profitability ratios are positively correlated. In panel A, gross profitability is increasing from the value to growth portfolios, and in panel B, the market-to-book ratio is monotonically increasing in the gross profitability. In our model, this is because both the gross profitability and the market-to-book ratio are increasing in the level of productivity. This is the key element in the puzzle of the profitability premium, and it is the challenge for traditional production-based asset pricing models with only one factor in firm-level productivity. In one-factor-based models, both the gross profitability and the market-to-book load positively on the level of firm-level productivity. As a result, these models cannot simultaneously explain the value and profitability premiums.

Third, both the permanent and the transitory components of productivity are increasing from value to growth portfolios as well as from low to high profitability ratio portfolios. However, the spread in the permanent component, $\ln X$, is much larger for market-to-book sorted portfolios than for the profitability sorted ones. The spread in the transitory component $\ln Z$ is more significant for profitability sorted portfolios than for market-to-book sorted ones. As we explained earlier, in our model, sorting on the market-to-book ratio identifies the permanent component of productivity shock X, as shocks to X are persistent, and the market-to-book is a forward-looking measure of the present value of the marginal product of capital. In panel A, the average $\ln X$ for the value portfolio is -0.15, and that for the growth portfolio is 0.33, generating a difference of 0.48. By comparison, the gross profitability sorted portfolios almost all have roughly the same average level of $\ln X$, with a spread of -0.10 between the high and low profitability portfolios. Because the conditional volatility of $\ln X$ is small, its impact on profitability is limited.

The average level of $\ln Z$ exhibit a pattern different from that of $\ln X$. Because the conditional volatility of $\ln Z$ is much larger than that of $\ln X$, it has a large impact on current-period profitability. Differences in Z manifest themselves significantly in profitability sorted portfolios, with a spread of 2.11 between the high and low profitability portfolios. Due to the lack of persistence, however, Z has a much smaller impact on the market-to-book ratio. The difference between the average $\ln Z$ of the growth portfolio and that of the value portfolio is only 0.23.

Fourth, consistent with the empirical evidence we present in Figure 2, the market-to-book sorted portfolios differ significantly in terms of average investment rate, while the spread between the investment rate in profitability sorted portfolios is much smaller by comparison. In our model, because market-to-book sorting identifies the permanent component of the productivity shock, X, and investment responds strongly to permanent productivity shocks, the average investment rate increases from 7% per year to 24% per year from value to growth portfolios. In contrast, because profitability sorted portfolios do not have a significant difference in their average $\ln X$, the investment rate also does not vary significantly across portfolios. Overall, the above results, therefore, confirm the empirical evidence we document in Section 2 and the asset pricing implications A1-A4 we list in Section 4.

5.4 The composition effect of risk premium

Our model is a general equilibrium one with heterogeneous firms and can be used to study the feedback mechanism between the cross-section and the aggregate economy. In our model, the cross-section distribution of firms is a slow-moving state variable and predicts aggregate stock market returns. In the cross section, firm-level expected returns are increasing in k and Z. The composition effect also manifests itself in time series: whenever high k firms compose a higher fraction of market capitalization than low k firms, the overall market expected return is high. Similarly, whenever the relative weight of high-Z firms and low-Z firms increases, so does the market expected return.

In our formulation, firm heterogeneity is summarized by the high-dimensional state variable m. Our solution method allows us to provide accurate descriptions of the equilibrium dynamics of the distribution m and study the above feedback mechanism between cross-section and aggregate time series. To illustrate this composition effect, in Table 6 we report return predictability regressions where we use variables constructed from the distribution m to predict model-implied expected return over horizons from one year to six years. We construct two distribution related variable that reflect this composition effect: the relative share of high k firms is computed as the ratio of the average market value of the high k firms (top tercile) over low k firms (bottom tercile), and the relative share of high Z firms is the ratio of the average market value of the high Z firms (top tercile) over low Z firms (bottom tercile).

Both the relative shares of high k firms and high Z firms have strong predictive powers for expected returns. The R^2 from the regression using the relative share of high k and high Z firms are around 76% and 11%, respectively. This result implies that the composition effect accounts for almost all of the return predictability in our model. In fact, most of the predictability comes from the general equilibrium effect of distribution on the stochastic discount factor. When we regress conditional volatility of stochastic discount factor on the relative shares of high k or high Z firms in the model, we obtain very similar R^2 as those in Table 6. We do not report t-statistics in Table 6 because we have a large number of simulations and all coefficients are highly significant.

Below we present empirical evidence that supports the above composition effect implied by our model. As we demonstrate before, book-to-market ratio is a good empirical proxy for the state variable k and the profitability ratio is a good empirical proxy for Z. In Table

Table 6: Distribution and risk premium in the model

This table shows the return predictive regressions using the relative high k share and relative high Z share in the model. The predictive regression is

$$E_t[R^M_{t\to t+h}] - R^f_{t\to t+h} = \text{const} + b \cdot \text{relative share}_t + \varepsilon_{t,t+h},$$

where h is the horizon, $E_t[R^M_{t\to t+h}] - R^f_{t\to t+h}$ is the cumulative expected market excess return from period t to $t+h$. We firstly sort firms into tercile portfolios each year based on firms' state k and Z, respectively. The relative share of high k firms is the ratio of the average market value of high k firms over low k firms. The relative share of high Z firms is the ratio of the average market value of high Z firms over low Z firms.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
Relative share of high k firms	0.003	0.006	0.009	0.012	0.014	0.017
R^2	0.77	0.77	0.77	0.76	0.76	0.75
Relative share of high Z firms	0.001	0.002	0.003	0.004	0.005	0.006
R^2	0.11	0.11	0.11	0.11	0.11	0.11

7, we report predictability regressions for market returns over horizons of one to six years where we approximate the relative shares of high-k and high-Z firms in the data using the relative shares of value and high profitability firms, respectively. To construct the relative share of value firms and the relative share of high-profitability firms, we firstly sort firms into tercile portfolios each year based on their book-to-market ratio and gross profitability, respectively. We define the relative share of value firms as the ratio of the average asset value of value firms (in the top tercile of book-to-market sorted portfolios) to that of growth firms (in the bottom tercile of the book-to-market sorted portfolios). Similarly, we define the relative share of high-profitability firms as the ratio of the average asset value of top-tercile high profitability firms to that of bottom-tercile low-profitability firms. The asset value of a firm in the data is sum of its market capitalization and total value of debt.

Clearly, as shown in the above table, both relative value share and relative profitability share have significant predictive powers for the market excess returns, even after controlling for conventional predictive variables such as the dividend yield, market volatility, and credit spread. The t-statistics are quite significant across the predictability regressions of almost all horizons from one to six years.

Table 7: Return predictability in the data

This table shows the return predictive regressions using the relative shares of value and profitable firms in the data. The predictive regression is

$$R^M_{t\to t+h} - R^f_{t\to t+h} = \text{const} + b \cdot \text{relative share}_t + c \cdot \text{controls}_t + \varepsilon_{t,t+h},$$

where h is the horizon, and $R^M_{t\to t+h} - R^f_{t\to t+h}$ is the cumulative market excess return from period t to $t+h$. We firstly sort firms into tercile portfolios each year based on their market-to-book and gross profitability, respectively. The relative share of value firms is the ratio of the average asset value of the value firms over growth firms. The relative share of profitable firms is the ratio of the average asset value of the profitable firms over unprofitable firms. The asset value of a firm is the sum of its market capitalization and debt. We also control for dividend yield, market volatility, and credit spread. The sample period is from 1963 to 2020. We report t-statistics in parenthesis, which are Newey-West adjusted allowing for 8 lags.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
Relative share of value firms	0.015	0.077	0.109	0.122	0.126	0.151
(t)	(0.70)	(2.39)	(3.12)	(3.61)	(3.23)	(2.77)
R^2	0.06	0.15	0.11	0.15	0.20	0.21
Relative share of profitable firms	0.072	0.146	0.173	0.217	0.277	0.290
(t)	(3.95)	(5.23)	(4.44)	(3.85)	(3.93)	(3.30)
R^2	0.14	0.26	0.19	0.25	0.33	0.29

5.5 Double sorted portfolios

In this section, we examine our model's ability to account for the expected returns on market-to-book and profitability ratio double sorted portfolios. Empirically, as shown by Novy-Marx (2013), controlling for profitability can significantly improve the value premium, and controlling for book-to-market can significantly improve the spread between high and low profitability sorted portfolios. This is an important aspect of the puzzle to address, because it presents an additional challenge for production-based asset pricing models.

We perform the same double-sorting procedure as in Novy-Marx (2013) using model-simulated data. We form portfolios by sorting firms independently into quintiles based on gross profitability and market-to-book ratio. The value-weighted average excess returns of the double sorted portfolios are presented in Table 8. Panel A reports the results using the empirical data between 1963 and 2020, and panel B reports the results from the model simulation.

As shown in Panel B, after controlling for value, the profitability strategy earns a

Table 8: Double sorted portfolios using model-simulated data

This table shows the value-weighted average excess returns of portfolios independently double sorted on gross profitability and market-to-book ratio. We also report the average returns of high-minus-low portfolios on each dimension. Panel A reports the results from the model-simulated data, Panel B reports the results from the empirical data. For the model-simulated excess returns, we assume a financial leverage of two. The t-statistics are in parenthesis, which are Newey-West adjusted allowing for 3 lags.

Panel A: Data

	GP/A1	GP/A2	GP/A3	GP/A4	GP/A5	H-L	(t)
MB1	8.44	11.55	13.06	12.08	12.32	4.13	(1.72)
MB2	5.49	7.54	11.60	13.81	11.07	5.58	(2.35)
MB3	6.26	6.04	8.11	10.13	9.83	3.57	(1.63)
MB4	3.81	4.85	6.04	8.88	9.17	5.35	(2.12)
MB5	1.46	3.12	4.91	4.66	8.98	7.52	(2.88)
H-L	-6.98	-8.43	-8.36	-7.55	-3.65		
(t)	(-2.21)	(-3.41)	(-2.82)	(-2.60)	(-1.22)		

Panel B: Model

	GP/A1	GP/A2	GP/A3	GP/A4	GP/A5	H-L	(t)
MB1	4.48	5.99	5.52	7.93	10.41	5.94	(9.11)
MB2	3.65	2.29	4.89	6.97	7.11	3.46	(7.18)
MB3	2.06	2.59	3.75	3.72	4.96	2.90	(7.69)
MB4	0.99	1.99	2.24	3.48	4.49	3.54	(10.87)
MB5	0.48	-1.32	-0.69	-2.82	0.13	0.81	(2.19)
H-L	-4.74	-7.01	-7.70	-10.75	-10.36		
(t)	(4.99)	(8.31)	(6.70)	(20.49)	(38.12)		

higher average return than the same strategy for univariate sorted portfolios. The average profitability spread across market-to-book ratio quintiles is 3.33% per annum (the second to last column), which is significantly higher than the unconditional profitability premium, 2.27%, as shown in Table 5. The same pattern holds for the value spread. As shown in Panel B of Table 8, the average value-growth spread across the five profitability quintiles is 8.11% (second to last row), which is higher than the value premium for univariate sorted portfolios, 6.85%.

The above pattern highlights the economic mechanism of our model. In our model, the most efficient way to generate return spreads is to sort directly on the two components of productivity, $\ln X$ and $\ln Z$, respectively. The market-to-book ratio and the gross profitability are noisy measures of $\ln X$ and $\ln Z$, and they contaminate each other. Firms with higher market-to-book ratios have higher permanent components of productivity on average. As a result, they have lower average returns than value firms due to the role of adjustment costs, as we discuss in Section 4. However, higher levels of the transitory component of productivity, Z, will also increase the market-to-book ratio. Because high Z is associated with a higher risk exposure to the capital depreciation shock, this effect dampens the value premium in univariate sorted portfolios. The same dampening mechanism is true for the univariate sorted profitability portfolios. Double sorting allows our model to separate the permanent component $\ln X$ from the transitory component $\ln Z$ and create a stronger spread in expected returns. In fact, in our model, the spread in the average $\ln X$ between value and growth portfolios is significantly higher for the double sorted portfolios than that for book-to-market single sorted portfolios. Similar patterns hold true for the profitability ratio sorted portfolios as well.

6 Conclusion

In this paper, we develop a production-based general equilibrium model with heterogeneous firms to jointly explain the value and profitability premiums. We emphasize the importance of a two-factor structure of productivity shocks to distinguish value and profitability as distinct firm-level characteristics. In addition, our model features two independent sources of aggregate shocks: productivity shocks and capital depreciation shocks. In equilibrium, the value factor endogenously loads more on productivity shocks, and the profitability factor loads more on capital depreciation shocks. Our model not only matches the key features of macroeconomic quantities and the dynamics of aggregate asset market, but more importantly, it captures well the coexistence of the value and profitability premium, as well as the different investment behaviors of value and growth sorted portfolios.

Our model features a general equilibrium setup, and we develop a method to numerically solve the model by exploiting the analytical expressions of the law of motion of the cross-section distribution of firms. This allows us to identify an important general equilibrium implication of return predictability from the cross-section distribution of firms. We show that consistent with the data, in our model, the relative weights of value and growth firms and that of high- and low- profitability firms have strong predictive powers for aggregate stock market returns.

References

Ai, H. and A. Bhandari (2021): "Asset Pricing With Endogenously Uninsurable Tail Risk," Econometrica, 89, 1471–1505.

Ai, H. and D. Kiku (2013): "Growth to value: Option exercise and the cross section of equity returns," Journal of Financial Economics, 107, 325–349.

Atkeson, A. and P. J. Kehoe (2005): "Modeling and Measuring Organization Capital," Journal of Political Economy, 113, 1026–53.

Bai, H., K. Hou, H. Kung, E. X. Li, and L. Zhang (2019): "The CAPM strikes back? An equilibrium model with disasters," Journal of Financial Economics, 131, 269–298.

Ball, R., J. Gerakos, J. T. Linnainmaa, and V. Nikolaev (2016): "Accruals, cash flows, and operating profitability in the cross section of stock returns," Journal of Financial Economics, 121, 28–45.

Belo, F., X. Lin, and S. Bazdresch (2014): "Labor Hiring, Investment, and Stock Return Predictability in the Cross Section," Journal of Political Economy, 122, 129–177.

Brunnermeier, M. K. and Y. Sannikov (2014): "A macroeconomic model with a financial sector," American Economic Review, 104, 379–421.

Byun, S. K., V. Polkovnichenko, and M. J. Rebello (2019): "Composition of Cash Flow Shocks, Firm Investment, and Cash Holdings," Available at SSRN 2852545.

Chen, A. Y. (2018): "A general equilibrium model of the value premium with time-varying risk premia," The Review of Asset Pricing Studies, 8, 337–374.

Cooper, R. W. and J. C. Haltiwanger (2006): "On the Nature of Capital Adjustment Costs," The Review of Economic Studies, 73, 611–633.

Croce, M. (2014): "Long-run productivity risk: A new hope for production-based asset pricing?" Journal of Monetary Economics, 66, 13–31.

Croce, M. M. (2014): "Long-run productivity risk: A new hope for production-based asset pricing?" Journal of Monetary Economics, 66, 13–31.

David, J. M. and V. Venkateswaran (2019): "The Sources of Capital Misallocation," American Economic Review, 109, 2531–67.

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

Donangelo, A., F. Gourio, M. Kehrig, and M. Palacios (2019): "The cross-section of labor leverage and equity returns," Journal of Financial Economics, 132, 497–518.

Dou, W., Y. Ji, and W. Wu (2020): "The Oligopoly Lucas Tree: Consumption Risk and Industry-Level Risk Exposure," working paper.

Dou, W. W., Y. Ji, and W. Wu (2021): "Competition, profitability, and discount rates," Journal of Financial Economics, 140, 582–620.

Eisfeldt, A. l. and D. Papanikolaou (2013): "Organization Capital and the Cross-Section of Expected Returns," The Journal of Finance, 68, 1365–1406.

Epstein, L. G. and S. E. Zin (1989): "Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework," Econometrica, 57, 937–969.

Fama, E. F. and K. R. French (2015): "A five-factor asset pricing model," Journal of Financial Economics, 116, 1–22.

Favilukis, J. and X. Lin (2015): "Wage rigidity: A quantitative solution to several asset pricing puzzles," Review of Financial Studies, 29, 148–192.

Favilukis, J., X. Lin, and X. Zhao (2020): "The Elephant in the Room: The Impact of Labor Obligations on Credit Markets," American Economic Review, 110, 1673–1712.

Garlappi, L. and Z. Song (2017): "Capital utilization, market power, and the pricing of investment shocks," Journal of Financial Economics, 126, 447–470.

Gertler, M. and P. Karadi (2011): "A Model of Unconventional Monetary Policy," Journal of Monetary Economics, 58, 17–34.

Gertler, M. and N. Kiyotaki (2010): "Financial intermediation and credit policy in business cycle analysis," Handbook of Monetary Economics, 3, 547–599.

Gomes, J., L. Kogan, and L. Zhang (2003): "Equilibrium Cross Section of Returns," Journal of Political Economy, 111, 693–732.

Gomes, j. f. and l. Schmid (2021): "Equilibrium Asset Pricing with Leverage and Default," The Journal of Finance, 76, 977–1018.

Gourio, F. (2008): "Estimating Firm-Level Risk," working paper.

Greenwood, J., Z. Hercowitz, and G. W. Huffman (1988): "Investment, Capacity Utilization, and the Real Business Cycle," The American Economic Review, 78, 402–417.

Grigoris, F. and G. Segal (2020): "The Utilization Premium," working paper.

Harvey, A. C. (1985): "Trends and cycles in macroeconomic time series," Journal of Business & Economic Statistics, 3, 216–227.

He, Z. and A. Krishnamurthy (2013): "Intermediary Asset Pricing," American Economic Review, 103, 732–70.

Herskovic, B., T. Kind, and H. Kung (2018): "Micro uncertainty and asset prices," Available at SSRN 3220825.

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and manufacturing TFP in China and India," The Quarterly journal of economics, 124, 1403–1448.

İmrohoroğlu, A. and Ş. Tüzel (2014): "Firm-level productivity, risk, and return," Management Science, 60, 2073–2090.

Jaimovich, N. and S. Rebelo (2009): "Can News about the Future Drive the Business Cycle?" American Economic Review, 99, 1097–1118.

Kogan, L., J. Li, and H. Zhang (2020): "Operating Hedge and Gross Profitability Premium," working paper.

Kogan, L. and D. Papanikolaou (2012): "Economic Activity of Firms and Asset Prices," Annual Review of Financial Economics, 4, 1–24.

——— (2013): "Firm Characteristics and Stock Returns: The Role of Investment-Specific Shocks," The Review of Financial Studies, 26, 2718–2759.

——— (2014): "Growth opportunities, technology shocks, and asset prices," Journal of Finance, 69, 675–718.

Kozlowski, J., L. Veldkamp, and V. Venkateswaran (2018): "The Tail that Keeps the Riskless Rate Low," .

Krusell, P. and A. A. Smith, Jr (1998): "Income and wealth heterogeneity in the macroeconomy," Journal of Political Economy, 106, 867–896.

Kuehn, L.-A. and L. Schmid (2014): "Investment-Based Corporate Bond Pricing," The Journal of Finance.

Kuehn, L.-a., M. Simutin, and J. J. Wang (2017): "A Labor Capital Asset Pricing Model," The Journal of Finance, 72, 2131–2178.

Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin (1992): "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" Journal of econometrics, 54, 159–178.

Kydland, F. E. and E. C. Prescott (1982): "Time to build and aggregate fluctuations," Econometrica, 1345–1370.

Li, J. (2018): "Explaining Momentum and Value Simultaneously," Management Science, 64, 4239–4260.

Luttmer, E. G. (2007): "Selection, growth, and the size distribution of firms," Quarterly Journal of Economics, 122, 1103–1144.

Novy-Marx, R. (2013): "The other side of value: The gross profitability premium," Journal of Financial Economics, 108, 1–28.

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," Econometrica, 64, 1263–1297.

Papanikolaou, D. (2011): "Investment shocks and asset prices," Journal of Political Economy, 119, 639–685.

Tong, J. and C. Ying (2020): "A Dynamic Agency Based Asset Pricing Model with Production," working paper.

Zhang, L. (2005): "The value premium," Journal of Finance, 60, 67–103.

Appendix

A Decomposition of firm-level productivity

A.1 Estimating firm-level productivity

We first construct a measure of productivity at the firm level by assuming a profit function of the Cobb-Douglas form

$$\log(\Pi_{j,t}) = \text{constant} + \ln TFP_{j,t} + \nu \log K_{j,t} + \epsilon_{j,t},$$

where $\ln TFP_{j,t}$ is the productivity of firm i at time t, $K_{j,t}$ is its capital stock, ν is the capital share, and $\varepsilon_{j,t}$ is the measurement error. This specification is consistent with the operating profits from the firm's optimal condition, as shown in equation (14). We use Compustat item PPENT for capital. Following Ball et al. (2016), we use operating profit for profit $\Pi_{j,t}$.[13]

However, by taking the Solow residuals, the productivity $a_{j,t}$ cannot be separated from the measurement error $\epsilon_{j,t}$. The key difference between $a_{j,t}$ and $\epsilon_{j,t}$ is that the former is a state variable that affects the firm's optimal decisions, whereas the latter does not. Following the industrial organization literature (e.g., Olley and Pakes (1996) and De Loecker et al. (2020)), we first project the observed operating profits on a set of polynomials based on capital $K_{j,t}$ and an instrument variable $\zeta_{j,t}$ to obtain the expected profits. This step helps us to remove the measurement error:

$$\log(\Pi_{j,t}) = c_0 + \sum_{m=0}^{2} \sum_{n=0}^{2-n} b_{m,n} (\log K_{j,t})^m (\log \zeta_{j,t})^n + e_{j,t}.$$

Following De Loecker et al. (2020), we use variable cost as the instrument variable. We define the Solow residual as

$$\begin{aligned} \text{Solow}_{j,t} &= \widehat{\log(\Pi_{it})} - \nu \log(K_{j,t}) \\ &= \widehat{c_0} + \sum_{m=0}^{2} \sum_{n=0}^{2-n} \widehat{b_{m,n}} (\log K_{j,t})^m (\log \zeta_{j,t})^n - \nu \log K_{j,t}. \end{aligned}$$

[13] In the model we developed in Section 3, the profit function, defined as total revenue less labor income, takes the same Cobb-Douglas functional form. We use profit function rather than production function for imputing firm-level productivity because the Compustat database does not have good coverage of employment data at the firm level.

We use the calibrated value $\nu = 0.85$, as in our model.[14]

To take into account the aggregate trend in the economy, we regress firm-level Solow residuals on a common linear trend. Therefore, the firm-level productivity $\ln TFP_{j,t}$ is obtained by

$$\text{Solow}_{j,t} = c \cdot t + \ln TFP_{j,t}.$$

A.2 Unobserved components decomposition

In Section 2.2, we assume for each firm j, its productivity takes the following form:

$$\begin{aligned}
\ln TFP_{j,t} &= \ln X_{j,t} + \ln Z_{j,t}, \\
\ln X_{j,t} &= \mu_j + \ln X_{j,t-1} + \sigma_{X,j}\varepsilon_{j,t}, \\
\ln Z_{j,t} &= \rho_{Z,j} \ln Z_{j,t-1} + \sigma_{Z,j}\eta_{j,t},
\end{aligned}$$

where $\ln TFP_{j,t}$ is the observed productivity series in logarithm, $\ln X_{j,t}$ is the unobserved trend component, which is assumed to be a random walk with average growth rate μ_j, and $\ln Z_{j,t}$ is the unobserved stationary component. To ensure that our decomposition is consistent with the productivity specification in our general equilibrium model, we assume that the innovations to the trend and transitory components are independent, thus $cov(\varepsilon_{j,t}, \eta_{j,t+s}) = 0, \forall s$.

We rewrite the productivity process into the state space representation:

$$\ln TFP_{j,t} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \ln X_{j,t} \\ \ln Z_{j,t} \end{pmatrix}, \tag{A1}$$

$$\begin{bmatrix} \ln X_{j,t} \\ \ln Z_{j,t} \\ \mu_j \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & \rho_{Z,j} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \ln X_{j,t-1} \\ \ln Z_{j,t-1} \\ \mu_j \end{bmatrix} + \begin{bmatrix} \sigma_{X,j}\varepsilon_{j,t} \\ \sigma_{Z,j}\eta_{j,t} \\ 0 \end{bmatrix}, \tag{A2}$$

where equation (A1) is the observation equation and equation (A2) is the state equation. To account for the firm-level fixed effect, we estimate the parameters separately for each firm. We estimate this state space model using the unobserved component package in `statsmodels`.

[14]We do not jointly estimate the capital share in profit function ν. The industrial organization literature focuses on estimating the production function with productivity following a Markov process. As in Section 2.2, we are interested in a two-factor structure in productivity, which contains both a transitory (AR(1)) and a permanent (random walk) component. Therefore, the total productivity no longer follows a Markov process. Fully estimating the production function of a non-Markov process in productivity is beyond the scope of this paper.

A.3 Nonstationarity of firm-level productivity

To empirically show that the firm-level productivity contains a permanent component, we perform two tests: the augmented Dickey-Fuller (ADF) test and the KPSS test proposed by Kwiatkowski et al. (1992). Both tests are used to test if a time series contains a unit root process. The null hypothesis of the ADF test is that the time series contains a unit root process. However, many firms have a relatively small sample size. The small sample bias of the ADF test leads us to accept that there is a unit root process too often. To address this issue, we also perform the KPSS test, in which the null hypothesis is that the time series is stationary. The small sample bias of the KPSS test is toward accepting stationarity. By combing the results from the ADF and KPSS tests, we can have a balanced view of the stationarity of the firm-level productivity process.

We perform both the ADF and KPSS tests, and the summary statistics of the p-values are reported in Table A.9. We focus on the case that we restrict firms to have at least 30 observations of productivity. The results of the ADF test suggest that for 95.58% of the firms, we cannot reject the null hypothesis that the firm-level productivity is nonstationary at the 1% level. The KPSS tests suggest a similar conclusion that for 99.99% of the firms, we can reject the null hypothesis that the productivity is stationary. Our results are robust if we perform these two tests on firms with fewer observations.

Table A.9: Tests of nonstationarity of firm-level productivity

This table presents the results of the ADF and KPSS tests. We report the percentage of the times that the ADF test is not rejected and the KPSS test is rejected for different confidence levels, 1%, 5%, and 10%, respectively. The null hypothesis for the ADF test is that the time series contains a unit root process. The null hypothesis for the KPSS test is that the time series is stationary. The tests are performed for each firm's productivity when the number of observations exceeds certain thresholds.

ADF			KPSS				
1%	5%	10%	1%	5%	10%	Minimum observations	Number of firms
95.58	90.07	85.30	58.18	82.15	99.99	30	1833
91.30	85.04	79.97	30.27	66.87	99.99	15	5739

B Proof of Propositions

B.1 Proof of proposition 1

To prove the homogeneity property of firm dividend and valuation, we start by showing that the operating profit function Π has a similar property. The operating profit function is given by

$$\Pi(K_t, Z_t, X_t; S_t) = \alpha Y(K_t, Z_t, X_t; S_t),$$

which can be rewritten as

$$\Pi(K_t, Z_t, X_t; S_t) = \alpha(A_t X_t Z_t)^{1-\nu}(uK)^{\nu}\Psi(S_t)^{\alpha-1} = A_t X_t \pi(k_t, Z_t; S_t),$$

where the normalized operating profit function

$$\pi(k_t, Z_t; S_t) = \alpha\left[Z_t^{1-\nu}(u_t k_t)^{\nu}\right]\Psi(S_t)^{\alpha-1},$$

Therefore, operating profit function satisfy the homogeneity property.

A firm's dividend is given by the budget constraint 10. Due to the homogeneity property of both the operating profit Π and convex adjustment H, the budget constraint 10 can be written as

$$d_t A_t X_t + i_t A_t X_t + H(i_t, k_t) A_t X_t \leq \pi(k_t, Z_t; S_t) A_t X_t,$$

This gives rise to the normalized firm budget constraint

$$d_t + i_t + H(i_t, k_t) + \xi k_t \leq \pi(k_t, Z_t; S_t) \tag{B3}$$

Finally, the normalized firm value $v(k_t, Z_t; S_t)$ follows as the cash-flow is homogeneous of degree one with respect to $A_t X_t$ and is given by

$$v(k_t, Z_t; S_t) = \max_{i_t, u_t, d_t} d_t + E\left[M_{t+1} v(k_{t+1}, Z_{t+1}; S_{t+1}) e^{\mu_A + \zeta_t + \epsilon^A_{t+1} + \mu + \sigma_x \epsilon_{t+1}}\right], \tag{B4}$$

B.2 Proof of proposition 2 and 3

a. The lagrangian associated with the dynamic programming problem 9 is

$$\mathcal{L} = E_t\left[\sum_{s=1}^{\infty}(1-\kappa_D)^s M_{t,t+s}\Big(\Pi_{t+s} - I_{t+s} - H(I_{t+s}, K_{t+s}) + \tilde{q}_{t+s}\left((1-\delta(u_{t+s-1}, \theta_t) K_{t+s-1} + I_{t+s-1} - K_{t+s})\right)\Big)\right], \tag{B5}$$

where the Lagrangian multiplier for the capital accumulation equation is denoted as $\tilde{q}_t$.

- first order condition for investment I_t

$$1 + H_I(I_t, K_t) = (1-\kappa_D)E_t\big[M_{t+1}\tilde{q}_{t+1}\big], \tag{B6}$$

- first order condition for utilization u_t

$$\Pi_{u,t} = (1-\kappa_D)u_t^{\lambda}E_t\big[M_{t+1}\theta_{t+1}\tilde{q}_{t+1}\big], \tag{B7}$$

- first order condition for K_{t+1}

$$\tilde{q}_{t+1} = \frac{\partial \Pi_{t+1}}{\partial K_{t+1}} - H_K(I_{t+1}, K_{t+1}) + (1-\kappa_D)E_{t+1}\big[M_{t+1,t+2}(1-\delta(u_{t+1},\theta_{t+2}))\tilde{q}_{t+2}\big], \tag{B8}$$

where the marginal product of capital MPK is given by $\frac{\partial \Pi}{\partial K}$. The economic interpretation of Lagrangian multiplier $\tilde{q}_t$ is that it is the marginal value of capital at the beginning of the period t before depreciation. Its net present value is the marginal Q, as shown in equation (B6). Combine equation (B6) and (B8) we have that

$$\begin{aligned} 1 + H_I(I_t, K_t) =& (1-\kappa_D)E_t\Big\{M_{t+1}\Big(\frac{\partial \Pi_{t+1}}{\partial K_{t+1}} - H_K(I_{t+1}, K_{t+1})\Big) \\ & +(1-\kappa_D)E_{t+1}\big[M_{t+1,t+2}\big(1-\delta(u_{t+1},\theta_{t+2})\big)\tilde{q}_{t+2}\big]\Big\}, \end{aligned} \tag{B9}$$

Use the law of iterated expectation iterate q forward using equation (B8), we have that

$$1 + H_I(I_t, K_t) = E_t\Big[\sum_{s=1}^{\infty}\hat{\beta}_s M_{t,t+s} MPK_{t+s}\Big] - E_t\Big[\sum_{s=1}^{\infty}\hat{\beta}_s M_{t,t+s} H_K(I_{t+s}, K_{t+s})\Big],$$

where

$$\hat{\beta}_s = \begin{cases} (1-\kappa_D), & \text{when s} = 1 \\ (1-\kappa_D)^s \prod_{\tau=1}^{s-1}\Big(1-\delta(u_{t+\tau},\theta_{t+\tau+1})\Big), & \text{when s} \geq 2 \end{cases}$$

Because the convex adjustment cost function is homogeneous of degree 1, therefore $H_I(I_t, K_t) = H_i(i_t, k_t)$ and $H_K(I_t, K_t) = H_k(i_t, k_t)$. This completes the proof for part a. of Proposition 2. □

b. To prove the relation between investment and firm value, we multiply K_{t+1} on both sides of equation (B6), this gives

$$
\begin{aligned}
\big[1 + H_I(I_t, K_t)\big] K_{t+1} =&(1-\kappa_D)E_t\Bigg[M_{t+1}\bigg(\frac{\partial \Pi_{t+1}}{\partial K_{t+1}}K_{t+1} - H_K(I_{t+1}, K_{t+1})K_{t+1} \\
&+(1-\kappa_D)E_{t+1}\big[M_{t+1,t+2}(1-\delta(u_{t+1}, \theta_{t+2}))\tilde{q}_{t+2}K_{t+1}\big]\bigg)\Bigg] \\
=&(1-\kappa_D)E_t\Bigg[M_{t+1}\bigg(\frac{\partial \Pi_{t+1}}{\partial K_{t+1}}K_{t+1} - H_K(I_{t+1}, K_{t+1})K_{t+1} \\
&+(1-\kappa_D)E_{t+1}\big[M_{t+1,t+2}\tilde{q}_{t+2}(K_{t+2} - I_{t+1})\big]\bigg)\Bigg],
\end{aligned}
$$

where we use the capital accumulation equation (7) to replace the term $(1-\delta(u_{t+1}, \theta_{t+2})K_{t+1})$. Apply equation (B6), the proof above can be written as

$$
\begin{aligned}
\big[1 + H_I(I_t, K_t)\big] K_{t+1} =&(1-\kappa_D)E_t\Bigg[M_{t+1}\bigg(\frac{\partial \Pi_{t+1}}{\partial K_{t+1}}K_{t+1} - H_K(I_{t+1}, K_{t+1})K_{t+1} \\
&-I_{t+1} - H_I(I_{t+1}, K_{t+1})K_{t+1} + \big[1 + H_I(I_{t+1}, K_{t+1})\big]K_{t+2}\bigg)\Bigg],
\end{aligned}
$$

Because investment adjustment cost $H(I, K)$ is homogeneous of degree 1, this can be further expressed as

$$
\begin{aligned}
\big[1 + H_I(I_t, K_t)\big] K_{t+1} =&(1-\kappa_D)E_t\Bigg[M_{t+1}\bigg(\Pi_{t+1} - I_{t+1} - H(I_{t+1}, K_{t+1}) + \frac{\partial \Pi_{t+1}}{\partial K_{t+1}}K_{t+1} - \Pi_{t+1} \\
&+\big[1 + H_I(I_{t+1}, K_{t+1})\big]K_{t+2}\bigg)\Bigg],
\end{aligned}
$$

Iterate $\big[1 + H_I(I_{t+1}, K_{t+1})\big]K_{t+2}$ forward and use the definition of firm value, we have

$$
\begin{aligned}
\big[1 + H_I(I_t, K_t)\big] K_{t+1} =&V_t - D_t + E_t\Bigg[\sum_{s=1}^{\infty}(1-\kappa_D)^s M_{t,t+s}\bigg(\frac{\partial \Pi_{t+s}}{\partial K_{t+s}}K_{t+s} - \Pi_{t+s}\bigg)\Bigg] \\
=&V_t - D_t - (1-\nu)E_t\Bigg[\sum_{T=1}^{\infty}(1-\kappa_D)^s M_{t,t+s}\Pi_{t+s}\Bigg].
\end{aligned}
$$

$\square$

Proposition 3 Using the optimal utilization equation (B7), the left-hand variable Π_u is

$$\Pi_{u,t} = \nu \frac{\Pi_t}{u_t}, \tag{B10}$$

Combine with the right hand side of optimal utilization equation, we have that

$$u_t^{1+\lambda} = \frac{\nu \Pi_t}{(1-\kappa_D)E[M_{t+1}\theta_{t+1}\tilde{q}_{t+1}]}.$$

□

C Normalized firm problem and the summary measure

Thanks to the homogeneity property of the problem, the aggregate state vector S_t can be further simplified and denoted as $s_t = (\zeta_t, \theta_t, m_t)$. Now the normalized aggregate state vector includes the predictable component of aggregate TFP ζ_t, the capital utilization shock θ_t, and an aggregate state variable m_t that summarizes and replaces the higher dimensional firm distribution Γ_t. We term m_t as the summary measure and we explain the construction of it below, after we introduce normalized firm maximization problem.

Given the normalized state vector s_t and its law of motion, the normalized firm value $v(k_t, Z_t; s_t)$ satisfies the following Bellman equation:

$$v(k_t, Z_t; s_t) = \max_{i_t, u_t, d_t} d_t + (1-\kappa_D)E\left[M_{t+1}v(k_{t+1}, Z_{t+1}; s_{t+1})e^{\mu_A+\zeta_t+\epsilon^A_{t+1}+\mu+\sigma_x\epsilon_{t+1}}\right], \tag{C11}$$

subject to

$$d_t + i_t + H(i_t, k_t) \leq \pi(k_t, Z_t; s_t), \tag{C12}$$

$$e^{\mu_A+\zeta_t+\epsilon^A_{t+1}+\mu+\sigma_x\epsilon_{t+1}}k_{t+1} = (1-\delta(u_t, \theta_{t+1}))k_t + i_t, \tag{C13}$$

and the depreciation function δ is given by equation (8).

For new entrant firms, their normalized value function is given by

$$\bar{v}(s_t) = \max_{\bar{k}_t} -\bar{k}_t - \bar{H}(\bar{k}_t) + (1-\kappa_D)E\left[M_{t+1}v(\bar{k}_t, Z_{t+1}; s_{t+1})e^{\mu_A+\zeta_t+\epsilon^A_{t+1}+\mu+\sigma_x\epsilon_{t+1}}\right]. \tag{C14}$$

Finally, we describe the construction of the aggregate state variable m which we refer to as the "summary measure". Let $\tilde{\Gamma}(k, Z, X)$ denote the joint distribution of (k, Z, X) for firm's state variable which is derived from the distribution Γ as follows,

$$\Gamma(K, Z, X) = \Gamma(k \cdot X, dZ, dX) = \tilde{\Gamma}(k, Z, X),$$

where $d\tilde{\Gamma}(\bar{X})$ denotes the distribution of new entrant firms' productivity. In general, $\tilde{\Gamma}$ is a state variable in the construction of a recursive competitive equilibrium because the aggregate resource constraint (C15) depends on the distribution $\tilde{\Gamma}$,

$$\begin{aligned} c(s) + \int \Big(i(k, Z; s) + H(i(k, Z; s), k)\Big) X d\tilde{\Gamma}(k, Z, X) + \int \Big(\bar{k}(s) + \bar{H}(k(s))\Big)\bar{X} d\tilde{\Gamma}(\bar{X}) \\ = \int \Big[u(k, Z; s)^{\nu} k^{\nu} Z^{1-\nu}\Big]^{\alpha} X d\tilde{\Gamma}(k, Z, X), \end{aligned} \tag{C15}$$

where $c(s) = \frac{C}{A}$ is the normalized aggregate consumption.

The aggregate investment, including adjustment costs, can be written as

$$\begin{aligned} &\int \Big(i(k, Z; s) + H(i(k, Z; s), k)\Big) X d\tilde{\Gamma}(k, Z, X) \\ &\qquad = \int \Big(i(k, Z; s) + H(i(k, Z; s), k)\Big) \Big[\int X d\tilde{\Gamma}(X|k, Z)\Big] d\tilde{\Gamma}(k, Z), \end{aligned} \tag{C16}$$

where we decompose the joint distribution into a marginal distribution and a conditional distribution: $d\tilde{\Gamma}(k, Z, X) = d\tilde{\Gamma}(X|k, Z) \cdot d\tilde{\Gamma}(k, Z)$. We define the summary measure $m(k, Z)$ as

$$m(k, Z) = \int X d\tilde{\Gamma}(X|k, Z).$$

Thus, $m(k, Z)$ is the average amount of permanent productivity X for firms with (k, Z). Similarly, the initial amount of permanent productivity for new entrants is referred to as $m(\bar{k})$. The aggregate resource constraint (C15) can be written as

$$\begin{aligned} c(s) + \int \Big(i(k, Z; s) + H(i(k, Z; s), k)\Big) m(k, Z) d\tilde{\Gamma}(k, Z) + \Big(\bar{k}(s) + \bar{H}(k(s))\Big) m(\bar{k}(s)) \\ = \int \Big[u(k, Z; s)^{\nu} k^{\nu} Z^{1-\nu}\Big]^{\alpha} m(k, Z) d\tilde{\Gamma}(k, Z). \end{aligned} \tag{C17}$$

The summary measure m reduces the three-dimensional distribution Γ into a two-dimensional one. This greatly simplifies our analysis. Now we are ready to define a recursive competitive equilibrium with normalized variables and summary measures, which we solve numerically. A recursive competitive equilibrium consists of:

- A value function for the representative household's utility $\tilde{u}(s)$ and the associated consumption policy function $c(s)$, where $u(s) = \frac{\mathbf{U}(S)}{A}$ and $c(s) = \frac{\mathbf{C}(S)}{A}$.

- A value function for incumbent firms, $v(k, Z; s)$, and the associated policy functions, $i(k, Z; s)$ and $u(k, Z; s)$.

- A value function for new firms, $\bar{v}(s)$, and the associated policy function, $\bar{k}(s)$, where $\bar{v}(s) = \frac{\bar{V}(S)}{X}$ and $\bar{k}(s) = \frac{\bar{K}(S)}{X}$.

- A stochastic discount factor, $M(s'|s)$, which is given by

$$M(s'|s) = \beta e^{-\gamma(\mu_A+\xi+\epsilon^{A}\prime)} \left(\frac{c\,(s')}{c\,(s)}\right)^{-\frac{1}{\psi}} \left(\frac{\tilde{u}\,(s')^{1-\gamma}}{E\left[\tilde{u}\,(s')^{1-\gamma}\, e^{(1-\gamma)(\mu_A+\xi+\epsilon^{A}\prime)}\right]}\right)^{\frac{1/\psi-\gamma}{1-\gamma}}, \tag{C18}$$

- A low of motion of the summary measure m, that is given by

$$\begin{aligned} m'\left(k', Z'\right) = (1-\kappa_D) \int \mathbb{I}_{k'}(k, Z) \cdot \phi\left(\frac{\ln Z' - \rho_z \ln Z}{\sigma_z}\right) m(k, Z) \\ + \kappa_D \bar{\mathbb{I}}_{k'}(\bar{k}, \bar{Z}) \phi\left(\frac{\ln Z' - \rho_z \ln \bar{Z}}{\sigma_z}\right) m(\bar{k}(s)), \end{aligned} \tag{C19}$$

such that the following conditions are met:

1. given the equilibrium stochastic discount factor and law of motion of summary measure m, the value function $v(k, Z; s)$ and the policy functions, $i(k, Z; s)$ and $u(k, Z; s)$, solve the normalized maximization problem (C11), (C12), and (C13) for incumbent firms.

2. the value function $\bar{v}(s)$ and the policy function $\bar{k}(s)$ solve the profit maximization problem for new entrants (C14).

3. the equilibrium SDF is consistent with household consumption.

4. the law of motion of summary measure (C19) is consistent with firms' optimal policies

5. Market clearing conditions (C15) holds.

6. labor market clearing gives rise to the optimal labor allocation across firms.

D Data construction

D.1 Firm-level variables

We obtain firm-level balance sheet data from the Compustat Annual Database, for the period 1963 to 2020. Following standard practice in the literature, we exclude utility firms (SIC codes between 4900-4999) and financial firms (SIC codes between 6000-6999). Firm-level variables are deflated using the GDP deflator, obtained from the Bureau of Economic Analysis (BEA). Monthly data stock prices are obtained from CRSP database, from 1963 to 2020. Detailed definitions of variables are presented in Table D.10.

Table D.10: Variable definitions

Variables	Definition	Sources
BE	Book value of equity, computed as the book value of stockholders' equity, plus balance sheet deferred taxes and investment tax credit (if available), minus the book value of preferred stock. Depending on availability, we use the redemption, liquidation, or par value (in that order) as the book value of preferred stock.	Compustat
ME	Market value of equity is copmputed as price per share times the number of shares outstanding. The share price is taken from CRSP, the number of shares outstandings from Compustat or CRSP, depending on availability.	Compustat
BM	Book to market value of equity ratio.	Compustat
GP/A	Compustat item REVT minus COGS divided by AT.	Compustat
ΔSales	Growth rate of Compustat item SALE.	Compustat
I/K	Compustat item CAPX minus SPPE divided by PPEGT.	Compustat
ΔAT	Growth rate of Compustat item AT.	Compustat
ΔCF	Growth rate of Compustat item EBITDA.	Compustat
Operating profits	Compustat item REVT minus the sum of COGS and XSGA. See Ball et al. (2016).	Compustat

D.2 Aggregate capital depreciation rates

We use the flow budget constraint for capital to compute the aggregate depreciation rates,

$$\delta_t = \frac{K_t + I_t - K_{t+1}}{K_t},$$

where I_t is the real investment and K_t is the real capital stock. Real investment is defined as the private nonresidential investment divided by the investment goods deflator. Real aggregate capital is the private nonresidential fixed assets deflated by the investment goods deflator.