Essays in Finance



Bingcun Dai
Green Templeton College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity Term 2023

Abstract

This thesis consists of three studies in asset pricing. There are three principles in financial economics which well known even by non-economists: (1) stock return and risk should be positively correlated; (2) people's happiness should depend on their own consumption, and not on other people's wealth, in other words, wealth inequality should not matter if people can think and behave rationally; (3) stock prices only reflect agents' rational forecasts of future cashflows, so should not affect the real economy. Although these claims sound very reasonable, why do they often contradict to how we feel, or what financial economics found? My three thesis chapters respectively answer these three questions, and with one common ingredient: heterogeneity. The first chapter *Heterogeneous Beliefs and the Cross-section of Stock Returns*[1] theoretically and empirically shows that when disagreement about a stock return among investors is either very low or very high, expected stock return is positively correlated with its risk; when disagreement is moderate, the security market line is downward sloping. The second chapter *Indivisible Goods and Relative Wealth Concerns: Gambling, Under-Participation and Lotteries.*[2] introduces heterogeneity in divisibility of consumption goods in a model with an endowment economy. When some goods are indivisible like houses, people's utility rationally depends on wealth inequality, and the relatively rich invest more in risky assets than the poor, even while the poor exhibit a preference for gambles with positively skewed payoffs. The third chapter *The stock market isn't a sideshow: How granular instrumental variables reveal the effect of stock market volatility on the real economy.* shows that stock market returns have a large impact on aggregate GDP, consumption and investment growth, using a granular instrumental variable based on idiosyncratic flows from different investment sectors. These idiosyncratic flows reflect deviations of investors' own beliefs or preferences from the common component and are not correlated with common shocks to the aggregate economy, therefore they are good instruments for stock prices.

[1]This is joint work with Mungo Wilson.

[2]This is joint work with Pavel Savor and Mungo Wilson.

Contents

List of Figures **v**

List of Tables **vii**

1 Introduction **1**

2 Heterogeneous Beliefs and the Cross-section of Stock Returns **3**

2.1 Introduction 3

2.2 Model 8

2.3 Empirical Analysis 17

Data 17

Results 19

2.4 Conclusion 20

2.5 Appendix 22

3 Indivisible Goods and Relative Wealth Concerns: Gambling, Under-Participation and Lotteries. **33**

3.1 Introduction 33

3.2 Competition for indivisible goods 38

Inequality does not matter in an endowment economy with divisible goods 38

Inequality matters when some goods are indivisible 40

3.3 Gambles without aggregate risk 44

3.4 Aggregate risk 48

Aggregate risk and welfare 49
Aggregate risk and gambling 51
The choice of how much risk to bear 55
3.5 Lotteries 58
Populism 60
Cincinnati kid 61
Populist kid 63
3.6 Conclusion 65
3.7 Appendix 67
Proof of $p < q$ for all values in $(0, 0.5)$. 67
Proof that $Var[C_{2,1}] > Var[C_{1,1}]$ for $1 < k < 1 + 2P'(\widetilde{E}_{2,1})$. . . . 68
3.8 Figures 70

4 The stock market isn't a sideshow: How granular instrumental variables reveal the effect of stock market volatility on the real economy. 1
4.1 Introduction 1
4.2 Instrumental variables for stock market prices 5
Dividends 5
Why is GIV a good instrument for stock price to address endogeneity concern? 6
Data sources and construction 9
GIVs: Data and first-stage results 10
4.3 The real effects of stock market volatility 12
4.4 Effects through the financing channel 17
4.5 An alternative specification using dividend flows 19
4.6 Conclusion 21
4.7 Tables 23

5 Conclusion 33

Bibliography **37**

List of Figures

2.1 Stock Market Turnover 18

2.2 SML on Low-turnover Days. This figure plots average daily excess returns in bps against market betas for ten beta-sorted portfolios of all NYSE, Amex, and Nasdaq stocks (the solid points), and the implied ordinary least squares estimates of the SML (the solid line). The sample covers the 1964–2011 period. In addition, the SML predicted by the standard CAPM is displayed (the dotted line). 28

2.3 SML on Moderate-turnover Days. This figure plots average daily excess returns in bps against market betas for ten beta-sorted portfolios of all NYSE, Amex, and Nasdaq stocks (the solid points), and the implied ordinary least squares estimates of the SML (the solid line). The sample covers the 1964–2011 period. In addition, the SML predicted by the standard CAPM is displayed (the dotted line). 29

2.4 SML on High-turnover Days. This figure plots average daily excess returns in bps against market betas for ten beta-sorted portfolios of all NYSE, Amex, and Nasdaq stocks (the solid points), and the implied ordinary least squares estimates of the SML (the solid line). The sample covers the 1964–2011 period. In addition, the SML predicted by the standard CAPM is displayed (the dotted line). 30

2.5 Borrowing Constraints 31

3.1 43
3.2 70
3.3 70
3.4 71
3.5 72
3.6 73
3.7 74
3.8 74
3.9 75
3.10 75
3.11 76

List of Tables

2.1 Summary Statistics for Time-Series Variables 19

4.1 Summary Statistics 24
4.2 GIV and Stock Price 25
4.3 GDP 26
4.5 Nonresidential Investment 27
4.4 Consumption and Investment 30
4.6 Financing Channels 31
4.7 Dividend Payout and Change in Stock Price 31
4.8 Dividend Payment-Driven Returns 32

Chapter 1

Introduction

In this thesis I study the role of heterogeneity between agents in fairly simple economic settings. For decades, the study of economics studied representative agent models in model economies of varying degrees of complexity, from endowment to production economies. More recently, the literature began to focus on heterogeneous agent models as offering a key to understadning the puzzles which representative agent models could not explain.

In my first thesis chapter, I explore how heterogeneity in beliefs between investors in stocks, combined with a simple margin constraint on leveraged investment and short selling in shares can generate interesting and novel implications for the relationship between share trading volume and expected return. I show that the Securites Market Line of the Capital Asset Pricing Model is not straightforwardly behaved in settings where, inevitably given disagreement, some agents want to borrow to invest in stocks while other agents want to short sell the same stocks. I derive a general relationship between such disagreement, manifested in leverage, short interest, and trading volume and confirm some of my fidnings in the data. Previous studies focussed on either leverage (optimism relative ot the mean) or short selling (pessimism). To my knowledge, no study has unified these in one setting.

In my second chapter I take a different path. Some goods are in fact indivisible, and this fact has often been deliberately overlooked because the standar method of solving for optimal consumption and investment strategies requires

considering small changes near the optimum, which are non-feasible in the presence of indivisible goods. I show that heterogeneity in wealth delivers remarkable new predictions about how wealth inequality will impact preferneces for taking risk. In brief, the indriect utility functions of agents who compete for indivisible goods are those conjectured by Friedman and Savage way back in 1948. As a result, poorer agents take less aggregate risk (under-investing in stocks) while still wishing to insure against small risks (such as purchasing mobile phone insurance), but are very attracted to gambles with a positive skew, provided the positive state payoff is big enough. Thus the apparent impossibility of explaining why some agents under-invest in stocks, buy mobile phone insurance and yet also lottery tickets is resolved, prvoided those agents are relatively poor. Richer agents expect to get richer. This can provide some of the thoeretical underpinnings for the recent literatiure on capital and inequality.

In my final thesis chapter, I focus on how these differential demands for stocks, in a more behavioral, or at least more institutional, setting, can be used to deliver insights into how stock markets really matter in the twenty-first century. I show that different flows into stocks from differnet sectors of invesotrs holding stocks can be exploited to provide an instrument for stock market returns. Equipped with this instrument, I show that stock market returns, far from being passive responses to anticipated future events in the real economy, actually drive them. If anything, different flows from differentsectors appear to be the prime mover of economic, particularly investment, cycles. This finding has the potential to resolve an age-old mystery in real business cycle theory familiar to all economics students, namely, why is aggregate investment so volatile when it 'ought' to be very stable? The answer is, once again, heterogeneity in beliefs, preferences and financing needs, as shown to matter in my first two chapters.

My research agenda is to explore these ideas much more deeply in the rest of my career.

Chapter 2

Heterogeneous Beliefs and the Cross-section of Stock Returns

2.1 Introduction

The capital asset pricing model (CAPM) of Sharpe (1964) and Sharpe (1964) implies a cross-sectional relation between return and risk: expected returns on securities are a linear function of their market betas, and all cross-sectional variation in returns can be fully captured by betas, i.e. only systematic risk is priced. Thus, the relationship between risk and expected return is plotted on a single security market line (SML).

Some theoretical studies try to reconcile the deviations from the SML by relaxing the assumptions underlying the standard CAPM. One essential assumption is that investors can take long or short positions of any size in any asset, including risk-free asset. Black (1972) shows that when investors are prohibited from borrowing at risk-free rate, the slope of SML is smaller than in the unconstrained CAPM. Frazzini & Pedersen (2014) extend the Black (1972) model under a less restrictive condition that investors are allowed to leverage but face margin requirement, and show that the slope of the SML decreases in the the tightness of the leverage constraint. In the models with restricted borrowings, the constrained investors who desire high returns cannot lever up the tangency portfolio, and thus have to increase the holdings in riskier assets (high-beta stocks). The increase in demand for high-beta stocks bids up

prices, and in turn flattens the SML. Although those models yield a flatter SML relative to the CAPM, they cannot explain the empirical evidence that high-beta stocks deliver lower returns than low-beta stocks, which implies a downward-sloping SML (Black (1972) Baker et al. (2011) Frazzini & Pedersen (2014)).

On the other hand, to obtain a downward-sloping SML, Hong & Sraer (2016) construct a model with short-sales restrictions, i.e. constraints on borrowing risky assets rather than risk-free asset. Moreover, they relax the other CAPM assumption of complete agreement and allow agents to disagree about the common cash flows. High-beta assets are more sensitive to the aggregate disagreement and thus experience greater divergence of opinion. Thus, pessimists, who would take short positions in high-beta assets, are sidelined due to short-sales constraint. As a consequence, those assets are held only by optimists and hence overpriced compared to low-beta assets.

We find supporting evidence for the implication of the Hong & Sraer (2016) model, i.e. expected returns decrease with beta for high-beta assets (Figure 2.3). However, our finding is at odds with their implication when disagreement is very high: on days with very high disagreement (defined as the top decile of the market turnover), the SML runs very near to its CAPM prediction (Figure 2.4), contrary to the Hong & Sraer (2016)'s prediction of a downward-sloping SML. In other words, the market portfolio is *efficient* when disagreement is very high. If the market frictions (i.e. short-sales constraint) lead to the inefficiency of the market portfolio (i.e. the deviation from standard CAPM), it seems like a *puzzle* that, when the stock market experience extremely high disagreement and therefore short-sales constraint is more binding, the market portfolio is instead efficient.

In addressing the *"efficient frictions" puzzle* described above, this paper offers a theoretical framework incorporating heterogeneity and borrowing constraints on both risky assets and risk-free assets, i.e. short-sales constraint and leverage constraint. We model investors as being heterogeneous in their be-

liefs about expected payoffs and covariance matrix of risky assets, as in Jarrow (1980), Lintner (1969). Investors can use leverage (or short selling) but face a margin requirement that dictates the minimum value of collateral (i.e. part of investors' wealth) needed in a borrowing contract, limiting the maximum amount of borrowing. In addition to being realistic, it is necessary to assume margin-type constraints on these two types of borrowing, because borrowing constraints arise endogenously from belief disagreements (Simsek (2013)): the optimists who think assets are underpriced may want to borrow money from pessimists to take more positions; however, pessimists do not value the collateral (i.e. optimists' wealth) as much as optimists do, thus become reluctant to lend, which creates leverage constraint for optimists. Short-sales constraint arises in a similar fashion.

Our main result is that under consensus belief, the expected excess return on a risky asset relates to its expected market beta in the linear form as the standard CAPM, plus an extra term, which is related to the shadow price difference between leverage constraint and short-sales constraint. This extra term defined as the *distance-to-CAPM* explains the variation of the SML for different levels of disagreement. In particular, when leverage constraint and short-sales constraint bind to a similar degree (i.e. the *distance-to-CAPM* is close to zero), the SML runs very near to its CAPM prediction, which rationalize the *"efficient frictions" puzzle*. The intuition is as follows. The leverage-constrained investors who have a preference for high return will apply leverage to riskier (high-beta) stocks instead of the tangency portfolio. Similarly, the short-sales constrained investors will take more short positions in high-beta stocks to achieve a higher return. When the price pressure on high-beta stocks from leveraging is balanced by that from short selling, the stocks are efficiently priced as if there were no borrowing frictions.

The reason why the "efficient frictions" can be observed only when disagreement is very high is that, disagreement-induced leverage constraint is generally tighter than short-sales constraint. The source of the endogenous borrowing

constraints is that lenders want to avoid the potential loss that they would suffer if they satisfied borrowers' demand. Pessimistic lenders worry, that if borrowers default and give up the collateral (i.e. stocks), the drop of stock price will leads to a real reduction in their wealth, all of which can be lost when stock is worthless. In contrast, the nominal value of the collateral for optimistic lenders is guaranteed. Thus, the potential loss that optimistic lenders care about is not a real loss with a negative net income; instead, it is like a loss of lottery prize for betting correctly. Therefore, the potential loss for pessimistic lenders of more real impact, and thus they will would impose a tighter constraint (leverage constraint) on borrowers than optimists do. The investors who hold biased beliefs may bid up high-beta stock prices to the point of having lower returns than low-beta stocks, which implies a downward-sloping SML. When belief dispersion is very high, there is hardly borrowing between pessimists and optimists, which leads to maximum tightness of borrowing constraints, i.e. leverage constraint and short-sales constraint will bind to a similarly large degree.

Our model also implies that heterogeneous beliefs command a positive equity premium, consistent with the findings in Anderson et al. (2005), Banerjee & Kremer (2010), David (2008). We find that the shadow price difference between leverage constraint and short-sales constraint, which result endogenously from heterogeneous beliefs, is a determinant of equity premium. Two constraints have opposite effects on asset prices. One one hand, leverage-constrained investors would like to borrow cash and invest in equity but the constraint prevents them from doing so. Therefore, the leverage constraint decreases the demand for equity, and lowers stock price (increase stock return) as a result. On the other hand, when short-sales constraint is binding for some pessimists, equity prices will not reflect the valuations of those pessimists but reflect the valuations of relatively more optimistic investors, which pushes up stock price (lower stock return). As discussed earlier, leverage constraint is generally tighter than short-sales constraint, and thus net effect of borrowing

constraints is to increase the stock return.

Moreover, we show that the effect of disagreement on trading volume can be either positive or negative: trading volume increases in disagreement since investors with dispersed beliefs have high speculative demand; but disagreement also has a negative effect which limits the growth in trading volume, because the increase in disagreement tightens borrowing constraints and thus prohibits the amount of volume which would be traded if the constraints did not bind. Although it is difficult to obtain an analytical derivative of trading volume with respect to disagreement, given the empirical evidence (Atiase (1991), Bessembinder et al. (1996), Goetzmann & Massa (2005)), we could plausibly conjecture that the positive direct effect is stronger in general: trading volume increases in disagreement.

Our model yields the following key testable implications. (i) When disagreement is very low or very high, the standard CAPM holds because two constraints do not bind or bind to a similarly large degree. (ii) When disagreement is moderate, the effect of leverage constraint is dominant over short-sales constraint; the average leverage-constrained investor would strategically substitute low-beta stocks for high-beta stocks. Consequently, less risky assets will be underpriced relative to the standard CAPM (*distance-to-CAPM*>0), while riskier assets will be overpriced (*distance-to-CAPM*<0). When the net effect of borrowing constraint is large enough, the SML would be downward-sloping as in Hong & Sraer (2016).

Next, we conduct a preliminary empirical analysis to test our model predictions using the daily trading volume in the stock market as our measure of disagreement. Consistent with our theoretical predictions, we find that (i) on days of very low and very high stock market turnover, the empirical Security Market Line runs very near its CAPM prediction; (ii) in contrast, when turnover is not extreme, the empirical SML is above that predicted by the standard CAPM for low-beta assets and below it for high-beta assets.

The rest of the paper is organized as follows. In Section 2 we present the

model. Section 3 describes the data and shows the empirical results. We conclude in Section 4.

2.2 Model

A. Securities Market

Consider a market with one risk-free rate asset and K risky assets. There are two periods, $t = 0$ and $t = 1$. At $t = 0$, let the price of the risk-free asset be 1 and the price of risky assets be $P = (p_1, \cdots, p_K)^T$. At $t = 1$, the payoff of the risk-free asset is $R_f = 1 + r_f$ and the payoff of risky assets are $X = (x_1, \cdots x_k, \cdots, x_K)^T$.

B. Investors

There are I investors in the market, indexed by $i = 1, 2, \cdots, I$. The investors differ in their initial endowments, risk aversion, and probability assessments of prices which are joint normally distributed. Let the belief of investor i represented by $B_i = \{E_i(X), \Omega_i\}$ be the expected payoff and the covariance matrix as given:

$$E_i(X) = (E_i(x_1), \cdots, E_i(x_K))^T$$

$$\Omega_i = (\sigma_{i,kl})_{K \times K}, \qquad \sigma_{i,kl} = cov_i(x_k, x_l).$$

At $t = 0$ investor i invest her initial wealth $W_{i,0}$ in a portfolio which contains $Z_i = (z_{i,1}, z_{i,2}, \ldots, z_{i,K})^T$ shares of risky assets and the risk-free asset. The wealth of the portfolio for investor i at $t = 1$ is

$$W_i = R_f\left(W_{i,0} - P^T Z_i\right) + X^T Z_i,$$

and thus

$$E_i(W_i) = R_f\left(W_{i,0} - P^T Z_i\right) + E_i^T(X) Z_i, \qquad \sigma_i^2(W_i) = Z_i^T \Omega_i Z_i. \tag{2.1}$$

Investor i has a risk aversion coefficient of γ_i and a constant absolute risk aversion (CARA) utility function $U_i(W_i) = -e^{-\gamma_i W_i}$. As W_i is normally

distributed, the investor's optimal portfolio is obtained by maximizing the certainty-equivalent of future wealth:

$$V_i\left(W_i\right) = E_i\left(W_i\right) - \frac{\gamma_i}{2}\sigma_i^2\left(W_i\right)$$

Investors can borrow cash and stocks, but need to post an initial margin of m^L and m^S. Thus, the maximum investment in risky assets is $W_{i,0}/m^L$ and the maximum borrowing of risky asset is $W_{i,0}/m^S$.[1]

C. Portfolio Choice

Investors choose portfolios by maximizing their date-1 utility under leverage constraint and short-sales constraint. Thus the optimization problem is given by

$$\max V_i\left(W_i\right) = E_i\left(W_i\right) - \frac{\gamma_i}{2}\sigma_i^2\left(W_i\right)$$

subject to

$$m^L P^T Z_i \leq W_{i,0}$$

$$m^S P^T Z_i \geq -W_{i,0}$$

The Lagrangian is given by

$$\mathcal{L} \;=\; E_i\left(W_i\right) - \tfrac{\gamma_i}{2}\sigma_i^2\left(W_i\right) + \psi_i\left(W_{i,0} - m^L P^T Z_i\right) + \phi_i\left(m^S P^T Z_i + W_{i,0}\right)$$

where ψ_i and ϕ_i are non-negative shadow prices of the leverage constraint and the short-sales constraint, respectively.

Therefore the first order conditions are

$$\frac{\partial \mathcal{L}}{\partial Z_i} = \left[\frac{\partial E_i\left(W_i\right)}{\partial Z_i} - \frac{\gamma_i}{2}\frac{\partial \sigma_i^2\left(W_i\right)}{\partial Z_i}\right] - m^L\psi_i P + m^S\phi_i P = 0 \qquad (2.2)$$

[1] In the fully worked version, m^L and m^S will be derived from difference in beliefs, and will vary across investors and assets. We are not quite there yet.

$$\psi_i \geq 0, \qquad m^L P^T Z_i \leq W_{i,0}, \qquad \psi_i \left(W_{i,0} - m^L P^T Z_i\right) = 0$$

$$\phi_i \geq 0, \qquad m^S P^T Z_i \geq -W_{i,0}, \qquad \phi_i \left(m^S P^T Z_i + W_{i,0}\right) = 0$$

Given the expression (2.1) we have

$$\frac{\partial E_i\left(W_i\right)}{\partial Z_i} = E_i\left(X\right) - R_f P \qquad \frac{\partial \sigma_i^2\left(W_i\right)}{\partial Z_i} = 2\Omega_i Z_i. \tag{2.3}$$

Substituting (2.3) into (2.2) leads to the optimal demand for risky assets

$$Z_i = \tau_i \Omega_i^{-1}\left[E_i\left(X\right) - \left(R_f + m^L\psi_i - m^S\phi_i\right)P\right] \tag{2.4}$$

where $\tau_i = 1/\gamma_i$ is the risk tolerance coefficient.

D. Equilibrium

Investors choose the optimal portfolio strategies subject to the equilibrium prices based on their own assessments of the joint distribution for risky asset prices. The equilibrium prices are achieved when the market clears and the sum of individual demands equals aggregate market supply:

$$\sum_{i=1}^{I} Z_i = Z_m \tag{2.5}$$

where Z_m is the number of shares outstanding of risky assets. To establish the result, we represent heterogeneous beliefs by a consensus belief which, if held by all investors, would result in the same equilibrium prices as in the actual heterogeneous economy, as given by the definition in Verrecchia (1979).

Definition 1. *The consensus belief about the expected payoff or covariance matrix of the risky assets is an average of individual beliefs, weighted by their risk tolerances and the inverses of the covariance matrices:*

$$\Omega^{-1} = \sum_{i=1}^{I} \frac{\tau_i}{\tau} \Omega_i^{-1} \tag{2.6}$$

$$E(X) = \Omega \sum_{i=1}^{I} \frac{\tau_i}{\tau} \Omega_i^{-1} E_i(X) \tag{2.7}$$

where τ *is the aggregate risk tolerance* $\tau = \sum_{i=1}^{I} \tau_i$.
Similarly define the aggregate shadow price of short-sales constraint and leverage constraint as

$$\Phi = m^S \Omega \sum_{i=1}^{I} \frac{\tau_i}{\tau} \Omega_i^{-1} \phi_i, \qquad \Psi = m^L \Omega \sum_{i=1}^{I} \frac{\tau_i}{\tau} \Omega_i^{-1} \psi_i. \tag{2.8}$$

We denote by r_k *the return per share for asset* k *and by* r_m *the return on the market portfolio which is the value-weighted average of all risky assets. Then, under the consensus belief, the expected return of asset* k *is given by* $E(r_k) = E(x_k)/p_k - 1$, *and the market beta of asset* k *is* $\beta_k = cov(r_m, r_k)/\sigma^2(r_m)$, *where* $cov(r_m, r_k)$ *and* $\sigma^2(r_m)$ *are obtained from the covariance matrix* Ω *(2.6).*

Lemma 1. *Under the consensus belief as defined in Definition 1, equilibrium prices are given by*

$$P = (R_f I_K + \Psi - \Phi)^{-1} \left[E(X) - \tau^{-1} \Omega Z_m \right] \tag{2.9}$$

and the optimal portfolio of agent i *is*

$$Z_i = \tau_i \Omega_i^{-1} \left[E_i(X) - \chi_i \left(E(X) - \tau^{-1} \Omega Z_m \right) \right]. \tag{2.10}$$

where $\chi_i = (R_f + \psi_i - \phi_i)(R_f I_K + \Psi - \Phi)^{-1}$, *and* I_K *is an identity matrix.*

PROOF: See Appendix A.

Proposition 1. [CAPM with heterogeneous beliefs] *Under the consensus belief in Definition 1, the expected excess return on a risky asset relates to its expected market beta in the linear form as the standard CAPM, plus an extra term, which represents the net effect that borrowing constraints have on the Security Market Line:*

$$E(r_k) - r_f = \underbrace{\beta_k [E(r_m) - r_f]}_{standard\ CAPM} + \underbrace{f(\Psi - \Phi, \beta_k)}_{extra\ term}, \tag{2.11}$$

where

$$f(\Psi - \Phi, \beta_k) = \frac{[(\Psi - \Phi) P]_k}{p_k} - \beta_k \frac{P^T (\Psi - \Phi) Z_m}{P^T Z_m}.$$

The extra term $f(\Psi - \Phi, \beta_k)$ can be naturally interpreted as the mispricing relative to the standard CAPM, defined as the distance-to-CAPM (DtC). If DtC is positive (negative), the empirical SML is above (below) that predicted by the standard CAPM, and assets are underpriced (overpriced) relative to the standard CAPM.

PROOF: See Appendix B.

Corollary 1. *In order to obtain cleaner empirical predictions, we assume that investors agree upon the covariance matrix. Then Ψ and Φ are reduced to scalars, and thus $f(\Psi - \Phi, \beta_k) = (\Psi - \Phi)(1 - \beta_k)$. Consequently, equation (2.11) becomes*

$$E(r_k) - r_f = \beta_k [E(r_m) - r_f] + \chi (1 - \beta_k), \tag{2.12}$$

where $\chi = \Psi - \Phi$ is the difference in shadow price between the leverage constraint and the short-sales constraint. The distance-to-CAPM is given by

$$DtC = \chi (1 - \beta_k). \tag{2.13}$$

Remark 1. *(i) When $\chi = 0$, i.e. leverage constraint and short-sales constraint bind to a similar degree, $DtC = 0$ and thus the standard CAPM holds; (ii) when $\chi \neq 0$, i.e. either constraint dominates, DtC will flip the sign around $\beta = 1$. For example, when χ is positive, DtC will be positive for less risky assets ($\beta < 1$), and negative for riskier assets ($\beta > 1$). Therefore, χ, the net effect of borrowing constraint, is like a force that rotates the SML around the the fixed*

point which represents the market portfolio. The direction of rotation depends on which constraint dominates: clockwise for positive χ and counterclockwise for negative χ.

E. Disagreement and Borrowing Constraints

Remark 2. *The net effect of borrowing constraints on CAPM is captured by χ in equation (2.12), the value of which depends on disagreement. In the following discussion we will show that (i) when disagreement is very low or very high, χ is close to zero and thus the standard CAPM holds; (ii) when disagreement is not extreme, leverage constraint is tighter than short-sales constraint, and thus $\chi > 0$. As highlighted in Remark 1, when χ is positive, the SML will rotate clockwise and thereby becomes flatter. If the net effect of borrowing constraint χ is large enough, the SML will be downward-sloping.*

The tightness of borrowing constraints, Ψ and Φ, depends on the level of disagreement. Optimists want to make leveraged investments in risky assets by borrowing from pessimists using loans collateralized by the assets they have. However, pessimists do not value the collateral as much as optimists do and thus are reluctant to lend, which leads to an endogenous constraint on optimists' ability to borrow. Similarly, pessimists borrow assets from optimists to short sell using their cash holding as collateral. Since optimists expect an increase in asset prices, they would anticipate risks that pessimists' cash-collateral is deficient to pay off the debt. Consequently, optimists tend to reduce assets available to lend relative to demand, and thus pessimists face short-sales constraints. In summary, investors would face a limit on borrowing which resembles margin requirements.

Greater belief dispersion, i.e. the pessimists think that the asset is worth less and optimists think that it is worth more, increases borrowing demand so as to take positions in line with their expectations, which, in turn, tightens the borrowing constraints (ψ_i and ϕ_i increase). On the other hand, from lenders'

perspective, the collateral becomes less valuable relative to the loan contract, which leads lenders to reducing lending, and hence the tightness of borrowing constraints increases (m^L and m^S increase). Overall, a higher belief dispersion increases demand for borrowing (demand-channel) and reduces supply of lending (supply-channel), and thus the borrowing constraints get tighter.

Leverage constraint is generally tighter than short-sales constraint in terms of supply-channel effect. The intuition is as follows. Pessimistic lenders discount the borrowers' collateral by a discount rate, which can be proxied by the ratio of the current asset price to pessimists' valuation, and the inverse ratio applies to optimistic lenders. If pessimists and optimists expect the same absolute deviation from the current price, the discount rate used by pessimistic lenders would be larger than that used by optimistic lenders, meaning that pessimistic lenders would discount the collateral more than optimistic lenders do, and thus leverage constraint would be tighter than short-sales constraint (See Appendix C for a simple example illustrating this intuition).

When belief dispersion is very high, investors can hardly borrow cash or assets due to lenders' reluctance to lend, which leads to the maximum supply-channel effect. As dispersion keeps increasing, the borrowing constraints will be dominated by the demand-channel effect, and consequently both leverage constraint and short-sales constraint will bind to a similarly large degree.

F. Disagreement and Trading Volume

Suppose for simplicity that each investor has initial endowment of risky assets $Z_{i,0} = \frac{\tau_i}{\tau} Z_m$ at time 0, which is a market portfolio weighted by the risk tolerance coefficient so that the sum of individual endowments is equal to the aggregate supply: $\sum_{i=1}^{I} Z_{i,0} = Z_m$. It can be shown that $Z_{i,0}$ is the optimal portfolio when there is no disagreement about the expected payoff by imposing $E_i(X) = E(X)$, $\Omega_i = \Omega$ and $\chi_i = 1$ in equation (2.10).

In the next period, investors have the identical assessment of the variance-covariance matrix but their beliefs about expected payoff start to diverge ($E_i(X) \neq E(X)$). Define investor i's volume of trade in risky assets T_i as

the actual demand (2.10) minus the endowment $Z_{i,0}$:

$$T_i = Z_i - Z_{i,0} = \tau_i \Omega^{-1} \left[E_i(X) - \chi_i E(X) + (\chi_i - 1) \tau^{-1} \Omega Z_m \right], \qquad (2.14)$$

where $\chi_i = (R_f + \psi_i - \phi_i)(R_f + \Psi - \Phi)^{-1}$. Then the overall trading volume is given by the following corollary.

Corollary 2. *The equilibrium trading volume measure V is given by*

$$V = \frac{1}{2} \tau \Omega^{-1} \left[\triangle - (\Phi + \Psi) P \right],$$

where $\triangle = \sum_{i=1}^{I} \frac{\tau_i}{\tau} \mid E_i(X) - E(X) \mid$ measures the dispersion in beliefs.

PROOF: See Appendix D.

In the presence of high belief dispersion $\triangle$, trading volume tends to be high, as in Atmaz & Basak (2018); however, $(\Phi + \Psi)$, which increases with the level of disagreement, have negative effect on trading volume. Thus, disagreement generates two counteracting effects on trading volume: a positive direct effect characterized by the measure of belief dispersion; and a negative indirect effect through the shadow prices of borrowing constraints.

Although it is difficult to obtain an analytical derivative of trading volume with respect to disagreement, given the empirical evidence (Atiase (1991), Bessembinder et al. (1996), Goetzmann & Massa (2005)), we could plausibly conjecture that the positive direct effect is stronger in general. For example, consider a group of unconstrained optimists mildly increase their expectation, which in turn should generate trading volume. Suppose they finance the increase in the long positions by borrowing more from pessimists who have excess cash holdings. Since the pessimists keep their expectation unchanged, they would regard the optimists' collateral as before. Therefore, the pessimists are very likely to provide the additional liquidity to optimists, who would remain unconstrained thereafter. Consequently, the increase in the optimists' expectations do not change the aggregate shadow costs of borrowing constraints, and thus the indirect channel of disagreement is shut down.

Remark 3. *The effect of disagreement on trading volume can be either positive or negative: trading volume increases in disagreement since investors with dispersed beliefs have high speculative demand; but disagreement also has a negative effect which limits the growth in trading volume, because the increase in disagreement tightens borrowing constraints and thus prohibits the amount of volume which would be traded if the constraints did not bind. In general, the positive effect is dominant, which is supported by several empirical studies. Overall, trading volume increases in disagreement.*

G. Disagreement and Equity Premium

Rearranging (2.9) and normalizing the initial wealth of the market to one yields the equity premium

$$E\left(r_m\right) - r_f = \frac{1}{\tau}\sigma^2\left(r_m\right) + \left(\Psi - \Phi\right) \tag{2.15}$$

With the same expected market variance and risk-free rate as the economy with homogeneous beliefs, disagreement about expected payoffs leads to borrowing constraints binding for some investors, and the leverage constraint is generally tighter than the short-sales constraint (as discussed in Remark 2). Thus, $\Psi - \Phi$ is positive and increases the premium. As belief dispersion keeps increasing, two constraints tend to bind to the similar extent and thus the net effect approaches zero.

Moreover, high disagreement might be associated with higher expected market variance ($\sigma^2\left(r_m\right)$), which is the implication of Banerjee & Kremer (2010) following the assumption that investors' uncertainty about the interpretation of future payoff increases with the level of disagreement. The positive relation between belief dispersion and expected market variance is also confirmed empirically by Barinov (2013). A plausible explanation for this positive relation might be that, when the belief is highly dispersed in the stock market, investors would observe some other investors having very different beliefs, which makes them less confident about their own interpretation, that is, the expected market

variance increases. Equation (2.15) shows that this increase of investor confidence risk would be priced in the equity premium, as in Bansal & Shaliastovich (2010).[2]

Remark 4. *In sum, heterogeneous beliefs command a positive equity premium, either through increasing the shadow prices of borrowing constraints, or through decreasing investors' confidence (increasing the expected market variance).*

H. Model Predictions

Our model generates the following testable predictions:

(i) A high level of disagreement is associated with high trading volume and equity premium (Remark 3 and Remark 4);

(ii) On very low- (or high-) disagreement days, the net effect of two borrowing constraints is close to zero, and thus the standard CAPM will apply (Remark 2: (i)).

(iii) When disagreement is moderate, leverage constraint tends to be tighter than short-sales constraint. As a consequence, *distance-to-CAPM* is positive for small beta ($\beta < 1$) and negative for large beta ($\beta > 1$). The empirical SML is above that predicted by the standard CAPM for low-beta assets and below it for high-beta assets (Remark 2: (ii)).

We next provide evidence for our model predictions.

2.3 Empirical Analysis

Data

Our measure of disagreement is the daily trading volume of all listed stocks on New York Stock Exchange, American Stock Exchange, the Nasdaq Stock Market and the Arca Stock Market. The volume is defined as in Campbell

[2]This is not allowed in the model, which has dogmatic differences in beliefs. A more general model can allow for these effects with a considerable increase in complexity.

et al. (1993), which is the log ratio of number of shares traded to the number of outstanding, detrended by a one-year moving average filter. We obtain trading-volume data from the Center for Research in Security Prices (CRSP). Figure 2.1 plots the time series of our disagreement measure.

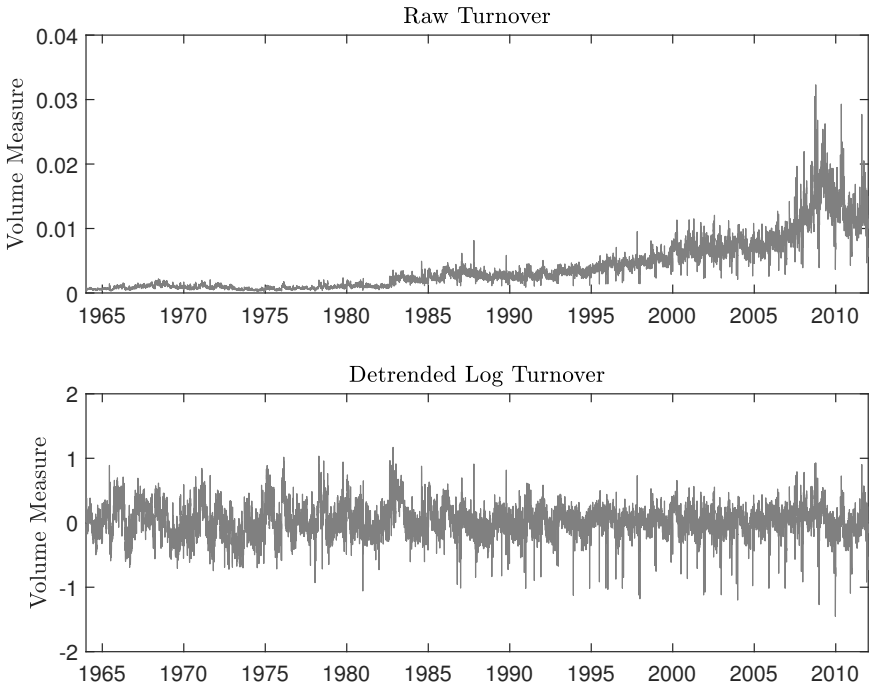


Figure 2.1: Stock Market Turnover

We study the relation between CAPM betas and return using ten beta-sorted portfolios provided by Savor & Wilson (2014). They first estimate stock market betas for all stocks on CRSP using rolling windows of 12 months of daily returns from 1964 to 2011, and then sort stocks into one of ten beta-decile value-weighted portfolios. Following Savor & Wilson (2014), our stock market proxy is the CRSP NYSE, Amex, and Nasdaq value-weighted index of all listed shares. Table 2.1 presents descriptive statistics for those variables.

Table 2.1: Summary Statistics for Time-Series Variables

This table presents descriptive statistics for key variables used in the paper. *Turnover* is the measure of disagreement. *MKTRF* is the excess return of the CRSP value-weighted index. *R(1)-R(10)* are the returns of ten beta-sorted portfolios. Returns are expressed in percent. The sample period is 1/2/1964 to 12/30/2011 with 12085 daily observations.

	Mean	Std.Dev.	P10	P25	Median	P75	P90
Turnover	0.031	0.238	-0.248	-0.107	0.029	0.164	0.428
MKTRF	0.021	1.008	-1.050	-0.430	0.050	0.490	1.473
R(1)	0.031	0.798	-0.692	-0.251	0.050	0.337	1.107
R(2)	0.044	0.592	-0.533	-0.201	0.060	0.309	0.866
R(3)	0.043	0.605	-0.562	-0.206	0.066	0.315	0.872
R(4)	0.039	0.682	-0.651	-0.243	0.066	0.349	0.993
R(5)	0.047	0.776	-0.736	-0.287	0.076	0.404	1.090
R(6)	0.042	0.894	-0.868	-0.339	0.069	0.454	1.264
R(7)	0.043	1.029	-1.008	-0.412	0.066	0.515	1.496
R(8)	0.039	1.165	-1.162	-0.480	0.064	0.573	1.704
R(9)	0.038	1.417	-1.414	-0.605	0.056	0.690	2.055
R(10)	0.036	1.900	-1.957	-0.833	0.074	0.901	2.734

Results

We evaluate our model implications as follows. First, we sort *Turnover* into three groups: very low-disagreement days (below the 10th percentile of *Turnover*), moderate-disagreement days (range from 10th percentile to the 90th percentile) and very high-disagreement days (above the 90th percentile). We then separately plot the SML for periods of low, moderate, and high disagreement. The results are summarized in Figure 2.2-2.4.

Prediction 1. *A high level of disagreement is associated with high equity premium.*

The average excess market return $\bar{r}_M^e$ is -9.8, 1.3, and 19.9 basis points (bps), respectively, for low, moderate, and high disagreement. This shows that equity premium increases with the level of disagreement, which is consistent with Prediction 1.

Prediction 2. *On very low or very high disagreement days, the standard CAPM holds.*

During the periods of very low or high disagreement (Figure 2.2 and Figure 2.4), the empirical SML (solid points and a solid line) runs very near its CAPM prediction (a dotted line).[3] Note that the SML on very low-disagreement days is downward-sloping; however, it does not indicate the rejection of the CAPM because the average excess return one low-disagreement days is negative. The R^2s of SMLs are 98% for low-disagreement days and 99% for high-disagreement days, respectively, indicating that most of the variation in average excess returns is accounted for by their stock market betas. This evidence is consistent with Prediction 2.

Prediction 3. *When disagreement is not extreme, the empirical SML is above that predicted by the standard CAPM for low-beta assets ($\beta < 1$) and below it for high-beta assets ($\beta > 1$).*

When disagreement is moderate (Figure 2.3), the empirical SML (solid points and a solid line) differs significantly from that predicted by the standard CAPM (a dotted line), and its R^2 is 54%.[4] Furthermore, average excess returns are plotted above the CAPM prediction for $\beta < 1$, and below it for $\beta > 1$, supporting Prediction 3.

2.4 Conclusion

This paper shows that the interaction of heterogeneous expectations, leverage constraint and short-sales constraint determines the Security Market Line,

[3]On low-disagreement days, the intercept is 0.1 bps and is not significant different from zero (t-statistics=0.2). The slope of the SML is -10.5 bps (t-statistics=-18.9), and it is not significantly different from the average market excess return of -9.8 bps (the t-statistics for the difference is 1.3). On high-disagreement days, the intercept is -2.3 bps and is just marginally significantly different from zero (t-statistics=-2.5). The slope of the SML is 22.6 bps (t-statistics=23.2), and it is just marginally significantly different from the average market excess return of 19.9 bps (the t-statistics for the difference is 2.8).

[4]On moderate-disagreement days, the intercept is 2.4 bps and is significantly different

both theoretically and empirically. Disagreement drives borrowing demand for both cash and risky assets: optimists who think assets are underpriced may want to borrow money from pessimists to take more positions; in contrast, pessimists may want to borrow more assets from optimists to sell short. However, disagreement reduces the supply of loans because lenders do not value the collateral which consists of borrowers' wealth as much as their borrowing counterparts do, and therefore are reluctant to meet demand fully. This leads to endogenous leverage constraint and short-sales constraint. In equilibrium, a modified version of CAPM holds which incorporates the shadow prices of these borrowing constraints.

from zero (t-statistics=6.6). The slope of the SML is -1.2 bps (t-statistics=-3.1), and it is significantly different from the average market excess return of 1.3 bps (the t-statistics for the difference is 6.5).

2.5 Appendix

Appendix A: Proof of Lemma 1

Substituting investors' asset demands (2.4) into the market clearing condition (2.5) yields

$$Z_m = \sum_{i=1}^{I} Z_i = \sum_{i=1}^{I} \tau_i \Omega_i^{-1} \left[E_i(X) - \left(R_f + m^L \psi_i - m^S \phi_i \right) P \right] \tag{2.16}$$

Using Definition 1, we rewrite (2.16) as

$$Z_m = \tau \Omega^{-1} \left[E(X) - \left(R_f I_K + \Psi - \Phi \right) P \right] \tag{2.17}$$

Rearranging (2.17) leads to

$$P = \left(R_f I_K + \Psi - \Phi \right)^{-1} \left[E(X) - \tau^{-1} \Omega Z_m \right].$$

Substituting P into the demand function (2.4) gives the optimal demand of investor i:

$$Z_i = \tau_i \Omega_i^{-1} \left[E_i(X) - \chi_i \left(E(X) - \tau^{-1} \Omega Z_m \right) \right].$$

where $\chi_i = \left(R_f + \psi_i - \phi_i \right) \left(R_f I_K + \Psi - \Phi \right)^{-1}$.

■

Appendix B: Proof of Proposition 1

Under consensus belief the the variance of market portfolio is expressed as $\sigma^2(W_m) = Z_m^T \Omega Z_m$. Substituting (2.17) into $\sigma^2(W_m)$ gives

$$\sigma^2(W_m) = \tau \left[E(W_m) - R_f W_{m,0} - \Psi_m + \Phi_m \right] \tag{2.18}$$

where $W_{m,0} = P^T Z_m$, $\Psi_m = P^T \Psi Z_m$ and $\Phi_m = P^T \Phi Z_m$. Rearranging 2.18 gives

$$\frac{1}{\tau} = \frac{E(W_m) - R_f W_{m,0} - \Psi_m + \Phi_m}{\sigma^2(W_m)}. \tag{2.19}$$

Substituting 2.19 into 2.9 yields

$$E(X) - R_f P = \frac{1}{\sigma^2(W_m)} \Omega Z_m \left[E(W_m) - R_f W_{m,0} - \Psi_m + \Phi_m\right] + (\Psi - \Phi) P,$$

or equivalently

$$E(x_k) - R_f p_k = \frac{\sigma(W_m, x_k)}{\sigma^2(W_m)} \left[E(W_m) - R_f W_{m,0} - \Psi_m + \Phi_m\right] + \left((\Psi - \Phi) P\right)_k. \tag{2.20}$$

Dividing (**??**) throughout by p_k gives

$$E(r_k) - r_f = \beta_k \left[E(r_m) - r_f\right] + \frac{\left[(\Psi - \Phi) P\right]_k}{p_k} - \beta_k \frac{P^T (\Psi - \Phi) Z_m}{P^T Z_m}. \tag{2.21}$$

■

Appendix C: An example for Remark 2

To exemplify the intuition behind Remark 2, consider a simple example with one asset, two periods (0,1), and two investors who disagree on the future price of the asset. Suppose the price of the asset per unit at date 0 is P. The optimist (Anna) believes the price will increase by Δ_A at date 1 and the pessimist (Bob) expects a decrease Δ_B. Suppose both of them have identical wealth W, which consists only of cash for Anna and the risky asset for Bob. Given their expectations, Anna will borrow cash from Bob to buy more shares of the asset; Bob will borrow the asset from Anna to short sell. Assume their optimal amounts of borrowing in terms of cash are identically equal to W; however, their demands for loans may not be satisfied due to the borrowing constraints.

Suppose Anna borrows L dollars of cash using W/P units of asset as col-

lateral at date 0, and promises to repay L dollars at date 1, assuming she can borrow at zero cost. Since Bob believes the asset price at date 1 is $P-\Delta_B$ and hence expects the value of Anna's collateral at date 1 is $(P-\Delta_B)W/P$. Bob is willing to lend only if the expected value of collateral is greater or equal to repayment. Thus, the maximum loan amount is $L = W(P-\Delta_B)/P$.

Similarly, Bob borrows S/P units of the asset using W dollars of cash as collateral at date 0 and promises to repay S/P units of the asset at date 1. Since Anna believes the asset price at date 1 is $P+\Delta_A$, and therefore expects the value of repayment at date 1 is $(P+\Delta_A)S/P$. Anna is willing to lend only if the expected value of collateral is greater or equal to repayment. Thus the maximum loan amount is $S = WP/(P+\Delta_A)$.

Define the tightness of borrowing constraints as the fraction of demands for loans which are not satisfied. Let C^L and C^S denote the tightness of leverage constraint and short-sales constraint, respectively

$$C^L = 1 - \frac{L}{W} = \frac{\Delta_B}{P} = \delta_B,$$

and

$$C^S = 1 - \frac{S}{W} = \frac{\Delta_A/P}{1+\Delta_A/P} = \frac{\delta_A}{1+\delta_A},$$

where $\delta_{A(B)} = \Delta_{A(B)}/P$ is the rate of return on the asset expected by Anna (Bob). δ_B ranges from 0 to 1 because the largest price decline expected by Bob is P, while δ_A ranges from 0 to positive infinity because there is no upper bound for the optimistic expectation. As illustrated in Figure 2.5, C^L and C^S are both monotonic functions of δ. When there is no disagreement, lenders evaluate borrowers as wealthy as themselves and thus are willing to lend. Consequently, borrowers can borrow the full amount with the borrowing constraint not binding. An increase in disagreement leads to an increase in the discount of the collateral value by lenders, and hence lenders are more reluctant to lend and tighten the borrowing constraints.

With the same level of δ, the leverage constraint is always tighter than the short-sales constraint. When pessimists expect the asset to be worthless ($\delta_B = 1$) and thus the collateral value to be zero, they will not lend cash at all and $C^L = 1$. As optimism increases to a very high level, optimists will not lend assets out because they think pessimists cannot afford to buy back the asset which was sold short, which leads to C^S approaching one. In other words, when belief dispersion is very high, both the leverage constraint and the short-sales constraint will bind to a similarly large degree.

Appendix D: Proof of Corollary 2

We rewrite investor i's expectation as $E_i(X) = E(X) + \triangle_i$, where $\triangle_i$ is the investor i's bias . Substitute this expression of $E_i(X)$ into (2.14) yields

$$T_i = \tau_i \Omega^{-1} \left[(1 - \chi_i) \left(E(X) - \tau^{-1} \Omega Z_m \right) + \triangle_i \right]. \tag{2.22}$$

Plugging (2.9) into (2.22) yields

$$T_i = \tau_i \Omega^{-1} \left[(\Psi - \Phi - \psi_i + \phi_i) P + \triangle_i \right]. \tag{2.23}$$

Now consider an investor who has the correct belief and thus $\triangle_i = 0$. For simplicity we assume the investor is not constrained, $\psi_i = \phi_i = 0$, and thus $T_i = \tau_i (\Psi - \Phi) \Omega^{-1} P$. Therefore, the investor holds a portfolio that deviates from the market portfolio due to the aggregate effect of borrowing constraints for other investors who hold biased beliefs. Consequently, the market portfolio is no more the tangency portfolio. We further assume that optimistic (pessimistic) investors have positive (negative) biases for all risky assets, and the number of shares traded in each asset is nonnegative (nonpositive), i.e. $\triangle_{i,k} > 0$ ($\triangle_{i,k} < 0$) and $T_{i,k} \geq 0$ ($T_{i,k} \leq 0$) for $k = 1, \cdots, K$.

First consider unconstrained investors for whom the shadow prices of borrowing constraints are zero: $\psi_i = \phi_i = 0$. Thus, his trading volume becomes $T_i = \tau_i \Omega^{-1} [(\Psi - \Phi) P + \triangle_i]$. Suppose each individual has negligible influ-

ence on the overall tightness of constraints. All else equal, an increase in the absolute value of investor's bias leads to a higher trading volume. Next, consider the leverage (short-sales) constrained investors. Note that constrained investors' dollar volume is determined by the margin requirement and their initial wealth. Let D_i^L (D_i^S) be the dollar trading volume for a leverage (short-sales) constrained investor:

$$D_i^L = W_{i,0}/m^L - (W_{i,0} - z_f) = P^T T_i$$

$$D_i^S = W_{i,0}/m^S + (W_{i,0} - z_f) = -P^T T_i,$$

which together with (2.23) yields ψ_i and ϕ_i

$$\psi_i = (\Psi - \Phi) + \frac{P^T \Omega^{-1} \triangle_i - D_i^L/\tau_i}{P^T \Omega^{-1} P}$$

$$\phi_i = (\Phi - \Psi) - \frac{P^T \Omega^{-1} \triangle_i + D_i^S/\tau_i}{P^T \Omega^{-1} P}.$$

Then the overall trading volume is measured as the sum of the absolute trading volume of all investors:

$$V = \frac{1}{2} \sum_{i=1}^{I} \mid T_i \mid,$$

where the adjustment $1/2$ prevents double summation of the shares across investors. Investors are naturally divided into four groups: (i) unconstrained optimists $i \in U+$; (ii) unconstrained pessimists $i \in U-$; (iii) constrained optimists $i \in C+$; and (iv) constrained pessimists $i \in C-$. Suppose four groups are of equal importance in determining the consensus belief, i.e. $\sum_{i \in U+} \tau_i = \sum_{i \in U-} \tau_i = \sum_{i \in C+} \tau_i = \sum_{i \in C-} \tau_i$. Then the trading volume can be shown to be

$$
\begin{aligned}
V &= \tfrac{1}{2}\left(\sum_{i\in U+} \mid T_i \mid + \sum_{i\in U-} \mid T_i \mid + \sum_{i\in C+} \mid T_i \mid + \sum_{i\in C-} \mid T_i \mid\right) \\
&= \tfrac{1}{2}(\sum_{i\in U+} \tau_i \Omega^{-1}\left[(\Psi-\Phi)\,P + \triangle_i\right] + \sum_{i\in U-} \tau_i \Omega^{-1}\left[-(\Psi-\Phi)\,P - \triangle_i\right] \cdots \\
&\quad + \sum_{i\in C+} \tau_i \Omega^{-1}\left[(\Psi-\Phi)\,P + \triangle_i - \psi_i P\right] + \sum_{i\in C-} \tau_i \Omega^{-1}\left[-(\Psi-\Phi)\,P - \triangle_i - \phi_i P\right]) \\
&= \tfrac{1}{2}\Omega^{-1}\left(\textstyle\sum_{i=1}^{I} \tau_i \mid \triangle_i \mid - \sum_{i\in C+} \tau_i \psi_i P - \sum_{i\in C-} \tau_i \phi_i P\right) \\
&= \tfrac{1}{2}\tau \Omega^{-1}\left[\Delta - (\Psi+\Phi)\,P\right]
\end{aligned}
$$

where $\triangle = \sum_{i=1}^{I} \frac{\tau_i}{\tau} \mid \triangle_i \mid$ measures the dispersion in beliefs. ■

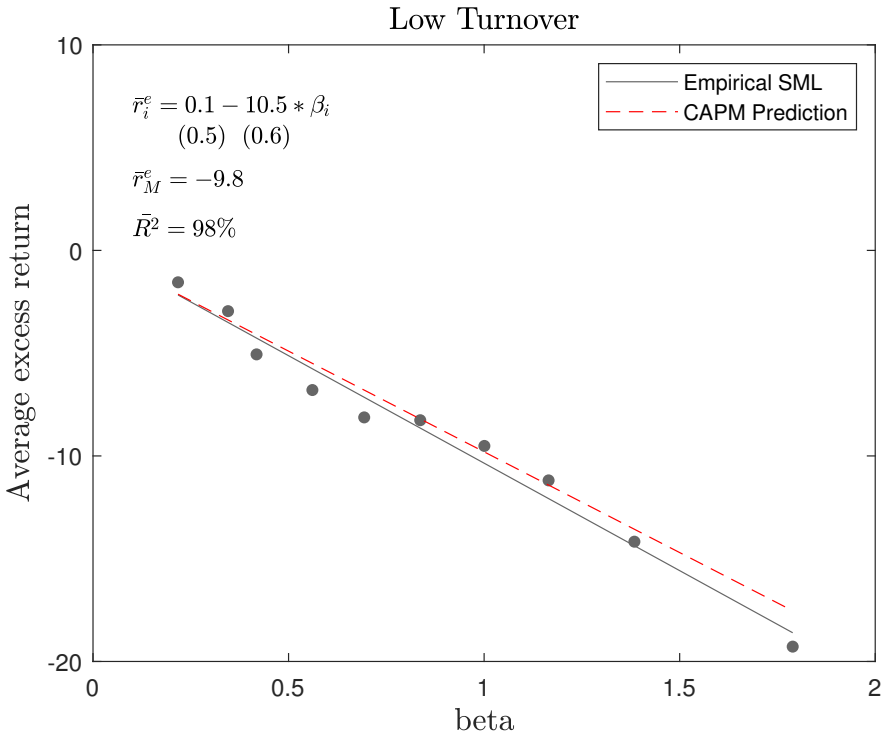


Figure 2.2: SML on Low-turnover Days. This figure plots average daily excess returns in bps against market betas for ten beta-sorted portfolios of all NYSE, Amex, and Nasdaq stocks (the solid points), and the implied ordinary least squares estimates of the SML (the solid line). The sample covers the 1964–2011 period. In addition, the SML predicted by the standard CAPM is displayed (the dotted line).

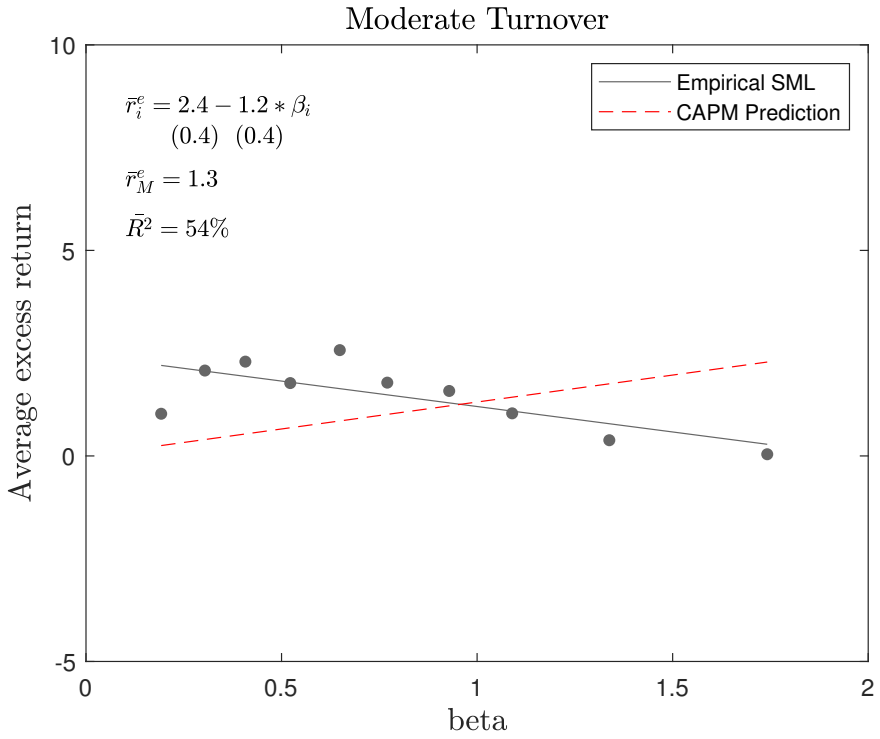


Figure 2.3: SML on Moderate-turnover Days. This figure plots average daily excess returns in bps against market betas for ten beta-sorted portfolios of all NYSE, Amex, and Nasdaq stocks (the solid points), and the implied ordinary least squares estimates of the SML (the solid line). The sample covers the 1964–2011 period. In addition, the SML predicted by the standard CAPM is displayed (the dotted line).

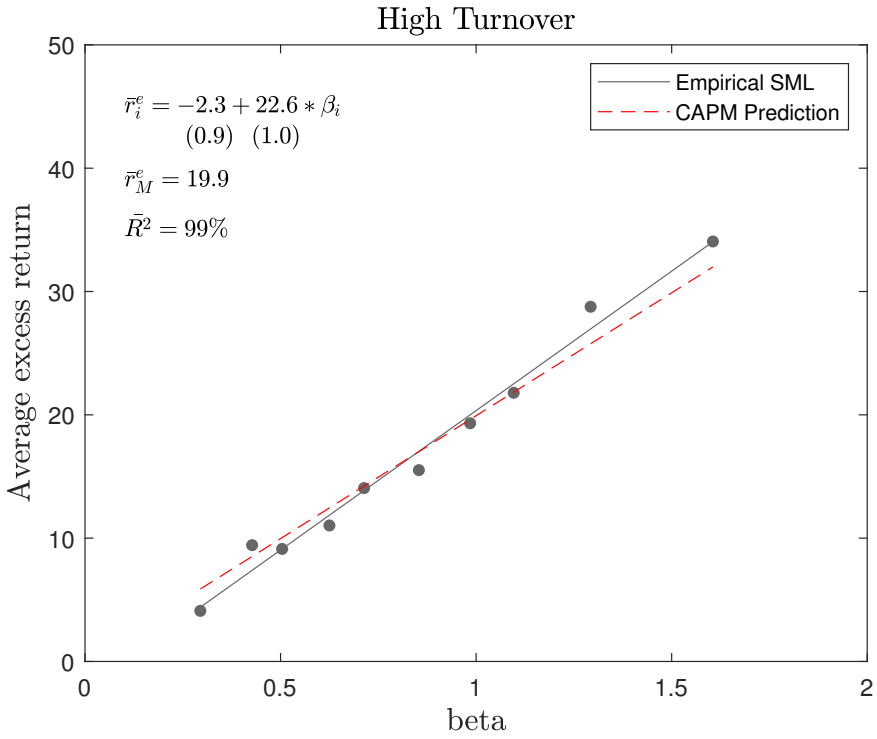


Figure 2.4: SML on High-turnover Days. This figure plots average daily excess returns in bps against market betas for ten beta-sorted portfolios of all NYSE, Amex, and Nasdaq stocks (the solid points), and the implied ordinary least squares estimates of the SML (the solid line). The sample covers the 1964–2011 period. In addition, the SML predicted by the standard CAPM is displayed (the dotted line).

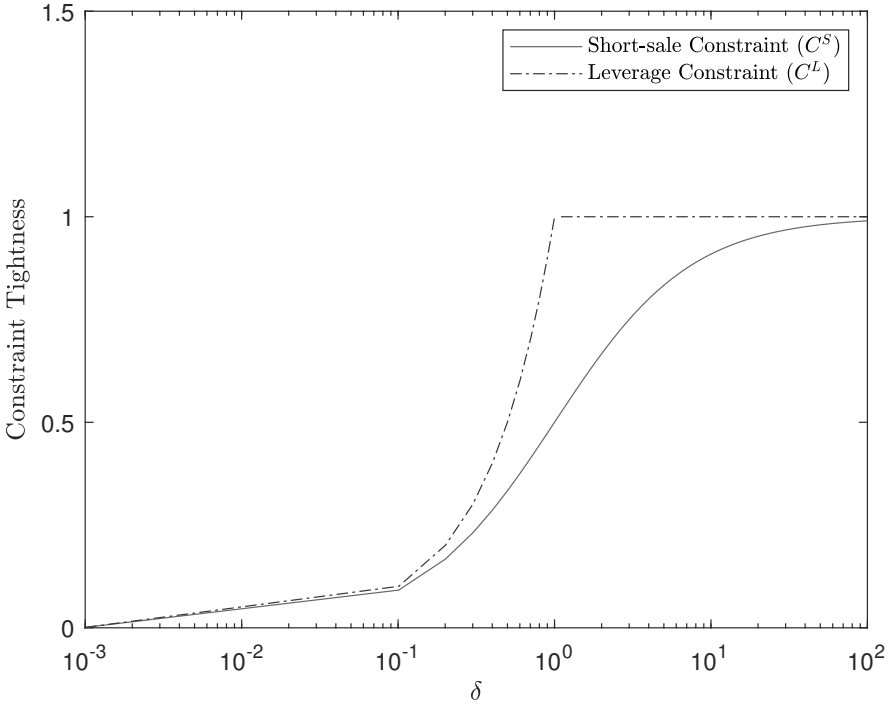


Figure 2.5: Borrowing Constraints

Chapter 3

Indivisible Goods and Relative Wealth Concerns: Gambling, Under-Participation and Lotteries.

3.1 Introduction

Why do economic agents care about inequality? In a standard endowment economy, agents care about the levels of their own endowments and about prices, but not otherwise about relative endowments. Since prices depend on aggregate relative endowments, it immediately follows that they do not care about inequality in itself. Yet plenty of evidence suggests that agents care very much about relative endowments holding their own levels and prices constant, as shown by much recent research, notably the work of Piketty (2014), Saez (2021). For example, Saez argues that the tremendous increase in the provision of public services by governments of developed economies in the past century is very strong evidence that voters dislike inequality. The failure of standard economic theory to account for inequality concerns is perhaps one of its most perplexing aspects to non-economists. Furthermore, if agents have strong concerns about relative wealth, some predictions of standard models will not be correct, such as the prediction that as living standards rise, people will consume more leisure (see e.g. Keynes (1930)). If, as we shall argue below,

inequality concerns may themselves induce inequality to increase, then understanding why agents care about inequality is crucial to reducing it, and also to understanding which policy interventions are likely to be effective.

Motivated by this failure, many papers assume that agents care about relative wealth and proceed from there. The problems with this approach are: First (obviously) that assuming that agents care about inequality is not an explanation of why they do so and; Second, we can't be sure that these assumptions correctly capture the ways in which agents care, and therefore what are the consequences for behaviour of such inequality concerns. For example, Roussanov (2010) assumes agents have a utility function which is concave in an agent's own wealth and linear in the ratio of her own wealth to average wealth. Such agents will underdiversify their investments, relative to a standard model, but not gamble. By contrast, Friedman and Savage (1948) famously assumed that agents had preferences which were concave in their own wealth but convex in *status*, giving rise to a desire to gamble. Other papers make assumptions about keeping up with the Joneses, or DeMarzo et al. (2008), which result in herd-like investment behaviour, and therefore bubbles, but not gambling in the sense of Friedman and Savage. Such assumptions are apparently all fairly radical departures from standard economic models.

In this paper, we seek to characterize a minimal departure from the standard endowment economy required to generate inequality (relative wealth) concerns, and then to explore properties of equilibria of such a minimally different economy: to see which predictions of standard models no longer hold and which novel predictions arise. The crucial, usually implicit, assumption of standard models which we claim is unrealistic is the assumption that all goods are divisible. Allowing for the existence of at least one complementary indivisible good in a closed endowment economy with multiple goods is a sufficient condition for some agents with standard preferences to have strong relative wealth concerns. These concerns are particularly strong if the indivisible goods are scarce.

Indivisibility appears to be a key feature of many important goods in the

real world. For example most reasonably well-off households appear not to wish to share their primary accommodation, so housing is often treated as indivisible, even if it varies in quality (Piazzesi et al. (2007)). Matching relationships such as spousal partnerships, job matches and students to schools also appear to be difficult or impossible to share without removing their utility (Cole et al. (1992)) and medical treatment programs are usually of limited usefulness unless substantially completed. Furthermore these important goods are complementary to more everyday consumption goods such as food or energy. From our perspective, the key aspect of such goods is that the agent with the highest valuation secures the additional available unit of the good, and that agent is usually the richer agent. Even when wealth is not necessarily what determines who wins the additional unit, as in the case of spousal or job matching for example, there is a strong tournament-like element. Finally, in our setup, players' relative wealth concerns are cardinal, not ordinal: The greater their relative wealth, the greater their welfare, in contrast to agents who only care about being ranked relatively wealthiest.

Indivisible complements exert much stronger effects in our model when they are scarce: As the quantity available of such goods increases, inequality concerns abate much faster than prices decline. A priori, scarcity of such goods should often be the case, since it makes no sense for example to provide far more college degree courses than there are students who want to enrol on them, so that on average there will be more students applying than available college places. To us this seems realistic: our impression is that people care more about inequality of access to elite educational institutions or cutting-edge medical treatments (both scarce, and indivisible) than inequality of access to fast cars or fancy club memberships (arguably indivisible but not so scarce). Relative to some previous related work, a key part of our approach is to relax the assumption of one-to-one matching, under which every agent can obtain one unit in equilibrium, albeit of varying quality. In our model, some agents can obtain no units while others obtain more than one.

Once inequality concerns are generated by indivisible goods, they give rise to other interesting phenomena such as a desire to gamble. In a two-agent economy, for example, agents have indirect utility of wealth of the Friedman-Savage form. These preferences are locally concave, except at equality of wealth, but convex over a wider range. As noted by Friedman & Savage (1948), Becker et al. (2005) and Chetty & Szeidl (2007), gambling increases expected utility for both agents provided the poorer agent can become the richer agent if she wins. In contrast to other papers, we study gambles in which agents have equal bargaining power, or more generally in which each agent is paid her Shapley value in expectation. As a result, these gambles are not in general actuarially fair, with better terms for the initially richer agent. Thus in economies in which inequality concerns arise from the existence of scarce indivisible complements, inequality is expected to increase, and to accelerate. Thus our explanation of inequality concerns has fairly strong predictions about how inequality should evolve over time.

In the economy which we study, richer agents pay poorer agents a price which depends positively on the poorer agent's wealth. Thus richer agents have relative wealth concerns: Holding their own wealth constant, they prefer the poorer agent's wealth to be lower. The poorer agent's welfare is a leveraged function of her own wealth, while the richer agent's welfare is a develeraged function of the poorer agent's wealth. In the presence of aggregate risk, the richer agent's wealth is lower in bad times, making her worse off, but so is the poorer agent's wealth, making the richer agent better off. For poorer agents there is no such offsetting effect. In consequence, richer agents are naturally hedged against common risks, which reduce the wealth of all agents, while the opposite is true for poorer agents. Richer agents are therefore more willing to bear aggregate risk, which in equilibrium has a positive expected return, under precisely the same conditions and for the same reasons that 'background' labour income risk can increase agent's risk aversion (Eeckhoudt & Kimball (1992)). This greater participation in positive expected return risk further

increases expected growth in inequality. These results suggest that an important application of this model may be in finance, since the same mechanism that generates inequality concerns also generates lower participation by the relatively poor.

Finally, our model can explain mass participation in lotteries by rational agents with standard preferences. In an extension with one rich agent and many identical poor agents, too poor for the rich agent to be willing to gamble with any of them, the poor agents can engage in a variety of lotteries which are designed to produce one rich winner. This rich winner can then force a much higher price for the indivisible goods, paid to the mass of poor agents, or simply enter into a gamble with the rich agent (or both). Either the desire to force a higher price, or to be able to gamble with the rich, are sufficient motivations to play the lottery for the poor agents, but the first is much stronger. One potential application of this section is that our model may be able to explain populism, defined as negative present value policies which implement redistributions towards 'the people' (Edwards (2019)).

What measures could mitigate these inequality concerns and thus, paradoxically, reduce the expected increase in inequality? One answer is to produce more units of the indivisible good, but this is likely to be inefficient. Another may be to produce different types of scarce indivisible goods, or goods of the same type but different qualities. Consistent with this idea, we find that as the number of available goods increases, inequality concerns abate at low levels of inequality, but not at higher levels. One intervention that is unlikely to be effective on its own is tax-funded redistribution, since much of the inequality is 'manufactured' in the sense of Rosen (1997) through gambling.

Ng (1965) showed how complementary indivisible goods give rise to indirect utility of the Friedman-Savage type once wealth becomes large enough to buy an additional unit. Our contribution is to show how, in equilibrium, competition for these complementary goods gives rise to relative wealth concerns, and how these relative wealth concerns alter behavior relative to a standard model with

divisible goods.

The next section explains how agents compete for goods in our setting. Section 3 describes gambling in a setting without aggregate risk. Section 4 details how aggregate risk changes investment choices. Section 5 considers an extension to lotteries with multiple identical poor agents. Section 6 concludes.

3.2 Competition for indivisible goods

In this section we show how the presence of indivisible goods can give rise to relative wealth concerns, in the sense that agents' indirect utility of consumption depends on the distribution of wealth across agents. First, however, we show briefly how no such concerns arise in an economy in which all goods are divisible.

Inequality does not matter in an endowment economy with divisible goods

Trivially, in an endowment economy with a single good, where more is better, no agent's welfare can be altered by changing any other agent's endowment, holding the first agent's endowment constant. All agents simply consume their own endowments, indivisible or not. In a two-good economy, agents care about the levels of each good in their own endowment, and about the relative price of the two goods. This relative price depends on the ratio of the aggregate endowment of each. An agent whose own relative endowment is the same as the aggregate relative endowment cannot be made better or worse off by changing the distribution of the aggregate endowment across agents, holding her own endowments constant. We can also give a simple example in which making an agent relatively poorer (holding her own endowment constant) can increase her welfare, while making her relatively richer can reduce it.

We can give a simple example of how an increase in inequality can make the relatively poorer of two agents better off in terms of increasing her welfare.

Suppose $j = 1, 2$ two agents (agent 1 always being the richest and so on down) have CES utility of consumption of both goods of the form

$$U(C_j, H_j) = \left(C_j^{\rho} + H_j^{\rho}\right)^{\frac{1}{\rho}}$$

where ρ is the elasticity of substitution between the two goods C_j and H_j, assumed to equal minus one, so that the two goods are complements. Each agent has an initial endowment of two units of the first good and 1.5 units of the second (to be consistent with what comes later): $(e_{C_j}, e_{H_j}) = (2, 1.5)$, $j = 1, 2$. We then change agent 1's endowment to $(e'_{C_1}, e'_{H_1}) = (4, 1.5)$ while holding agent 2's endowment unchanged. Standard calculations imply that the relative price of good H in terms of good C increases from $P = 1.78$ to $P = \left((e'_{C_1} + e'_{C_2})/(e'_{H_1} + e'_{H_2})\right)^2 = 4$. Thus agent 2 becomes relatively poorer, while agent 1 becomes both relatively richer and absolutely richer in terms of the first good.

Agent 2's share of each good drops from half to one third of good C, while her share of aggregate wealth drops from half to 44%. Agent 1's share of each increases, so this is a case of an increase in inequality in which agent 1 becomes relatively and absolutely richer and agent 2 relatively poorer. However, agent 2's maximized utility (welfare) from optimal consumption increases, making her better off. The reason is that the change in endowments increases the relative scarcity of good H, with which agent 2 is relatively well endowed compared to agent 1 after the change in endowments. By selling some of her scarce endowment of good H at high prices to agent 1, agent 2 can acquire increased consumption of good C which more than compensates. Thus, in this two good endowment economy, an increase in inequality is not sufficient to reduce welfare.

By a reverse argument, let $(e_{C_2}, e_{H_2}) = (2, 1.5)$and $(e_{C_1}, e_{H_1}) = (4, 1.5)$ and change these endowments to $(e'_{C_2}, e'_{H_2}) = (e'_{C_1}, e'_{H_1}) = (2, 1.5)$. So inequality reduces (to zero), and agent 2 becomes relatively better off, but her welfare decreases. Therefore an increase in inequality is not necessary to reduce welfare

either. Thus in this endowment economy with two complements, inequality has no correspondence with welfare and does not matter. These results clearly extend to other values of ρ and to multiple goods and agents. Instead what matters for each agent is whether the ratio of their endowment of the two goods is less than or greater than the ratio of the aggregate endowment, which determines the price. That is, the terms of trade are what matter, not inequality, given the level of each agent's endowments. Furthermore, in multi-agent economies, transfers between other agents, leaving the aggregate endowment of each good unchanged, have no effect on prices, and therefore leave agent 2's welfare unaltered. The distribution of endowments simply does not matter for individual welfare. As we will see in the next subsection, this is no longer true if one of the goods is indivisible: In that case the distribution of endowments directly affects welfare.

Inequality matters when some goods are indivisible

In this section, we show how complementary, indivisible goods alter agents' standard objective functions and explain how goods are allocated between agents. Agents have endowments of a divisible good C, (the numeraire) resulting from gambles made in a prior first stage, which we discuss in sections 3 and 4. Agent j's endowment of divisible wealth at date 1 is E_j. There are two agents. (We discuss extensions to more agents in the conclusion.)

The economy is also endowed with M units of an indivisible good H. As agents are assumed to have CES-preferences over combinations of goods, we need each agent to have at least one unit of the indivisible good so as to guarantee utility levels above negative infinity. However, this is just a baseline setting. We also assume agents compete for additional units of the indivisible good. They make bids to buy, and the highest offer wins and is paid. The winning agent pays the losing agent the winning bid and receives the additional unit. If there are more additional units, the process continues until no further trading occurs.

In defining total net worth, we treat each agent as having an implicit endowment of $M/2$ units of the indivisible goods. For example, in the case of three units, she consumes one unit whose price is normalized to zero and either buys or sells her claim to the other half for a price P. The whole unit is therefore worth $2P$ so that the total value of aggregate wealth is preserved.

Just as in Ng (1965) and in Becker et al. (2005) (BMW), we assume the cross-partial derivative of utility $U_{CH} > 0$, which is equivalent to complementarity of the divisible and indivisible good. It follows that the richest agent will have the highest reservation price and will buy the additional unit.

For the rest of this subsection we assume $M = 3$. At the end we discuss how increasing M alters our results.

At a price P, agent j prefers to buy the unit provided

$$U(E_j - P, 2) > U(E_j + P, 1)$$

and not to buy otherwise. Agent j's reservation price $\overline{P}_j$ is then determined by

$$U(E_j - \overline{P}_j, 2) = U(E_j + \overline{P}_j, 1).$$

As claimed above,

$$E_j > E_k \Leftrightarrow \overline{P}_j > \overline{P}_k$$

and in the case of equal levels of divisible wealth, valuations are equal too.

Since agents are assumed to have complete information, the richest agent bids the poorer agent's reservation price (plus a penny) and wins the object. The price of the object thus depends only on the wealth of the poorer agent. That is, $P(E_1, E_2) = P(\min(E_1, E_2))$ solves

$$U(\min(E_1, E_2) - P(E_1, E_2), 2) = U(\min(E_1, E_2) + P(E_1, E_2), 1).$$

Welfare is then given by

$$V(E_j | E_1, E_2) = \max\left\{U(E_j - P(E_1, E_2), 2), U(E_j + P(E_1, E_2), 1)\right\}.$$

Figure 3.1 shows how welfare depends on relative endowments of divisible wealth. The x-axis increases the agent's own endowment holding the aggregate endowment constant (i.e. decreasing the other agent's endowment at the same time). The y-axis shows welfare V. The main point is that the richer agent has utility of the form $U(E_R - P(E_P), 2)$, that is increasing in her own endowment E_R but decreasing in the poorer agent's endowment E_P. Thus the richer agent has relative wealth concerns, which arise entirely endogenously in equilibrium. The poorer agent has welfare of the form $U(E_P + P(E_P), 1)$. While the poorer agent has no relative wealth concerns (in this two-agent setting), her welfare is a leveraged function of her own wealth.

Except at exact equality, where there is an inflection point, agents are locally risk-averse at all levels of wealth. However, over bigger gambles with outcomes above and below equality, agents are risk-loving, as in Friedman and Savage. As noted by Chetty & Szeidl (2007) in a different context, this form of welfare function addresses the Rabin (2000) critique of expected utility theory. Rabin shows that investors' attitudes to small and large risks are generally inconsistent with standard expected utility, because given how much agents appear to dislike small risks, they should avoid almost all large risks. But in our setting agents can dislike small risks and still like larger risks.

Between divisible wealth levels $\underline{E}$ and $\overline{E}$, both agents prefer to gamble. The dotted line tangent to V at both $\underline{E}$ and $\overline{E}$ represents the expected welfare obtainable by optimal gambles between the agents. (All convex combinations of any other pairs of divisible wealth levels would result in dotted lines lying strictly below this tangent line.) As all the points on the dotted line between these levels of divisible wealth lie above the points on the solid line beneath, agents prefer to gamble, with the richer agent obtaining higher probabilities of winning (the exact magnitude depends on bargaining power or, equivalently, the equilibrium criterion). Importantly, in this case the marginal utility of divisible wealth at both tangency points is the same. Thus, given an aggregate

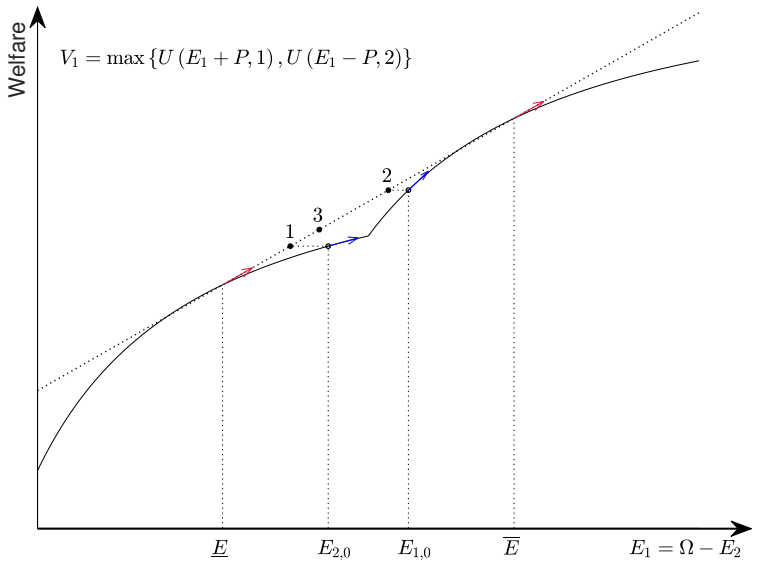


Figure 3.1

endowment of $\Omega = \underline{E} + \overline{E}$, the lower bound $\underline{E}$ is identified by

$$U'_C(\underline{E} + P(\underline{E}), 1) = U'_C(\Omega - \underline{E} - P(\underline{E}), 2).$$

The price function $P(E_1, E_2)$ goes to zero as the poorer agent's endowment of divisible wealth goes to zero, and is a positive and increasing function of the poorer agent's endowment. The richer agent, holding her own endowment constant, is better off as the poorer agent becomes poorer, since she pays less for the additional unit. Thus, the richer agent will accept gambles that offer either a positive expected return or a lower price for the additional unit if successful, provided either is sufficient to compensate for the risk of losing.

In the neighborhood of the inflection point, a small increase in the poorer agent's divisible endowment has a small effect on welfare, provided it doesn't thereby make her the richer agent, as she already has a lot of the divisible good.[1]

This observation can also rationalize status anxiety in this region: a small

[1]For the richer agent, marginal utility of divisible wealth is higher by

$$U_C(E_1 - P(E_2), 2) - U_C(E_2 + P(E_2), 1) \approx U_{CH}(E_2 + P(E_2), 1) + (E_1 - E_2 - 2P(E_2))U_{CC}(E_2 + P(E_2), 1) > 0.$$

decrease in the richer agent's wealth, holding the poorer agent's wealth constant, can have a large negative affect on welfare. In this respect, the welfare function depicted in figure 3.1 resembles habit close to the subsistence level (Campbell & Cochrane (1999)) or keeping up with the Joneses' preferences (Chan & Kogan (2002)). Unlike those utility functions, however, this status anxiety arises endogenously in equilibrium from competition between agents, and implies that agents wish to gamble.

When thinking about aggregate risk in section 4, we do not have analytic solutions and will use numerical methods. We therefore assume a particular log-CES form for utility[2]:

$$U(C,H) = \ln\left\{(C^\rho + H^\rho)^{\frac{1}{\rho}}\right\}.$$

3.3 Gambles without aggregate risk

We first consider a setting in which aggregate wealth is deterministic.An agent's outside option is not to gamble, in which case the richer agent (agent 1, whose wealth $E_1 = kE_2$, $k \geq 1$) acquires the additional unit from agent 2 for $P(E_2)$. For a gamble to happen, both agents must be weakly better off in expected utility terms compared to their outside option. Otherwise at least one agent refuses the gamble. A gamble is described by the triple $\{p, G_1, G_2\}$ where p is the probability that the poorer agent (agent 2) wins, G_1 is the transfer of wealth by 1 to 2 if 2 wins, and G_2 the transfer by 2 to 1 if 1 wins. As a

---

An identical increase in the richer agent's divisible wealth has a larger effect on her welfare because she has more units of the indivisible good (the first term on the RHS) but also because she pays P(E) and the other agent receives it (the second term when E_1 is close to E_2).

[2]The coefficient ρ is assumed to be negative, making the cross-partial derivative positive and the goods complementary. Setting $\rho = -1$ makes matters particularly simple, since then

$$P(E) = \sqrt{4 + E^2} - 2$$

and given

$$x = \sqrt{(2+\Omega)^2 + (\Omega^2 + 2\Omega)} - (2+\Omega)$$

$$\underline{E} = \frac{x^2 + 4x}{2(x+2)}.$$

solution criterion we use the Nash bargaining solution with equal bargaining power, which in the case of two agents just maximizes the sum of their expected utilities and then splits the surplus from gambling. Our results are qualitatively robust to using alternative solution criteria, such as giving one agent all of the bargaining power.

Given the discussion in the previous section, transfers G_1 and G_2 then simply move each agent to the two points $\underline{E}$ and $\overline{E}$, flipping the points depending on who wins. Thus, given initial wealth levels $E_{1,0}$ and $E_{2,0}$, $G_1 = E_{1,0} - \underline{E}$ and $G_2 = E_{2,0} - \underline{E}$.

If agent 1 has all the bargaining power, agent 2 is driven to her outside option, so that agent 2's expected welfare from the gamble V_2^G equals her utility from not gambling V_2^{NG}. This point is indicated, for an initial wealth level $E_{2,0}$, in Figure 3.1 as point 1. Not surprisingly, p is low at this point. Likewise if agent 2 has all the bargaining power, agent 1 is driven to her outside option, indicated by point 2 in Figure 3.1 for an initial wealth level $E_{1,0}$, with a high p. With equal bargaining power they split the surplus and the gamble is shown at point 3 in Figure 3.1. For two agents, this also corresponds to each agent receiving their Shapley value. Specifically, G_1 and G_2 maximize the total expected utility from gambling and p determines each agent's share of this surplus utility (relative to no gamble).

For extensions involving more than 2 agents, matters become slightly more complicated, so we use the Shapley value as our solution criterion. For 2 agents, these criteria give the same answers.

An alternative solution criterion (used by BMW and Garratt & Marshall (1994), among others) is to require all gambles to be actuarially fair in the sense that each agent's expected total wealth (including the value of implicit holdings of indivisible goods) is the same under the gamble as under no gambling. We use Nash bargaining instead for three reasons. First, we are interested in exploring circumstances in which standard risk-averse agents will accept gambles with negative expected returns. Second, the Nash Bargaining/Shapley criterion,

although not universal, is more standard in solving games of this type. Third, we will shortly introduce aggregate risk, in which agents' expected wealth varies across states under any criterion.

Solving for p gives

$$p = \frac{1}{2}\left(1 - \left(\frac{U(E_{1,0} - P(E_{2,0}), 2) - U(E_{2,0} + P(E_{2,0}), 1)}{U(\overline{E} - P(\underline{E}), 2) - U(\underline{E} + P(\underline{E}), 1)}\right)\right).$$

Figure 3.2 shows how p depends on k, the ratio $E_{1,0}/E_{2,0}$ when $\Omega = 4$ and $\rho = -1$. At $k = 1$, initial equality, $p = 0.5$. As k increases to its upper bound, $\overline{E}/\underline{E}$, p decreases to zero. At $k > \overline{E}/\underline{E}$, neither agent wishes to gamble. We show in the Appendix that the term in the numerator, $U(E_{1,0} - P(E_{2,0}), 2) - U(E_{2,0} + P(E_{2,0}), 1)$, increases from zero to an upper bound as E_1 increases (and E_2 decreases) from equality to $\overline{E}$.

Furthermore, provided the marginal utility of divisible consumption is convex, then we also show that $p < q$. That is

$$-\frac{U_{CCC}}{U_{CC}} \geq 0 \Rightarrow p < q \text{ for all } q \in (0, 0.5).$$

where q is the set of all actuarially fair gambles given by

$$q = \frac{1}{2} - \frac{1}{2}\frac{E_{1,0} - \Omega/2}{\overline{E} - \Omega/2}.$$

2n a one-good setting, the left hand side of this inequality is the coefficient of absolute prudence (Kimball (1990)). It holds for example for all constant and decreasing (DARA) absolute risk aversion functions. In a two-good model, the requirement relates not to the utility of the overall consumption bundle, but to the utility of the first good, divisible wealth. In pathological cases it does not hold even for DARA-preferences over the overall consumption bundle. Intuitively, if the complementarity between goods is strong enough, then additional consumption of the first good without additional consumption of the second good yields very little overall additional utility, and therefore may not be sufficient to decrease overall marginal utility of consumption at a sufficiently rapid rate.

For the case of CRRA-CES utility:

$$U(C,H) = \frac{x(C,H)^{1-\gamma}}{1-\gamma}$$

where

$$x(C,H) = (C^{\rho} + H^{\rho})^{\frac{1}{\rho}}$$

this sufficient condition holds, as we show in the Appendix, when

$$\gamma \geq \frac{-\rho - 1}{2}.$$

Thus for example, $\rho = -1$ requires $\gamma \geq 0$ and $\rho = -2$ requires $\gamma \geq 0.5$ before $-U_{CCC}/U_{CC}$ is non-negative. Since our main point is that there exist standard preferences under which risk-averse agents are willing to accept actuarially unfair gambles, we regard it reasonable to make the assumption that this inequality is satisfied in our model.

[FIGURE 2 ABOUT HERE]

The two agents split the surplus expected utility created by gambling, and aggregate wealth is fixed. At equality the two agent's outside options are equally good. As agent 1's wealth increases, agent 2's decreases, and so agent 1's outside option improves, while agent 2's deteriorates. Splitting the surplus therefore requires the deal obtained by agent 1 to be progressively better than the deal at equality of wealth, and so agent 1 must expect to make money.

[FIGURE 3 ABOUT HERE]

For agent 2, except at equality, the expected payoff on the gamble is negative, as shown below. Agent 2 expects to earn $pG_1 - (1-p)G_2$. This starts at zero when $k = 1$, and decreases as k increases, and then decreases again as $E_{2,0}$ approaches $\underline{E}$. This is because p decreases, but so does G_2. Figure 3.3 shows the relationship for the same case as Figure 3.2. Obviously, the expected payoff to agent 1 is positive. Therefore, in situations with scarce indivisible goods, or other real-world features which convexify equilibrium welfare functions such as status concerns, we should expect to see relatively poor agents with equal

bargaining power accept negative expected-return gambles when the gains from winning are greater than the losses from losing (i.e. when the payoffs from the gambles are positively skewed).

It is true that there still exist zero-expected-return gambles which both agents would strictly prefer to no gambling. However, given equal or superior bargaining power, agent 1 strictly prefers gambles with a positive expected return that agent 2 is also willing to accept over no gambling.

As an extension, in section 5 we consider a population of agents with many identical poor agents and one rich agent, all competing for one additional unit of the indivisible good. Individual poor agents are too poor to gamble directly with the rich agent, whose relative wealth is above the relevant upper bound. However, by first gambling among themselves, a winner can emerge from among the poorer agents whose winnings are large enough to allow them to compete with the rich agent. This possibility resembles the actual lotteries played by large numbers of poorer Amercians and Europeans, for example.

3.4 Aggregate risk

We now consider a scenario with aggregate risk. In particular, we allow the aggregate endowment to vary across two equally likely states $s = G, B$, such that for a given positive number ε and $\theta \geq 1$

$$\Omega(G) = (\theta + \varepsilon)\Omega$$

and

$$\Omega(B) = (\theta - \varepsilon)\Omega.$$

Here aggregate risk ε is a mean preserving spread of the zero-aggregate-risk case, and θ is the expected growth of the aggregate endowment. To begin, we consider what happens without gambles and without a riskfree asset, so that both agents are forced to bear all the aggregate risk. In the absence of gambles, inequality k cannot vary across states. (Since $E_{1,0} = kE_{2,0} = (k/(k+1))\Omega$,

$E_1(s) = kE_2(s)$ for all states s.) Since k is assumed to be greater than or equal to 1, agent 1 always acquires the additional unit and has higher welfare (except at equality).

Aggregate risk and welfare

The fact that agent 1 pays agent 2 her reservation price for the indivisible unit, which depends only on agent 2's wealth, makes the role of aggregate risk very different from an economy with all divisible goods. Essentially, agent 2's welfare is a leveraged function of her own wealth, while agent 1's welfare is a deleveraged function of agent 2's wealth. Essentially agent 2 provides insurance to agent 1 against the bad state through requiring a lower price for the indivisible good. This makes agent 2's marginal utility of divisible consumption much more sensitive to the aggregate endowment than agent 1's and therefore makes her more reluctant to bear aggregate risk. As far as we are aware, this insight is completely novel. It has the potential to explain lower participation in assets exposed to aggregate risk, such as equities, in a straightforward and yet non-behavioral fashion. It is also surely a very important contributing factor to why inequality matters for welfare.

Figure 3.4 plots the resulting welfare on the same axes as Figure 3.1, but showing welfare functions for two different states. In both states, the lower portion of the graph is the same. The point at which the two wealth levels are equal is, obviously, lower in the bad state. Above this point, the bad state welfare function lies above the good state welfare function until its upper bound at $E_j = (1 + \varepsilon)\Omega$. That is, holding divisible wealth E_1 constant across states, agent 1 is better off in the bad state, which is an intriguing result.

[FIGURE 4 ABOUT HERE]

Since relative divisible wealth cannot change, agent 1 always pays $P(E_2)$ to agent 2. Agent 2's welfare therefore depends only on her own wealth. Agent 1's welfare is increasing in her own wealth but decreasing in agent 2's wealth. Much of the benefit of an increase in aggregate wealth therefore accrues to

agent 2 through a higher price of the indivisible unit. In consequence, agent 1's consumption of the divisible good is not very sensitive to the state, whereas agent 2's is. The price that must be paid to agent 2 forms a natural hedge against aggregate risk for agent 1, and this hedge is provided by agent 2, for whom the opposite is true.

We thus have the following result: given consumption levels for each agent

$$\begin{aligned}\widetilde{C}_{1,1} &= \widetilde{E}_{1,1} - P\left(\widetilde{E}_{2,1}\right) \\ \widetilde{C}_{2,1} &= \widetilde{E}_{2,1} + P\left(\widetilde{E}_{2,1}\right)\end{aligned},$$

agent 2's volatility of consumption is greater than agent 1's provided

$$k < 1 + 2P'(\widetilde{E}_{2,1}).$$

This represents quite a high bound on k. Essentially, unless agent 1 is much richer than agent 2 at date zero, agent 2's consumption is more affected by the state at date 1. We provide the proof in the Appendix.

It follows that agent 1's marginal utility of consumption is not overly sensitive to the state over a wide range of k, whereas agent 2's marginal utility is much higher in the bad state than in the good state. Therefore, in this range, agent 2 has a stronger desire to rebalance resources across states, hence reducing the aggregate risk she bears.

It is worth recalling why this is not the case when the second good is divisible. In equilibrium, agents' marginal utilities will be very similar to each other whatever the state, even though they differ across states, because each agent can equate their marginal utilities of each form of consumption and each will have the same marginal rate of substitution. As a result, each agent has the same demand for rebalancing resources across states.

Linearizing each agent's expected utility under no gambles around their expected divisible wealth and taking expectations gives expected utility for each agent without gambling and being forced to bear all of aggregate risk:

$$E[U(C_{1,1}, 2)] = U(k\theta E_{2,0} - P(\theta E_{2,0}), 2) + \frac{1}{2}U_{CC}(k\theta E_{2,0} - P(\theta E_{2,0}), 2)Var[C_{1,1}]$$

and

$$E[U(C_{2,1},1)] = U(\theta E_{2,0} + P(\theta E_{2,0}),1) + \frac{1}{2}U_{CC}(\theta E_{2,0} + P(\theta E_{2,0}),1)Var[C_{2,1}].$$

The previous equation follows from the way in which $P(E_2)$ is determined. Both agents experience a reduction in expected welfare from an increase in aggregate risk, as is standard, of the form $0.5U_{CC}Var[C]$. Given the variance of agent 2's consumption increases by more than agent 1's, for a utility function with constant curvature with respect to divisible consumption we could immediately say that agent 2 is worse off. However, these curvatures are not constant, and therefore we cannot in general determine whether the higher consumption risk born by agent 2 reduces her welfare by more. It seems likely this is the case, and in our numerical example with $\rho = -1$ and $\theta\Omega = 4$ is always true.

Aggregate risk and gambling

Whenever agent 2 is worse off relative to agent 1 due to an increase in aggregate risk, her alternative to gambling is worse. Therefore, all else equal, the terms she can obtain for her gamble will worsen. We now discuss gambling in the presence of aggregate risk. We consider gambles entered into before the state is known, but whose outcome is determined after the state is realized.

Finding the optimal gamble in the presence of aggregate risk is a more involved problem than that discussed in section 3. A gambling contract C consists of$\{p, G1, G2\}$ just as before, but there are more constraints at work, and the utility of winning (losing) must be replaced by the expected utility of winning. That is agents choose C to solve

$$\begin{aligned}\max_C V_1^G + V_2^G &= pE[U_1(W,G_1,G_2)] + (1-p)E[U_1(L,G_1,G_2)] \\ &\quad + pE[U_2(W,G_1,G_2)] + (1-p)E[U_2(L,G_1,G_2)]\end{aligned}$$

where W denotes agent 2 winning and L agent 2 losing, so that for example $E[U_1(W, G_1, G_2)]$ is agent 1's expected utility under contract C when agent 2 wins (i.e. $E[U((\theta+\widetilde{\varepsilon})E_{1,0} - G_1 + P((\theta+\widetilde{\varepsilon})E_{1,0} + G_1), 1)]$). (In principle, agent 1 could still be richer than agent 2 in some states even after losing the gamble, but in our numerical analysis we find no case with an optimal gamble where this happens. With all optimal gambles that we observe, the poorer agent becomes relatively wealthier in both states if she wins[3].) Similarly $E[U_2(L, G_1, G_2)]$ is agent 2's expected utility under contract C when she herself loses, and so on.

The first two constraints require that neither agent can default in either state. Since both agents are worse off in the bad state, this requires that the transfers to be made under the gamble should be weakly less than the transferring agent's wealth in the bad state. That is

$$G_j < (\theta - \varepsilon)E_{j,0}.$$

Next, agents must increase their expected utility by gambling, as before. We again split the surplus between the players. This involves imposing a constraint on the maximization problem over C that

$$V_1^G - V_1^{NG} = V_2^G - V_2^{NG}.$$

For an optimal gambling contract, this constraint is binding, whereas none of the other constraints are. Let the Lagrange multiplier on this constraint be λ, then the Lagrangian is

$$\mathcal{L} = \max_C V_1^G + V_2^G + \lambda(V_1^{NG} - V_2^{NG} - (V_1^G - V_2^G)).$$

The following first order conditions define the optimal contract:

$$\frac{\partial \mathcal{L}}{\partial G_1} = (1-\lambda)\frac{\partial V_1^G}{\partial G_1} + (1+\lambda)\frac{\partial V_2^G}{\partial G_1} = 0$$

[3]We consider 'small' gambles in which the poorer agent only changes places with the richer in the bad state, conditional on winning the gamble, and find these small gambles are always dominated by large gambles.

which is equivalent to

$$(1+\lambda)E\left[\frac{dU(E_{2,1}(s,W)-P(E_{1,1}(s,W),2)}{dE_{2,1}(s,W)}\right]=(1-\lambda)E\left[\frac{dU(E_{1,1}(s,W)+P(E_{1,1}(s,W),1)}{dE_{1,1}(s,W)}\right].$$

Similarly $\partial\mathcal{L}/\partial G_2=0$ is equivalent to

$$(1+\lambda)E\left[\frac{dU(E_{2,1}(s,L)+P(E_{2,1}(s,L),1)}{dE_{2,1}(s,L)}\right]=(1-\lambda)E\left[\frac{dU(E_{1,1}(s,L)-P(E_{2,1}(s,L),2)}{dE_{1,1}(s,L)}\right].$$

In the zero risk case, the tangency wealth levels $\overline{E}$ and $\underline{E}$ are common to both agents. This is not the case in the more general setting with aggregate risk. Not only do these levels vary across agents, they also vary across states. These two first order conditions then identify two unkowns, G_1 and G_2, which pin down eight points, corresponding to optimal levels of divisible wealth for each agent in each state for each outcome of the gamble. Because these wealth levels vary across agents and states, the Lagrange multiplier λ is zero in the zero aggregate risk case, but not more generally. To solve for λ we also need a third first order condition which is $\delta\mathcal{L}/\delta p=0$. This is equivalent to

$$\begin{aligned}&(1-\lambda)E[U(E_{1,1}(s,W)+P(E_{1,1}(s,W),1)-U(E_{1,1}(s,L)-P(E_{2,1}(s,L),2)]\\ =\;&(1+\lambda)E[U(E_{2,1}(s,L)+P(E_{2,1}(s,L),1)-U(E_{2,1}(s,W)-P(E_{1,1}(s,W),2)].\end{aligned}$$

Given λ, G_1 and G_2, p is derived exactly as in section 3.

Then we have, for each state s

$$\underline{E}_{1,S}=E_{1,1}(s)-G_1,\ \overline{E}_{1,S}=E_{1,1}(s)+G_2$$

and

$$\underline{E}_{2,S}=E_{2,1}(s)-G_2,\ \overline{E}_{2,S}=E_{2,1}(s)+G_1.$$

These upper and lower wealth levels obtained by gambling are not tangency points or even expected tangency points, but are derived from the first order conditions when only a single gamble (involving one of two payoffs) is possible across both states. We could have considered state contingent gambles, in which case the contract C would depend on the state. We regard it as more realistic to set the terms of the gamble before the state is observed.

[FIGURE 5 ABOUT HERE]

Figure 3.5 shows how these upper and lower divisible wealth levels vary across states and across agents in the case of aggregate risk for $k = 1.2$, $\varepsilon = 0.4$ $\theta = 2$ (so $\theta\Omega_0 = 4$). First, the dotted line in Panel A shows welfare as a function of E_1 in the bad state, and the solid line in Panel B the same for the good state. The welfare functions of the other state are indicated in each panel to give the reader a sense of scale. If there were no aggregate risk, and instead the bad state were the only possible state, the lower and upper levels are indicated, and these are of course tangency points as in section 3. The corresponding expected utility levels for agents 1 and 2 are also indicated. Setting $\varepsilon = 0.4$, the blue squares indicate upper and lower wealth levels for agent 1, while the red squares indicate them for agent 2. Both agents' bounds are inflated outwards relative to the zero risk case, with both of agent 1's bounds lying below agent 2's. Agent 2's expected utility is slightly lower than in the absence of aggregate risk in this bad state, while agent 1's is only very slightly higher.

In the good state (panel B) these patterns are reversed. Agent 1's bounds both lie above agent 2's and both sets of bounds are deflated inwards, that is move towards equality, relative to the zero risk case. Agent 2's expected utility is unchanged relative to the zero risk case, while agent 1 is better off in expected utility terms.

These graphs show that the degree of manufactured inequality depends on the level of aggregate risk. In the good state, manufactured inequality is lower than when $\varepsilon = 0$, while in the bad state it goes up. However, Figure 3.5 clearly shows that the increase in the bad state is smaller than the decrease in the good state. So aggregate risk reduces the expected amount of manufactured inequality. (But of course requiring both agents to bear 100% of aggregate risk.)

Perhaps just as important, this exercise shows that the degree of manufactured inequality in the presence of aggregate risk also depends on the level of initial inequality k. Figure 3.6 is identical to Figure 3.5 in every way except

it also includes the case of $k = 1.6$. As can be discerned for $k = 1.2$ versus $k = 1.6$, some points move inwards and some outward in each state. Overall, conditional on the poorer agent winning, expected inequality increases when k increases. By contrast, conditional on agent 1 winning, expected inequality decreases. Because the probability of agent 2 winning is low, expected manufactured inequality is lower when initial inequality k is higher in the presence of aggregate risk.

[FIGURE 6 ABOUT HERE]

Figure 3.7 plots how p depends on k for different values of ε, thus generalizing Figure 3.2. p decreases more rapidly with k at higher levels of ε. Social mobility depends not just on inequality but also on the level of aggregate risk.

[FIGURE 7 ABOUT HERE]

Despite the findings in Figure 3.7, the expected return to agent 1 on the gamble, although still positive, is actually decreasing in aggregate risk. The reason is to do with the decline in manufactured inequality noted above: as aggregate risk increases, p decreases for any given level of $k > 1$. However, the transfers G_1 and G_2 also decrease, and in agent 2's favour. The expected return on the gamble though remains negative.

In summary, aggregate risk alters the terms of the optimal gamble between agents, reducing the amount of expected inequality manufactured by the gamble, and improving the terms for the poorer agent, but not by enough to make the gamble actuarially fair for her.

The choice of how much risk to bear

In this last section, we examine what happens when agents are allowed to choose the level of aggregate risk they bear. A riskless asset with initial price B pays riskless return $R_f = X_f/B$ at date 1, and a risky asset with initial price Q pays risky return $R(s) = X(s)/Q$. The supplies of the riskless and risky assets are respectively S_f and S_X.[4] Thus, we have

[4]The ratio S_f/S_X is arbitrary.

$$\begin{aligned}\Omega_0 &= BS_f + QS_X \\ \Omega(s) &= X_f S_f + X(s) S_X\end{aligned}$$

where

$$X(s) \equiv \frac{\theta\Omega - X_f S_f}{S_X} + \frac{\Omega_0}{S_X}\varepsilon(s).$$

Given a positive supply of the riskless asset, the risky asset is then a leveraged claim on the aggregate endowment and will have a higher expected growth rate.

At date 0, each agent allocates share α_i of initial wealth to the risky asset. Date 1 wealth $E_{i,1}(s)$ and return on wealth are

$$\begin{aligned}E^n_{i,1}(s, \alpha_i \mid R(s), R_f) &= r^n_i(s, \alpha_i \mid R(s), R_f) E_{0i} \\ r^n_i(s, \alpha_i \mid R(s), R_f) &= (1-\alpha_i)R_f + \alpha_i R(s).\end{aligned}$$

Market clearing requires

$$\begin{aligned}\alpha_1 E_{01} + \alpha_2 E_{02} &= QS_X \\ (1-\alpha_1)E_{01} + (1-\alpha_2)E_{02} &= BS_f\end{aligned}$$

which implies the equilibrium rates of return:

$$\begin{aligned}\tilde{R}(s, \alpha_1, \alpha_2) &= \tfrac{(\theta+\varepsilon(s))\eta(s)}{\eta_0(\alpha_1,\alpha_2)} \\ \tilde{R}_f(\alpha_1, \alpha_2) &= \tfrac{(\theta+\varepsilon(s))(1-\eta(s))}{1-\eta_0(\alpha_1,\alpha_2)}\end{aligned}$$

where

$$\eta(s) = \frac{X(s) S_X}{\Omega(s)}$$

and

$$\eta_0(\alpha_1, \alpha_2) = \frac{QS_X}{\Omega_0} = \frac{\alpha_1 k_0 + \alpha_2}{1+k_0}.$$

At date 0, agents may wish to enter a gamble $C \equiv \{p, G_1, G_2\}$. Consequently, the two agents face the following problems:

- In the asset market, given a gambling contract C and expected returns $E[R(s)]$ and R_f, two agents individually choose their optimal portfolio allocation $\hat{\alpha}_j(\alpha_{-j}, C, R(s), R_f)$ to maximize their expected utility.

- In the gambling market, given their equity allocations (α_1,α_2), two agents enter into a gambling contract subject to their both obtaining higher expected utilities than under no-gambling.The optimal gamble splits the surplus available from gambling (as in the rest of the paper) by Nash bargaining.

- Equilibrium occurs when both the asset market and the gambling market clear.

In the case of a divisible second good, both agents allocate the same proportion of wealth to the risky asset, and do not gamble, since they have identical risk preferences. By contrast, Figure 3.8 shows the share of wealth allocated to the risky asset by the richer and poorer agent against ε. As is standard, the allocation of both agents is decreasing in ε, but unlike the standard case, the poorer agent's allocation is lower than the richer agent's. These results are almost identical when the agents are unable to gamble.

[FIGURE 8 ABOUT HERE]

Since the risky asset is a leveraged claim on the aggregate endowment, an agent with a higher share of wealth invested in the risky asset will have a higher expected growth rate of wealth than the growth rate of the aggregate endowment, which in turn will grow faster (in expectation) than the wealth of the agent with the lower share. Thus, our results imply that the richer agent should on average increase their wealth faster than the economy can grow, resulting over time in a higher share of aggregate wealth. Thus, our equilibrium is consistent with Piketty's observation that returns on capital appear to be higher than returns for the aggregate economy.

Now that agents hold different shares of wealth in risky assets, given the previous observation, the degree of inequality of wealth k is expected to increase from date zero to date 1. In principle this alters the terms of the gamble entered into by the two agents. However, when we examine the optimal gamble in the presence of a risk-free asset we find very little difference from the gambles

discussed in the previous subsection, with all of our qualitative results still holding.

3.5 Lotteries

We now discuss a particular extension of our two-agent equilibrium, with N identically poor agents and one rich agent. The poor agents have initial wealth $E_{n,0}$ and the rich agent $E_{1,0}$. The poor agents are sufficiently poor that the rich agent prefers no gamble to a gamble with any individual poor agent (and therefore no individual poor agent wishes to gamble with the rich agent either). That is $E_{n,0} < \underline{E}$, where $\underline{E}$ solves

$$\frac{dU(E + P(E)/n, 1)}{dE} = -\frac{dU(E_{1,0} + E_{n,0} - E - P(E), 2)}{dE}.$$

The aggregate endowment Ω is fixed, as in section 2. Provided the aggregate wealth of the poor agents is sufficiently large, the poor agents strictly prefer to participate in an actuarially fair lottery with each other, excluding the rich agent, in which each bets G_n with a probability $1/n$ of winning the jackpot nG_n.

They have two separate motivations for doing so. Either of these is sufficient, under different constraints, to give rise to lottery participation. First, if all agents own an equal fraction $1/(n+1)$ of the additional unit of the indivisible good, they will participate in a lottery because the winner of the lottery is in a position to bid for the unit herself, with wealth $E_{2,0} = E_{n,0} + (n-1)G_n$. Because her reservation price is much higher than the mass of poor agents, every poor agent shares in the payout $P(\min(E_{2,0}, E_{1,0}))/n$. This is the case even if the lottery winner and the rich agent do not then go on to gamble between each other. This strikes us as a rational justification for populism (see **?**), which we define as any democratically mandated policy with a non-positive net present value. In this example, the populist policy increases the share of wealth of the vast mass of the population, particularly involving the elevation

of one of the masses to greatness. We term this case "Populism" and discuss it in detail below.

The second motivation each agent has to participate in such a lottery is that if they win, they may be rich enough to participate in a gamble with the rich agent. This can also be, on its own, sufficient motivation for all poor agents to participate in such a lottery. This is seen by the fact that even if the poor agents get no payout from the sale of the indivisible unit, with the two rich agents (the lottery winner and the intially rich agent) sharing the proceeds entirely between themselves, the poor agents still prefer to participate in a lottery. This reminds us of the famous movie The Cincinnati Kid (starring Steve McQueen and Edward G. Robinson) in which the Kid (McQueen) spends time building his stake by playing and beating many poor poker players in order to have enough to challenge the high-roller Lancey Howard (Robinson). We term this case "Cincinnati Kid."

Finally, we allow all agents to share in the payout and the winner and the rich agent to gamble. Such a lottery is strictly less beneficial to the lottery losers, since they receive $P(\underline{E})/n < P(\min(E_{2,0}, E_{1,0}))/n$. We call this lottery "Populist Kid." In this lottery both motivations to participate are present, but each weakens the other. We characterize the optimal lottery (in terms of G_n given Ω, $E_{1,0}$ and n) in each case.

Characterizing the full set of optimal gambles between all agents for any discrete distribution of initial wealth is beyond the scope of this paper. However, we consider it useful that a popular and universal phenomenon, the purchase of lottery tickets, should so naturally arise in our framework. Our only prediction is that lottery participants should be relatively poor, at least when all lottery players are identical. We also note that very rich agents would strictly prefer there to be no lottery. For this reason, the rich agent may decide to act as a spoiler by participating in a lottery on a large scale. Although we consider this an interesting possibility, and hope to explore it in future drafts, we assume for now that the rich agent is simply not allowed by the poor agents to participate.

Populism

In the absence of a lottery, poor agents receive

$$V_n^{NL} = U(E_{n,0} + P(E_{n,0})/n, 1)$$

where $P(E_{n,0})$ solves

$$U(E_{n,0} + P(E_{n,0})/n, 1) = U(E_{n,0} - P(E_{n,0}), 2).$$

Since these agents are too poor to gamble directly with the rich agent, $P(E_{n,0})$ will be small.

Under the rules of our populist lottery, the identically poor agents each buy one ticket for G_n. Losers will have divisible wealth of $E_{n,0} - G_n$ and the winner $E_{n,0} + (n-1)G_n = E_{2,0}$. Suppose that $E_{2,0} \leq E_{1,0}$, so that the lottery winner is still poorer than the rich agent. Then expected utility from playing this lottery is

$$V_n^L = \frac{n-1}{n} U\left(E_{n,0} - G_n + P\left(E_{2,0}\right)/n, 1\right) + \frac{1}{n} U\left(E_{2,0} + P\left(E_{2,0}\right)/n, 1\right).$$

A second-order Taylor expansion of the terms on the right-hand side of this expression around $E_{n,0}+P(E_{n,0})/n$ for the first argument shows that $V_n^L > V_n^{NL}$ whenever utility of consumption of divisible wealth is not too concave:

$$-\frac{U_{CC}\left(C_{n,0}, 1\right)}{U_C\left(C_{n,0}, 1\right)} < \frac{2P'\left(E_{n,0}\right) G_n^{-1}}{\left(-1 + (n-1)\, P'\left(E_{n,0}\right)/n\right)^2 + (n-1)\left(1 + P'\left(E_{n,0}\right)/n\right)^2}.$$

The right-hand side of this expression is an upper bound on the absolute curvature of the utility function with respect to divisible wealth (which is related to but not the same as risk aversion). This bound depends only on n (decreasing in n, as more players make the lottery less attractive) and $E_{n,0}$ (increasing in $E_{n,0}$) for a given ticket price G_n. This means that the willingness to play lotteries like this is independent of the rich agent's wealth, because the price the rich agent will end up paying depends only on the wealth of the winner of this lottery. For many triples $(n, E_{n,0}, G_n)$ this constraint is satisfied,

and so lotteries like this will occur in the presence of scarce, complementary indivisible goods.

This result may prove to be fruitful in explaining the popularity of populist policies. Because the poor agents are strictly better off participating in this lottery, they will also be willing to play a similar but actuarially unfair lottery with a slightly lower payoff or with a slightly lower probability of winning (where perhaps the lottery is costly to run, or requires a payment to a third party). This could rationalize policies which 'violate budget constraints' and appear to leave the poor and middle classes worse off on average (**?**) by forcing rich players to pay more. Of course, it depends on how much more, and who shares in the surplus.

Figure 3.9 shows how expected utility of this lottery varies with $G_n/E_{n,0}$ for a given level of $E_{1,0}$, $\gamma = 1$ (as in previous sections' graphs). We set $n = 100$, $E_{n,0} = 3$, and $E_{1,0} = 10$ (so that $\Omega = 310$). The dashed line is the baseline utility of poor agents from not participating in the lottery. The solid line plots their expected utility V_n^L from playing the lottery. What can be seen from this figure is that the optimal level of G_n is quite high: playing the lottery strictly dominates not playing for all ranges of G_n up to at least five per cent of wealth, but the optimal lottery fee is markedly higher than three per cent. Although a stylized example from a simple model, this looks encouraging in explaining why people play the much larger lotteries we actually observe in the presence of far greater inequality.

Finally we note that populist lotteries like this one may be vulnerable to free-riding and also to spoiling action by the richer agent. We intend to explore this issue in future drafts. There is no free riding problem with the next type of lottery we consider.

Cincinnati kid

An alternative motivation for participating is the chance to gamble with the rich agent if you win. In this lottery, losers receive nothing from the sale of the

additonal unit: the two richest agents share the pot between themselves.

With no lottery the poor agents' utility is $V_n^{NL} = U(E_{n,0}, 1)$ as the rich agent pays nothing for the additional unit. Given a lottery G_n, the poor agents' expected utility is

$$V_n^L = \frac{n-1}{n} U(E_{n,0} - G_n, 1) + \frac{1}{n} V_2^G$$

where

$$\begin{aligned} V_2^G &= \tfrac{1}{2}\left(U\left(\bar{E} - P(\underline{E}), 2\right) + U(\underline{E} + P(\underline{E}), 1)\right) \\ &\quad + \tfrac{1}{2}\left(U(E_{2,0} - P(E_{min}), 2) - U(E_{1,0} - P(E_{min}), 2)\right) \end{aligned}$$

and $E_{min} = \min\{E_{1,0}, E_{2,0}\}$, as in section 2.

Consider a lottery $G_n^\dagger$ in which the lottery winner is just able to gamble with agent 1: $E_{2,0} = E^\dagger = E_{n,0} + (n-1) G_n^\dagger$ where $E^\dagger$ is identified by

$$U_C\left(E^\dagger + P\left(E^\dagger\right), 1\right) = U_C\left(E_{1,0} - P\left(E^\dagger\right), 2\right).$$

$E^\dagger$ is a monotonically increasing function of $E_{1,0}$:

$$\frac{dE^\dagger}{dE_{1,0}} = \frac{U_{CC}\left(E_{1,0} - P\left(E^\dagger\right), 2\right)}{U_{CC}\left(E^\dagger + P\left(E^\dagger\right), 1\right)\left(1 + P'\left(E^\dagger\right)\right) + U_{CC}\left(E_{1,0} - P\left(E^\dagger\right), 2\right) P'\left(E^\dagger\right)} > 0$$

The poor agents' expected utility is therefore

$$V_n^L = \frac{n-1}{n} U\left(E_{n,0} - G_n^\dagger, 1\right) + \frac{1}{n} U\left(E^\dagger + P\left(E^\dagger\right), 1\right).$$

Whenever

$$-\frac{U_{CC}(C_{n,0}, 1)}{U_C(C_{n,0}, 1)} < \frac{2P\left(E^\dagger\right)}{\left(E^\dagger - E_{n,0}\right)^2 / (n-1) + \left(E^\dagger - E_{n,0} + P\left(E^\dagger\right)\right)^2}$$

poor agents will willingly participate in this lottery.

This lottery relies on a different motivation: the opportunity for a winner to gamble with the richer agent. This is also a sufficient reason to engage in a

lottery, but appears to hold over a much smaller range of lottery prices. Figure 3.10 shows how the optimal ticket price varies in this lottery, given the all the same paramters as in Figure 3.9. The optimal lottery ticket price G_n is 0.2% of initial wealth, and expected utility is decreasing in G_n over the entire range in which agents prefer playing the Cincinnati Kid lottery. At G_n greater than roughly 0.8% of initial wealth, agents no longer wish to play this lottery. This motivation is therefore much weaker than populism, although immune to free riding.

Populist kid

Finally we consider the combination of the two lotteries: all agents share in the payout when the richer agent buys, and the winner gambles with the rich agent.

With no lottery the poor agents' utility is $V_n^{NL} = U\left(E_{n,0} + P\left(E_{n,0}\right)/n, 1\right)$. Given a lottery G_n, their utility is

$$V_n^L = \frac{n-1}{n} U\left(E_{n,0} - G_n + P\left(\underline{E}\right)/n, 1\right) + \frac{1}{n} V_2^G$$

where

$$\begin{aligned} V_2^G &= \tfrac{1}{2}\left(U\left(\bar{E} - P\left(\underline{E}\right), 2\right) + U\left(\underline{E} + P\left(\underline{E}\right)/n, 1\right)\right) \\ &\quad + \tfrac{1}{2}\left(U\left(E_{2,0} - P\left(E_{min}\right), 2\right) - U\left(E_{1,0} - P\left(E_{min}\right), 2\right)\right) \end{aligned}$$

and $E_{min} = \min\left\{E_{1,0}, E_{2,0}\right\}$, as in section 2.

Consider a lottery $G_n^\dagger$ in which the lottery winner is just able to gamble with agent 1: $E_{2,0} = E^\dagger = E_{n,0} + (n-1)\, G_n^\dagger$ where $E^\dagger$ is identified by

$$U_C\left(E^\dagger + P\left(E^\dagger\right)/n, 1\right)\left(1 + P'\left(E^\dagger\right)/n\right) = U_C\left(E_{1,0} - P\left(E^\dagger\right), 2\right)\left(1 + P'\left(E^\dagger\right)\right).$$

$E^\dagger$ is again a monotonically increasing function of $E_{1,0}$[5]:

[5]The sufficient condition for $dE^\dagger/dE_{1,0} > 0$ is $P''(E) > 0$, which is always true for a CRRA-CES preference.

$$\frac{dE^{\dagger}}{dE_{1,0}} = \frac{A\left(1+P'\left(E^{\dagger}\right)\right)}{B\left(1+P'\left(E^{\dagger}\right)/n\right)^{2} + A\left(1+P'\left(E^{\dagger}\right)\right)P'\left(E^{\dagger}\right) - \left(1-1/n\right)CP''\left(E^{\dagger}\right)/\left(1+P'\left(E^{\dagger}\right)\right)} > 0$$

where $A = U_{CC}\left(E_{1,0} - P\left(E^{\dagger}\right), 2\right)$, $B = U_{CC}\left(E^{\dagger} + P\left(E^{\dagger}\right)/n, 1\right)$, and $C = U_{C}\left(E^{\dagger} + P\left(E^{\dagger}\right)/n, 1\right)$.

The poor agents' expected utility is thus

$$V_{n}^{L} = \frac{n-1}{n}U\left(E_{n,0} - G_{n}^{\dagger} + P\left(E^{\dagger}\right)/n, 1\right) + \frac{1}{n}U\left(E^{\dagger} + P\left(E^{\dagger}\right)/n, 1\right).$$

A sufficient condition for the existence of a lottery $G_{n}^{\dagger}$:

$$-\frac{U_{CC}\left(C_{n,0}, 1\right)}{U_{C}\left(C_{n,0}, 1\right)} < \frac{2P'\left(E_{n,0}\right)}{\left(1 + 1/\left(n-1\right) + P'\left(E_{n,0}\right)^{2}/n\right)\left(E^{\dagger} - E_{n,0}\right)}.$$

Figure 3.11 shows how expected utility for the poor agents varies with G_{n} for the same parameters as the previous two figures. Because the lottery is populist, the range of G_{n} over which players prefer to play this lottery is wider than for the Cincinnati Kid one (increasing from 0.25% to 1.2% of initial wealth, and again from 2.1% to 4.3%). The optimal ticket price is about 2.9% of initial wealth. Again quite high, but lower than for the purely Populist lottery.

Poorer agents strictly prefer the Populist to Populist Kid to Cincinnati Kid to no lottery. Winners of the lottery strictly prefer Cincinnati Kid to Populist Kid to Populist to no lottery. Rich Agents strictly prefer no lottery to Populist; prefer Populist to either of the other two types; and are indifferent between Cincinnati and Populist Kid.

In summary, when there is large number of identically poor agents, they will choose to participate in a lottery of one kind or another between themselves which will have at least one of two effects: first, to increase the price to be paid by the rich agent for the indivisible good by raising the wealth, and therefore the reservation price, of the lottery winner; second, to offer one lottery winner the opportunity to participate in gambles with the rich agent. Clearly the

optimal lottery, from the point of view of the poor agents, should have only one winner in our framework. For Populism, having more than one winner reduces the benefit to the non-winning agents. For CK, having more than one winner makes the lottery more expensive to play. In future drafts we will examine the scope for free-riding by poor agents and for spoiling behaviour by the rich agent. For example, the rich agent could herself participate in the lottery. Alternatively, she could pay a few poor agents not to participate, reducing the payout to the winner. We intend to explore the results in this section in the context of political economy. They could, for example, explain the popularity of populism, which is otherwise puzzling.

3.6 Conclusion

In this paper, we present a general result and some numerical examples of optimal risk-taking in an endowment economy with two complementary consumption goods, one of which is indivisible. With two agents and one additional unit of the indivisible good, the agent with greater divisible wealth acquires the unit and enjoys higher welfare. This can rationalize attitudes to inequality that are harder to rationalize in a standard divisible setting. Agents prefer to enter into gambles with each other, and the poorer agent expects to reduce their total wealth under such gambles, except in the limit of exact equality. In a setting without aggregate risk, p measures social mobility, and is strictly decreasing as the poorer agent becomes relatively poorer. In this setup, inequality is 'manufactured' as in Rosen (1987) and BMW, so taxation-funded residtribution of initial wealth will not change ultimate inequality, which over some range will increase.

Adding aggregate risk to such an equilibrium makes the poorer agent worse off in expected utility terms, relative to the richer agent, which worsens the terms of the gambles from the perspective of the poorer agent. The poorer agent also prefers to hold less aggregate risk than the richer agent. When

the aggregate endowment has positive expected growth, this means that the poorer agent invests less in 'good' risks. Thus the richer agent's total wealth is expected to grow faster than the aggregate endowment. This model can thus rationalize simultaneous participation in negative-expected-return lotteries and under-participation in positive-expected-return risks such as the aggregate endowment (at least in a two-agent, one-unit setting). Poorer agents are both more likely to buy lottery tickets and underinvest in stocks.

When there are three agents, the richest agent will pay the reservation price for the additional unit of the middle agent, rather than the poorest agent. Holding the bottom and top of the wealth distribution constant, poor (rich) agents are better (worse) off when the median agent's wealth level increases.

In an extension with multiple identical poor agents and one rich agent, we prove that agents prefer to participate in lotteries, either to create a rich winner who will drive up the price paid by the rich agent, or simply for the small chance of being able to gamble with the rich agent. This theory can easily rationalize participation in zero- or even negative-expected return lotteries by poorer agents with standard concave utility functions.

With more heterogenous agents, and more units, it is hard to say at present how these results change. As agents' wealth levels converge to each other, the prices agents pay converge to their own reservation prices and the incentives to gamble diminish. It is therefore probably correct to say that the crucial assumption in our model is that of a discrete wealth distribution.

In future papers we intend to study further implications of the basic framework, such as those for the risk-free rate and the equity premium. There are many studies that try to build on Friedman and Savage's original paper. Yet the literature remains oddly disparate and has failed to explain any large-scale puzzles in economics in a way that fits into the general canon. (Exceptions include **?** and also some contributions in corporate finance such as Basak et al. (2018).) We hope this paper will be a successful call to incorporate the demand for indivisible goods into the general framework.

3.7 Appendix

Proof of $p < q$ for all values in $(0, 0.5)$.

$$\begin{aligned} q\left(E_{1,0}\right) &= \tfrac{1}{2} - \tfrac{1}{2}\tfrac{E_{1,0}-\Omega/2}{\overline{E}-\Omega/2} \\ p\left(E_{1,0}\right) &= \tfrac{1}{2} - \tfrac{1}{2}\left(\tfrac{U(E_{1,0}-P(E_{2,0}),2)-U(E_{2,0}-P(E_{2,0}),2)}{U\left(\overline{E}-P(\underline{E}),2\right)-U(\underline{E}-P(\underline{E}),2)}\right) \end{aligned}$$

$$q > p \Leftrightarrow f\left(E_{1,0}\right) > f\left(\overline{E}\right) \Leftarrow f'\left(E_{1,0}\right) < 0$$

where $f\left(E_{1,0}\right) = \frac{U(E_{1,0}-P(E_{2,0}),2)-U(E_{2,0}-P(E_{2,0}),2)}{E_{1,0}-\Omega/2}$, given $U\left(E_{2,0} - P\left(E_{2,0}\right), 2\right) = U\left(E_{2,0} + P\left(E_{2,0}\right), 1\right)$ defines $P(E_{2,0})$.

$f\left(E_{1,0}\right)$ can be approximated by taking a Taylor expansion of $U\left(E_{1,0} - P\left(E_{2,0}\right), 2\right)$ around $\left(E_{2,0} - P\left(E_{2,0}\right)\right)$ and simplifying (noting that $E_{1,0} - E_{2,0} = E_{1,0} - (\Omega - E_{1,0}) = 2(E_{1,0} - \Omega/2)$:

$$f\left(E_{1,0}\right) \approx 2\left(U_C\left(E_{2,0} - P\left(E_{2,0}\right), 2\right) + U_{CC}\left(E_{2,0} - P\left(E_{2,0}\right), 2\right)\left(E_{1,0} - \Omega/2\right)\right).$$

Therefore,

$$q > p \Leftrightarrow f'\left(E_{1,0}\right) < 0 \Leftrightarrow -\frac{U_{CCC}\left(E_{2,0} - P\left(E_{2,0}\right), 2\right)}{U_{CC}\left(E_{2,0} - P\left(E_{2,0}\right), 2\right)} > \frac{P'\left(E_2\right)}{\left(P'\left(E_2\right) - 1\right)\left(E_{1,0} - \Omega/2\right)}.$$

The RHS of this inequality is negative, since differentiation of the identity $U\left(E_{2,0} - P\left(E_{2,0}\right), 2\right) = U\left(E_{2,0} + P\left(E_{2,0}\right), 1\right)$ with respect to $E_{2,0}$ establishes that $0 < P' < 1$. Therefore a sufficient condition for $q > p$ is that the LHS is non-negative, as claimed.

For CRRA-CES utility, $U = x(C, H)^{1-\gamma}/(1-\gamma)$

$$U_{CC}\left(C, H\right) = -\gamma x^{-\gamma-1} x_C^2 + x^{-\gamma} x_{CC} < 0$$

where $x_C = x^{1-\rho} C^{\rho-1} > 0$ and $x_{CC} = (1-\rho)\, x_C C^{-1}\left(x^{-\rho} C^{\rho} - 1\right) < 0$.

$$U_{CCC}\left(C, H\right) = x_C x^{-\gamma} C^{-2} \underbrace{\left(aY^2 + bY + c\right)}_{g(Y)})$$

where $Y = x^{-\rho} C^{\rho}$, $a = (2\rho + \gamma - 1)\,(\rho + \gamma - 1)$, $b = 3\,(1-\rho)\,(\gamma - 1 + \rho)$ and $c = (1-\rho)\,(2-\rho)$.

Note that $Y \in (0,1)$ and $g(0) = (1-\rho)(2-\rho) > 0$, $g(1) = \gamma(\gamma+1) > 0$. Therefore, a sufficient condition for $g(Y) > 0$ is $\gamma \geq \frac{-\rho-1}{2}$, in which case $-U_{CCC}/U_{CC} > 0$.

Proof that $Var[C_{2,1}] > Var[C_{1,1}]$ for $1 < k < 1 + 2P'(\widetilde{E}_{2,1})$.

Taylor expand $P(E_{2,1})$ around $\theta E_{2,0}$:

$$P(E_{2,1}) = P(\theta E_{2,0}) + P'(\theta E_{2,0})(\varepsilon E_{2,0}) = \alpha + \rho E_{2,0}\varepsilon$$

where $\alpha = P(\theta E_{2,0})$, $\rho = P'(\theta E_{2,0}) = \frac{\theta E_{2,0}}{\sqrt{4+(\theta E_{2,0})^2}}$ and $0 < \rho < 1$.

Let σ_ε^2 be the variance of shock ε. Then consumption variances are:

$$\begin{aligned} var(C_{1,1}) &= var(E_{1,1} - P(E_{2,1})) \\ &= var(E_{1,1}) + var(P(E_{2,1})) - 2cov(E_{1,1}, P(E_{2,1})) \\ &= E_{1,0}^2\sigma_\varepsilon^2 + \rho^2 E_{2,0}^2\sigma_\varepsilon^2 - 2\rho E_{1,0}E_{2,0}\sigma_\varepsilon^2 \\ &= (k^2 + \rho^2 - 2\rho k) E_{2,0}^2\sigma_\varepsilon^2 \\ &= (k-\rho)^2 E_{2,0}^2\sigma_\varepsilon^2 \end{aligned}$$

$$\begin{aligned} var(C_{2,1}) &= var(E_{2,1} + P(E_{2,1})) \\ &= var(E_{2,1}) + var(P(E_{2,1})) + 2cov(E_{2,1}, P(E_{2,1})) \\ &= E_{2,0}^2\sigma_\varepsilon^2 + \rho^2 E_{2,0}^2\sigma_\varepsilon^2 + 2\rho E_{2,0}^2\sigma_\varepsilon^2 \\ &= (\rho+1)^2 E_{2,0}^2\sigma_\varepsilon^2 \end{aligned}$$

where

$$\begin{aligned} &var(E_{1,1}) = var((\theta+\varepsilon)E_{1,0}) = E_{1,0}^2\sigma_\varepsilon^2 \\ &var(E_{2,1}) = var((\theta+\varepsilon)E_{2,0}) = E_{2,0}^2\sigma_\varepsilon^2 \\ &var(P(E_{2,1})) = var(\alpha + \rho E_{2,0}\varepsilon) = \rho^2 E_{2,0}^2\sigma_\varepsilon^2 \\ &cov(E_{1,1}, P(E_{2,1})) = cov((\theta+\varepsilon)E_{1,0}, \alpha + \rho E_{2,0}\varepsilon) = \rho E_{1,0}E_{2,0}\sigma_\varepsilon^2 \\ &cov(E_{2,1}, P(E_{2,1})) = cov((\theta+\varepsilon)E_{2,0}, \alpha + \rho E_{2,0}\varepsilon) = \rho E_{2,0}^2\sigma_\varepsilon^2 \end{aligned}$$

Therefore

$$var(C_{2,1}) > var(C_{1,1}) \Leftrightarrow \rho + 1 > k - \rho.$$

Proof:

Each agent's valuation is increasing in $U(E_j - P, 2) - U(E_j + P, 1)$. Agent 1's valuation is greater if and only if

$$U(E_1 - P, 2) - U(E_1 + P, 1) > (U(E_2 - P, 2) - U(E_2 + P, 1)).$$

Adding $U(E_1 - P, 1) - U(E_1 - P, 1)$ to the left-hand side of the inequality and $U(E_2 - P, 1) - U(E_2 - P, 1)$ to the right hand side gives

$$\begin{aligned} & U(E_1 - P, 2) - U(E_1 - P, 1) + U(E_1 - P, 1) - U(E_1 + P, 1) \\ > \; & U(E_2 - P, 2) - U(E_2 - P, 1) + U(E_2 - P, 1) - U(E_2 + P, 1). \end{aligned}$$

Rearranging gives

$$\begin{aligned} & \{(U(E_1 - P, 2) - U(E_1 - P, 1)) - (U(E_2 - P, 2) - U(E_2 - P, 1))\} \\ & + \{(U(E_2 + P, 1) - U(E_2 - P, 1)) - (U(E_1 + P, 1) - U(E_1 - P, 1))\} \\ > \; & 0. \end{aligned}$$

The first term in brackets is positive if and only if $E_1 > E_2$ since $U_{CH} > 0$ - the utility gain $U(E_j - P, 2) - U(E_j - P, 1)$ from an additional unit of the indivisible good is greater for an agent with more of the divisible good. The second term in brackets is also positive if and only $E_1 > E_2$ since $U_{CC} < 0$ - the utility gain $U(E_j + P, 1) - U(E_j - P, 1)$ from an additional $2P$ of the divisible good is greater for an agent with less of it. This argument shows why complementarity of the two goods is crucial for indivisible goods to give rise to a demand for gambles.This proof holds also in the extreme case of lexicographic preferences.

3.8 Figures

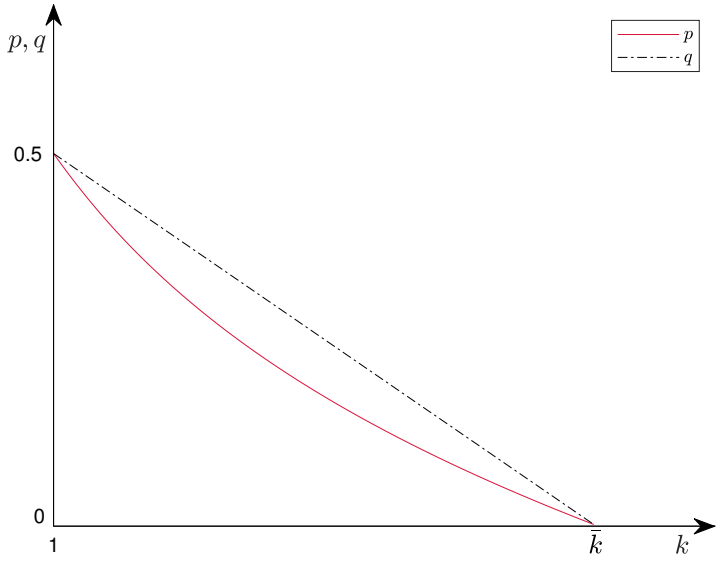


Figure 3.2

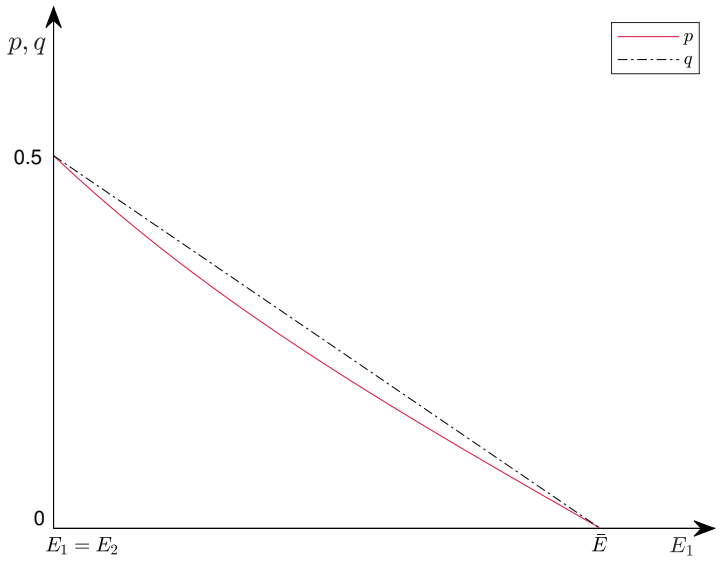


Figure 3.3

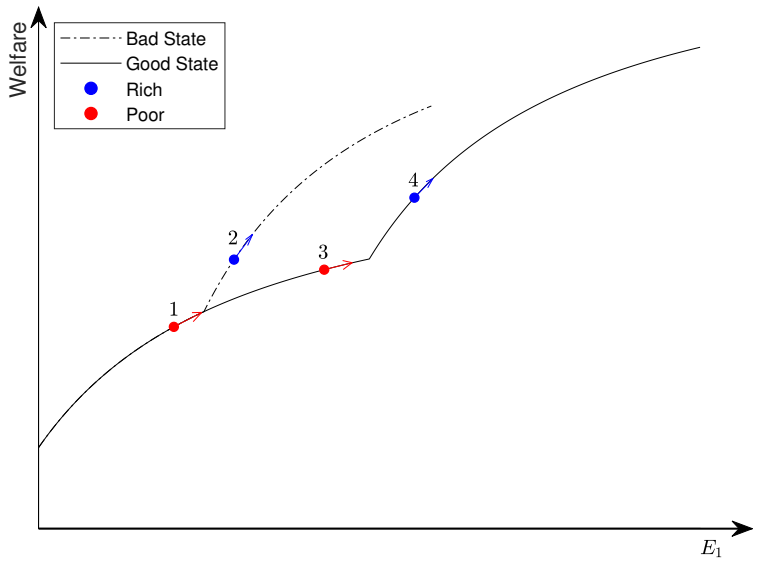
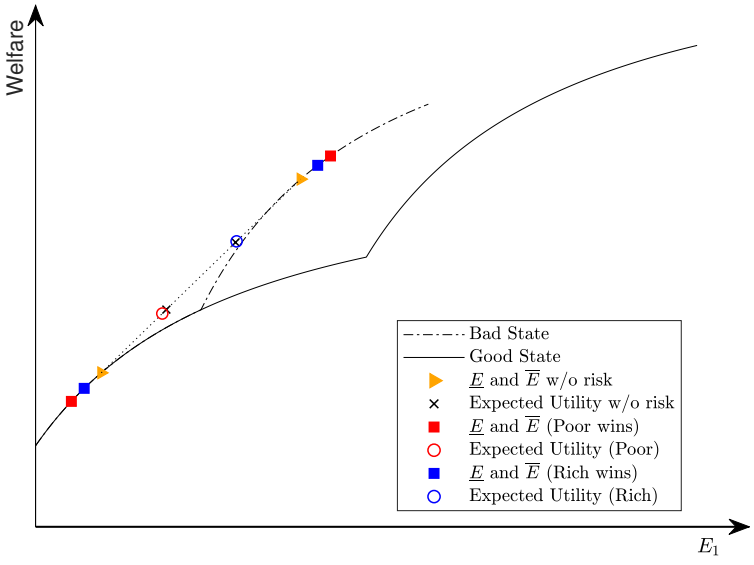
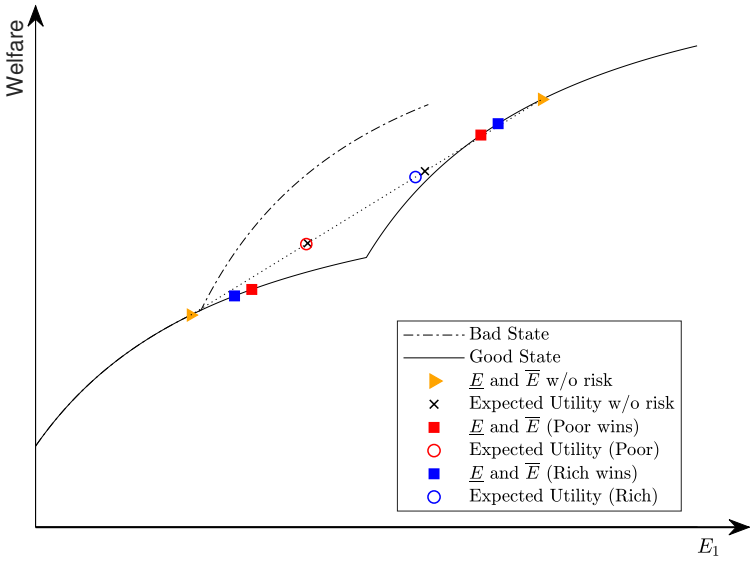


Figure 3.4

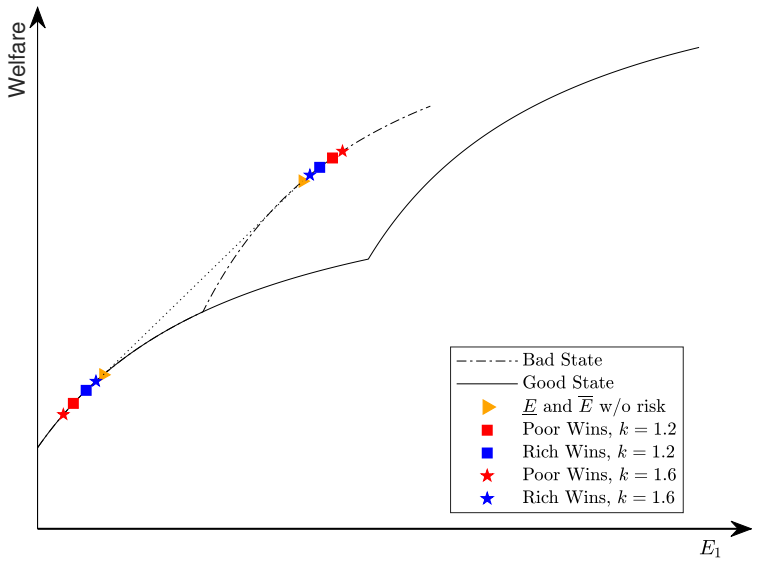

Welfare
Bad State
Good State
$\underline{E}$ and $\overline{E}$ w/o risk
Expected Utility w/o risk
$\underline{E}$ and $\overline{E}$ (Poor wins)
Expected Utility (Poor)
$\underline{E}$ and $\overline{E}$ (Rich wins)
Expected Utility (Rich)
E_1

(a)

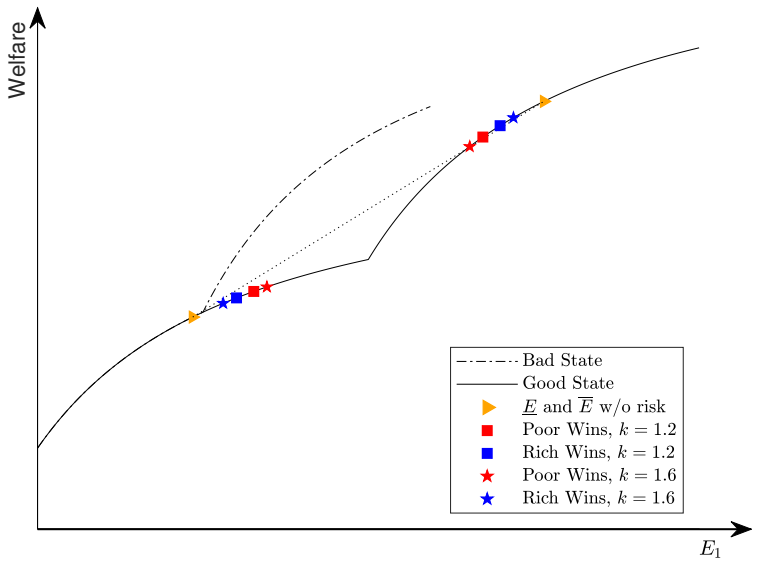

Welfare
Bad State
Good State
$\underline{E}$ and $\overline{E}$ w/o risk
Expected Utility w/o risk
$\underline{E}$ and $\overline{E}$ (Poor wins)
Expected Utility (Poor)
$\underline{E}$ and $\overline{E}$ (Rich wins)
Expected Utility (Rich)
E_1

(b)

Figure 3.5


Welfare
Bad State
Good State
$\underline{E}$ and $\overline{E}$ w/o risk
Poor Wins, $k = 1.2$
Rich Wins, $k = 1.2$
Poor Wins, $k = 1.6$
Rich Wins, $k = 1.6$
E_1

(a)


Welfare
Bad State
Good State
$\underline{E}$ and $\overline{E}$ w/o risk
Poor Wins, $k = 1.2$
Rich Wins, $k = 1.2$
Poor Wins, $k = 1.6$
Rich Wins, $k = 1.6$
E_1

(b)

Figure 3.6

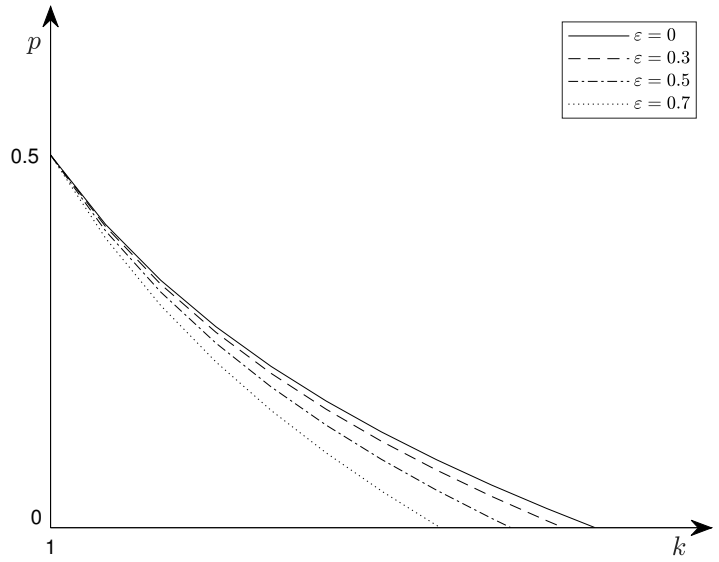


Figure 3.7

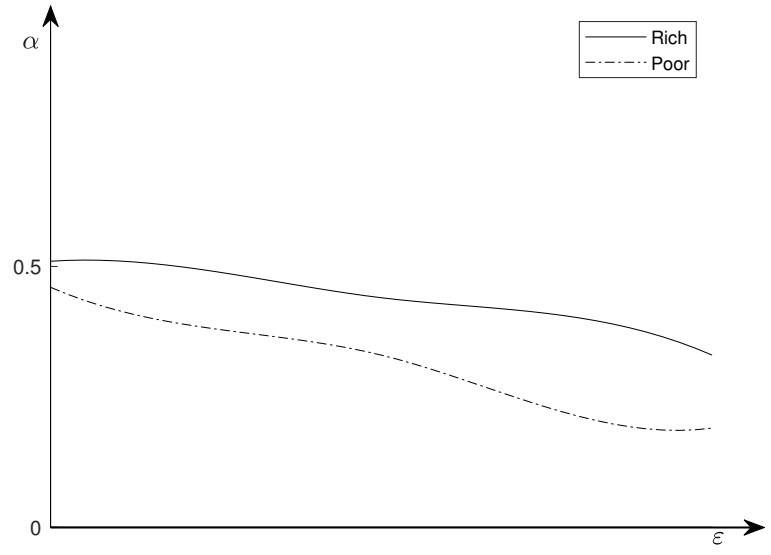


Figure 3.8

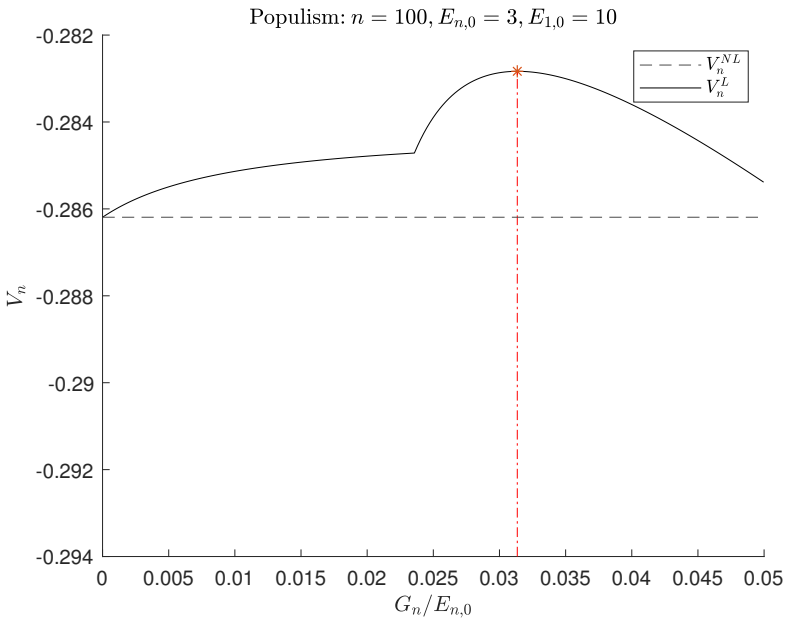


Figure 3.9

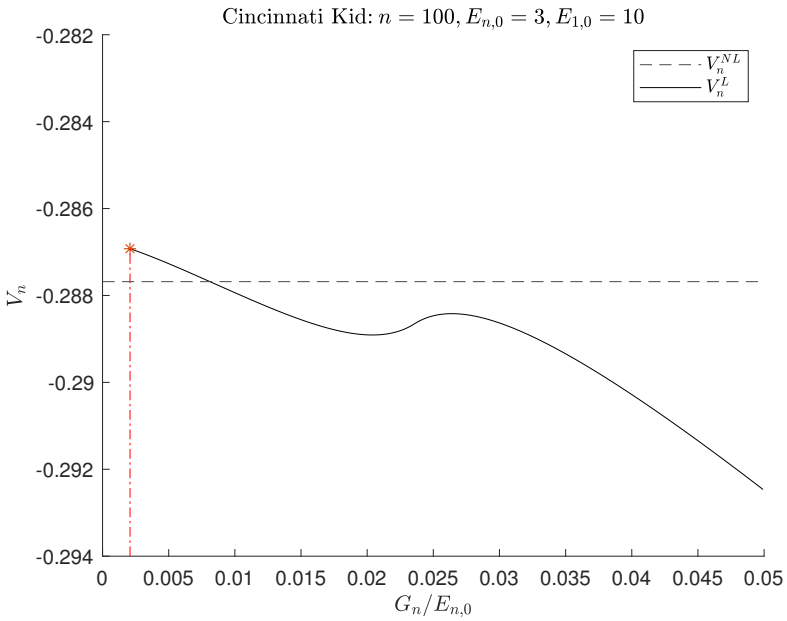


Figure 3.10

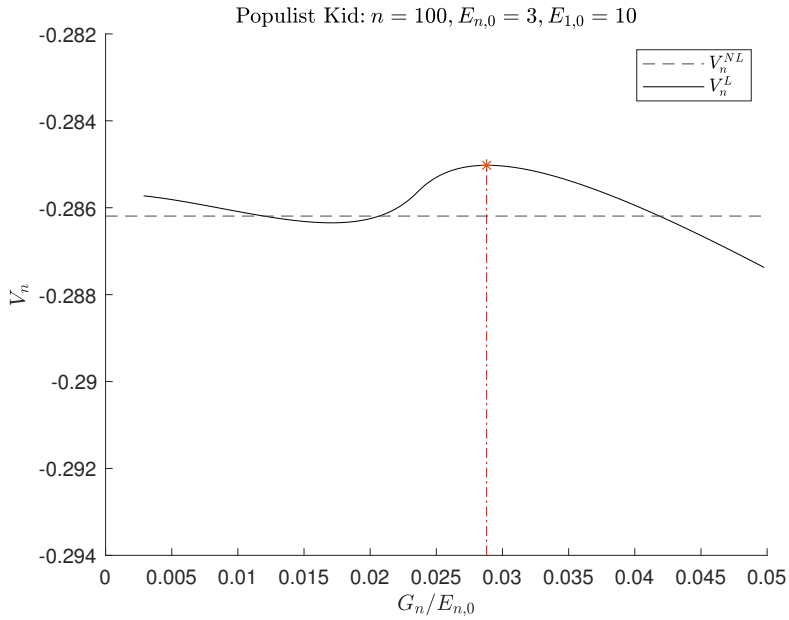

Populist Kid: $n = 100, E_{n,0} = 3, E_{1,0} = 10$
V_n^{NL}
V_n^L
-0.282
-0.284
-0.286
-0.288
-0.29
-0.292
-0.294
V_n
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
0.05
$G_n/E_{n,0}$

Figure 3.11

Chapter 4

The stock market isn't a sideshow: How granular instrumental variables reveal the effect of stock market volatility on the real economy.

4.1 Introduction

Do stock market price fluctuations cause fluctuations in real economic activity? This is surely the most important question in financial economics. It was also, until recently, considered to be a question that was impossible to answer: Stock market valuations are forward-looking and therefore changes in valuations anticipate changes in real economic activity as well as, possibly, causing them. Plausible identification of stock price fluctuations unrelated to expectations of future economic activity was practically impossible. Therefore it was impossible to be reasonably certain that a correlation between stock price fluctuations at date t followed later by changes in economic activity were evidence of any kind of causal relationship.

In this paper I propose instrumental variables which are plausibly unrelated to future economic activity such as GDP growth, consumption and investment. Using these instruments, I can identify shocks to stock market prices which are random (exogenous) with respect to subsequent changes in GDP, consumption,

investment etc. This enables me to offer a first answer to the question posed in the previous paragraph. I find that stock market index increases cause increases in consumption and investment, and therefore also in aggregate GDP of the United States between 1993 and 2019, and index decreases cause decreases, with a stronger effect on the side of decreases in some categories of investment. I go further and investigate some of the channels through which these stock price changes are transmitted to the real economy. Notably, I find that: Firms experiencing exogenous stock price increases exploit these increases to issue more equity, and use the subsequent proceeds to fund capital expenditure while at the same time reducing retained earnings growth. The magnitudes of the causal relationship are often somewhat larger than implied by naive estimates relying on correlations: The stock market is not a sideshow (Morck et al. (1990)). As I discuss in more detail below, in an average year, the incremental GDP created or destroyed in the USA by exogenous stock market return shocks amount to the size of the economy of Finland, or about $270 billion, about 1.3% of US GDP.

In an efficient market in which prices embed only rational forecasts of future cashflows, stock prices cannot affect real variables, but can only respond to revisions in expected future cashflows (or discount rates). Since Grossman & Stiglitz (1980) , financial economists have recognised that prices can also move in response to so-called 'noise' or 'noninformational' (Wang (1994)) trades, which (as their name implies) are not motivated by information about fundamental value. In the resulting equilibium, prices are not usually fully informationally efficient in the sense that agents can expect to profit from costly information-gathering about fundamental value. The question raised by Morck et al. (1990) was whether these noise trades affect real activity, or whether they merely redistribute wealth between lucky winners and unlucky losers. Ever since, the problem has been to identify a set of noise trades that move prices.

The literature has proposed four such sets of noise trades which are plausi-

ble: Dividends, mandate changes, index composition changes and flow-induced trades. Some of these proposals have been subject to recent criticism (see e.g. Wardlaw (2020). I use price pressure from dividend payments, following Harzmark and Solomon (2021) [REF], in particular, as an alternative instrument to check the robustness of my results. One of the problems with all of these sources of instruments for price changes is that they are weak (Staiger & Stock (1997)), making causal inference almost impossible. However, using an alternative identification strategy based on the insights in Harzmark et al., I estimate an almost identical impact of flow-induced price pressure on GDP growth as I do with my principal instrumental variable strategy, again implying an almost exactly Finland-sized average creation and destruction of US GDP in an average year.

Very recently, Gabaix & Koijen (2021) have proposed their so-called Granular IV (GIV) combing the equity-share-weighted idiosyncratic components of flows between major sectors of investors who hold listed equity. I use this instrument to construct an exogenous component of stock market price changes and then examine how this exogenous component affects the real economy, obtaining the results I describe above. The intutition behind GIV is that only the common component to flows in and out of equities should be related to changing common perceptions of fundamental values, with the idiosyncratic component only reflecting differences in beliefs, preferences, or immediate financing requirements, none of which should be related to fundamentals. These differences have an impact on stock prices, as opposed to the noise trades in say Grossman and Stiglitz, because they are intermediated by wealth management funds with manadates (tighter or looser) to hold a fairly constant position in equities. If all investors perfectly understood these manadates and sought to implement their own exactly calculated portfolios through offsetting positions in funds with difering mandates, the price-impact effects of these mandates would unravel, but in practice this clearly never happens. (Consider, for example, your own pension fund and your holdings of various actively and passively

managed mutual funds.)

There are several related literatures, most of which are effectively reviewed in Gabaix et al. For example, a host of papers such as Sirri & Tufano (1998) find evidence that retail investment in actively managed mutal funds resemble dumb money chasing a successful return period, that mutual funds use these flows to invest more in their existing portfolios (Pollet & Wilson (2008), and that these flows consequently impact prices (Lou (2012)), consistent with equity flows into even fairly discretionary mandates driving up demand and therefore prices of stocks. The rise of passive investing in recent decades has almost certainly strengthened this phenomenon, although that question is not the subject this paper.

If my results are correct, semi-informed flows in and out of stocks drive quite consequent changes in real activity, consistent with the claims made by Keynes (1936) almost a century ago. This has obvious implications for counter-cyclical monetary and fiscal policy, as well as being consistent with some of the findings in empirical corporate finance (e.g. Baker & Wurgler (2000)). I discuss some of these later in the paper.

Perhaps the first modern study to look at whether stock prices impact the real economy isFischer & Merton (1984). This was a purely theoretical paper arguing why stock market volatility might influence returns. Morck et al. is a later example that utilizes some evidence. Boadly, corporate managers may use stock markets mispricing to issue overpriced equity, may use higher stock prices to alleviate financial constraints, or may take their lead as to when and how to invest from high stock prices. My results remain somewhat agnostic as to which of these possiblilities predominate, although as mentioned I find very strong equity issuance increases following exogenous stock market increases.

The rest of the paper is organised as follows: Section 2 briefly recounts how Granular IV works in my setting, following Gabaix & Koijen (2021) and gives first-stage results; section 3 presents my main results; section 4 presents some results on how financing by firms responds to price shocks; section 5 presents

results from an alternative identification strategy using price pressure arising from dividend payments (as opposed to announcements); section 6 concludes.

4.2 Instrumental variables for stock market prices

This section discusses the problem of identifying price-driven changes in economic activity. I give an example using aggregate dividend flow based on Hartzmark & Solomon (2022) which illustrates all of the issues clearly. Then I explain the flow-based instrument I construct along the lines of Gabaix & Koijen (2021), and argue that it causes exogenous variation in prices with respect to real economic activity.

Dividends

Consider the following simple example. Suppose much of investor wealth is held through an index fund with a mandate to invest entirely in, for example, the Russell 2000 index in proportion to market weights. The index fund managers consider that their fiduciary duty to their shareholders is to minimize tracking error against their benchmark. When a constituent of the index announces a dividend, its own and other stock prices may move in response to the announcement, but this is clearly endogenous: Market participants revise their estimates of value based on their interpretations of the signal given by the dividend announcement. However, when the same dividend is paid, some weeks later, to the holder of record, no news is revealed by the latter event. Yet the index fund now has more cash than its target level, and so as to minimize tracking error will reinvest the dividend payment in the index as quickly as possible. Any increase in the index associated with the receipt of large dividends is therefore almost certainly exogenous, responding purely to a flow based on a cash receipt announced weeks earlier.

Hartzmark & Solomon (2022) show that the index price responds at a daily frequency to aggregate dividend receipts. I reproduce their results in table 4.7 for the period 1926-2022, which I discuss in the last section. The main point is that the buying pressure, and associated index price increases, induced by these dividend receipts is completely exogenous with respect to fundamental value of stock market-listed businesses, news about future economic activity, or indeed anything else.

Although exogenous, such dividend price pressure is too weak an instrument to identify an effect of stock market price pressure on real activity, as I show in the final section. This motivates a search for a stronger instrument,

Why is GIV a good instrument for stock price to address endogeneity concern?

There is a representative stock with a fixed number of shares Q and a risk-free bond in perfectly elastic supply. The economy lasts for two periods, $t = 0, 1$. At time 0, the stock price is P and the risk-free bond price is fixed and normalised to 1. The stock pays a dividend D per share at time 1. The equity premium is $\pi = \frac{E[D]}{P} - 1 - r_f$, where $E[D]$ is the expected dividend at time 0 and r_f is the risk-free rate. The dividend-price ratio is $\delta = \frac{E[D]}{P}$. The excess equity premium $\hat{\pi}$ is the deviation of the equity premium from the average equity premium $\hat{\pi} = \pi - \bar{\pi}$.

Assume that fund i's demand for stocks is given by a mandate to invest a fixed share of wealth in equity:

$$\frac{PQ_i}{W_i} = \theta_i e^{\kappa_i \hat{\pi}} \tag{4.1}$$

where Q_i is fund i's stock holding, W_i is fund i's total wealth (or assets under management), θ_i is the fraction of fund i's wealth invested in equity, and κ_i represents fund i's risk preference.

Suppose at time 0, a representative household invests $\triangle F_i$ extra dollars in fund i. The values at time 0^-, before this extra flow comes into the fund, are

baseline values and denoted by bars. At time 0^-, fund i has wealth $\bar{W}_i$ and $\bar{Q}_i$ shares, the stock price is $\bar{P}$, expected dividend is $\bar{D}$, dividend-price ratio is $\overline{\delta}$ and $\hat{\pi} = 0$. It can be shown that the (log) percentage change in fund i's wealth is

$$\triangle \ln W_i \approx \frac{\triangle W_i}{\bar{W}_i} = \theta_i \triangle p + f_i \tag{4.2}$$

,

where $\triangle p = \frac{P - \bar{P}}{\bar{P}}$ and $f_i = \frac{\triangle F_i}{\bar{W}_i}$.

Taking logs of (4.1) and subtracting its baseline value gives the (log) percentage change of equity demand:

$$\triangle \ln Q_i = \triangle \ln W_i - \triangle \ln P + \kappa_i \hat{\pi}. \tag{4.3}$$

(4.2) and (4.3) give the percentage change in fund's demand (with log approximation):

$$\triangle q_i := \frac{\triangle Q_i}{\bar{Q}_i} = -\zeta_i \triangle p + \bar{\delta} \kappa_i d + f_i \tag{4.4}$$

where $\zeta_i = 1 - \theta_i + \bar{\delta}\kappa_i$ and $d = \frac{E[D]}{\bar{D}} - 1$.

Define $\triangle q$ to be size-weighted aggregate demand:

$$\triangle q = \sum S_i \triangle q_i = -\zeta \triangle p + \bar{\delta} \kappa d + f$$

where $S_i = Q_i / Q$, $\zeta = \sum S_i \zeta_i$, $\kappa = \sum S_i \kappa_i$ and $f = \sum S_i f_i$.

Because the net amount of equity purchase is 0, i.e. $\Delta q = 0$, the equilibrium price (change) is

$$\triangle p = \frac{\bar{\delta} \kappa d}{\zeta} + \frac{f}{\zeta} \tag{4.5}$$

This simple two-period model (4.5) shows that the size-weighted flow f can have large price impact when the price elasticity ζ is low, which is negatively associated with investors' average mandate of equity share θ.

Next I will show how to use flow data to construct an instrumental variable for the time-series of stock price changes. Thus I will add time subscripts to variables to facilitate explanation. Assume that fund i's flow can be modeled as:

$$f_{it} = \lambda_i^{'} \eta_t + u_{it}$$

where η_t is a vector of common shocks to which all investors have exposure, $\lambda_i^{'}$ is a vector of factor loadings, and u_{it} is an idiosyncratic shock. The idiosyncratic shock is defined to be independent of the common shocks:

$$E\left[\eta_t u_{it}\right] = 0.$$

Thus the price can be decomposed into three parts:

$$\triangle p_t = \underbrace{\frac{1}{\zeta}\bar{\delta}\kappa d_t}_{\tilde{A}} + \underbrace{\frac{\lambda}{\zeta}\eta_t}_{\tilde{B}} + \underbrace{\frac{1}{\zeta}\sum S_i u_{it}}_{\tilde{C}}$$

$\tilde{A}$ relates to the stock's fundamentals; $\tilde{B}$ is the price impact of trading against common shocks, such as macroeconomics news; $\tilde{C}$ represents noise trading, which may be caused by difference in beliefs or preferences. The GIV instrument is defined as

$$Z_t := \sum S_i u_{it}.$$

Clearly, Z_t is correlated with $\triangle p_t$, and thus satisfies the relevance condition:

$$cov\left(\triangle p_t, Z_t\right) \neq 0.$$

Since d_t is a common source of variation in flows, news about it is contained in η_t and therefore Z_t is also independent of d_t.

Suppose that we want to identify the causal relationship between stock price and the real economy activity, such as GDP growth:

$$\triangle gdp_t = \alpha + \beta \triangle p_t + \xi_t$$

where ξ_t is a shock to GDP growth. If the shock to aggregate economy ξ_t is highly correlated with the common shock η_t to flows, which is highly possible, $\triangle p_t$ will be correlated with ξ_t through η_t:

$$cov\left(\triangle p_t, \xi_t\right) \propto cov\left(\triangle p_t, \eta_t\right) \neq 0.$$

This endogeneity problem, however, can be solved by GIV Z_t, under a weak assumption: $E\left[\xi_t u_{it}\right] = 0$, i.e., the idiosyncratic disturbance to funds' flow is independent of the shocks to aggregate economy.

Thus,

$$cov\left(\xi_t, Z_t\right) = 0,$$

which satisfies the exclusion restriction. As a result, Z_t is a valid instrument for $\triangle p_t$.

Data sources and construction

I construct an instrumental variable for stock price changes using the Flow of Funds (FoF) data. FoF reports quarterly flows and levels of equity holdings in all sectors of economy, such as Households and nonprofit organizations, Mutual funds, etc.

Let $W^{\varepsilon}_{i,t}$ and $\triangle F^{\varepsilon}_{i,t}$ denote the level and the flow of sector i's equity holding at time t, respectively. $\triangle F^{\varepsilon}_{i,t} = W^{\varepsilon}_{i,t} - W^{\varepsilon}_{i,t-1} R^X_t$, where R^X_t is the capital gain which the FOF use to calculate equity flows to adjust for price effects.

The relative change in equity demand for sector i, is:

$$\triangle q_{i,t} := \frac{\triangle Q_{i,t}}{Q_{i,t-1}} = \frac{\triangle F^{\varepsilon}_{i,t}}{W^{\varepsilon}_{i,t-1}} \left(R^X_t\right)^{-1}$$

I next isolate idiosyncratic flow shocks by removing the impact of common shocks to individual demand for stock. I run the panel regression

$$\triangle q_{i,t} = \alpha_i + \beta_t + \gamma_i \triangle y_t + \phi_i t + \triangle \hat{q}_{i,t}$$

where $\triangle y_t$ is real GDP growth. I take $\triangle \hat{q}_{i,t}$, the residuals from this regression and then extract the first principal component of $\triangle \hat{q}_{i,t}$. The principal component, denoted by η_t^{PC}, can capture more common shocks additional to that of total output. Then I regress $\triangle \hat{q}_{i,t}$ on η_t^{PC} and take the residuals as $\hat{u}_{i,t}$, the idiosyncratic demand shocks to stocks. The GIV instrument is calculated as either[1]:

$$Z_t = \sum S_{i,t-1} \widehat{q}_{i,t},$$

or else as

$$Z_t^{PC} = \sum S_{i,t-1} \widehat{u}_{i,t},$$

I then conduct two-stage least squares regression. In the first stage I run

$$\Delta p_t = a + bZ_t$$

or else replacing Z_t with Z_t^{PC}, in which case I find none of my results change materially.

and in the the second stage I use the fitted value from this regression, $\Delta \widehat{p}_t$ as my explanatory variable in regressions of changes in GDP and its components. Results are in the next section. If the assumptions about GIV are valid, then my results can be interpreted as causal impacts of aggregate stock market changes on real economic activity.

GIVs: Data and first-stage results

Table 4.1 presents summary statistics about my GIV variables, which are quarterly from 1993Q1 to 2019Q4. Z is on the first row and Z^{PC} on the second.

[1]The construction of GIV requires idiosyncratic shocks u_{it} to be homoskedastic. When $u_{i,t}$ are heteroskedastic with variance σ_i^2, $u_{i,t}$ need to be weighted by their inverse variance $\tilde{E}_i = \sigma_i^{-2} / \left(\sum_{n=1}^{N} \sigma_n^{-2} \right)$ to account for heteroskedasticity. If trading flows of some sectors are very stable, this inverse variance weighing method will put very high weight on these sectors,

Both variables have very similar distributions, having means of roughly zero, and standard deviations, respectively, of 0.6% and 0.5% in quarterly data. The diseprsion in returns is almost identical, with a lower minimum for Z of -2.1% versus -1.6% for Z^{PC}, and a slightly higher maximum for Z at 1.2% versus 1% for Z^{PC}. The 25th percentile is the same for both at -0.4%, median the same at zero, and 75th percentile at 0.3%. Z has a quarterly serial correlation of 0.35 and Z^{PC} of 0.37, so both variables are somehwat persistent compared to returns.

Average stock market returns, excluding dividends, are given in row 3. Mean quarterly returns averaged 1.4% with a standard deviation of 8.2%, a minimum of -24.6% and maximum of 19.7%. The 25th percentile quarterly return was -1.7%, the median 2.5%, the 75th percentile 5.7%. Serial correlation was less than 0.1. Thus this sample period is a fairly typical one for the stock market given the longer-term record.

Table 4.2 presents results of regressions of the CRSP value-weighted return excluding dividends on my GIV instrument, being the first stage of my two-stage least squares procedure. Column (1) gives a highly significant coefficient on Z^{PC} of 7.17 with a standard error of 1.9 and an R^2 of 21.6%. Column 2 gives a coefficient on Z_t of 8.1 with a standard error of 1.77 and an R^2 of 30.1%. Column (3) regresses Δp_t on the first pricinpal component of $\widehat{q}_{i,t}$ and reports a coefficient of 2.7 with a standard error of 0.5 and an R^2 of 41.2. In my second stage results, I do not find my results to be materially affected by using either Z or Z^{PC}, so I end up following Gabaix and Laibson and use Z because of its slightly greater strength.

Crucial to the rest of this study is that neither Z nor Z^{PC} are weak instruments for stock market return. The F-statstics for each are 14.1 for Z^{PC} and 20.9 for Z. Staiger & Stock (1997) recommend an F-statistic of at least

which will distort the estimation of the idiosyncratic flow shocks. I follow Gabaix & Koijen (2021) to winsorize the weights at $\frac{1.5}{N}$, which is an arbitrary and significant departure from equal weights. The panel regression uses these winsorized weights $\tilde{E}_i$ as regression weights, and the weighted residuals $\tilde{E}_i \triangle \hat{q}_{i,t}$ are used in the principal component analysis.

10 before an instrument is strong enough to carry out causal inference. This is one of the great advantages of flow-based GIVs.

Having expalined how my instrument works both theoretically and empirically and shown that it is not weak, I now present my main results on the impact of stock returns on the real economy in the second stage. Trhought I use the fited value, denoted Δp^{2SLS}, of stock market return, given by

$$\Delta p^{2SLS} = 8.104 Z_t.$$

4.3 The real effects of stock market volatility

In this section I present the main results on how exogenous flow-induced price changes affect real economic activity. I look at how exogenous stock market price changes relate to contemporaneous and next-quarter GDP changes, aggregate consumption and investment, and some of their componenents. I use two-stage least squares to examine these relationships.

Table 4.1 presents summary statistics on the dependent variables: Δgdp is annualized quarterly percentage change in real GDP, seasonally adjusted; $\Delta Consumption$ is defined in the same way for aggregate consumption, which is then broken down into Durable and Non-Durable consumption. $\Delta Investment$ is annualized quarterly percentage change in aggregate investment, again decomposed into Residential and Non-Residential investment. The table reports mean, standard deviation, minimum, 25th percentile, median, 75th percentile, maximum and first autocorrelation for each time series. Date are from the first quarter of 1993 until the first quarter of 2019. All data are from th website of the St. Louis Fed.

In this period mean quarterly real GDP growth averaged 2.5% annualized, a remarkable performance for such an advanced economy as that of the United States, with a standard deviation of 2.3%. The worst quarter saw a fall of 8.5% annualized, and the 25th percentile was 1.5%. Mean and median growth were almost identical, the 75th percentile was 3.8% and the maximum 7.5%. GDP

growth is highly serially correlated at a quarterly frequency with a coefficient of 0.38, so I include four lags of my dependent variable in all my specifications.

Mean consumption growth was slightly higher than that of GDP at 2.8%, 6.10% for Durable consumption and 2.3% for Non-Durable. Consumption is less volatile than GDP growth at 1.9%, although both Durable (8.5%) and Non-Durable (2.4%) were more volatile. Although consumption growth is highly serially correlated, with a coefficient of 0.5, its components are not: Durable Consumption is not serially correlated and Non-Durable Consumption has a coefficient of only 0.17. Durable consumption has a median of 5.5%, higher than the mean, its lowest value was -26.3% and its highest 37.7%, so it seems the component of GDP most responsive to shocks.

Investment growth has both a higher mean (4.6%, median 4.3%) and volatility (11.9%) than either GDP or Consumption, with a lower minimum (-38.8%) and a higher maximum (42.4%), with correspondingly widely-spaced 25h (-2.2%) and 75th (11.5%) percentiles. The serial correlation of aggregate consumption is 0.29, making it slightly less persistent than GDP or Consumption, but its components are both highly persistent: Residential Investment has an autocorrelation of 0.55 and Non-Residential 0.65. Although quite comparable to each other, Residential Investment growth was slightly lower (mean 1.9%, median 4.5%) and more volatile (12.7%) than Non-Residential (mean 5.0%, median 5.9% and standard deviation 7.4%). The distribution of growth rates was also more dispersed than that of Non-Residential at all points of the distribution in the table.

Table 4.3 presents estimates of how GDP responds to stock market price changes. In column (2), the independent variable, labelled Δp_t, is the same-quarter instrumented percentage change in stock market price from my first stage. I include the following controls: 4 lags of the dependent variable, percentage change in the CPI and the 3 month T-bill rate. In column (1), I run the same regression, but using the naive OLS specification which uses actual stock market return instead of instrumented return. Standard errors are adjusted

for heteroskedasticity and serial correlation, and are reported in parenthesis beneath the relevant coefficient estimate.

Column (1) reports a statistically significant coefficient of 0.145: a 1% instrumented stock market return causes an annualized change in same-quarter real GDP growth of 0.145%. The standard deviation of the instrumented return is 4.5%, which corresponds to twice that amount on an annualized basis, 9.0%, just over half the standard deviation of the actual stock market excess return. Multiplying by 0.145 gives an annual average percentage change in GDP due to exogenous price changes of 1.3%, i.e. just over half an average year's US GDP growth. Since US GDP was approximately $20.5 trillion in 2018 the dollar amount implied is about $267 billion per annum, about the GDP of Finland in 2018. In other words, exogenous price changes driven by noninformational trades cause an amount of US GDP to be created or destroyed in a typical year equivalent to the size of the Finnish economy or a typical year's US GDP growth.

Column (1) shows that the naive estimate using endogenous price changes is not very different, at 0.118. If much of the volatility of stock market prices is in fact due to exogenous flows, as argued by Gabaix & Koijen (2021), then this result should not be surprising: To a first approximation, stock market price changes are at least partly exogenous, at least with respect to contemporaneous macroeconomic events. Recall that the R^2 from the first stage was 30.1% in quarterly data. The 'true' impact is slightly larger than the naive estimate, suggesting other factors that positively affect GDP growth are negatively correlated with stock market returns, notably monetary policy (see e.g. Cieslak & Vissing-Jorgensen (2021)).

Column (3) reports the 2SLS estimate for a regression on lagged exogenous market return. The coefficient is very similar: a statistically significant 0.115, in contrast to the OLS estimate which is much smaller at 0.042 and not significant. This result needs to be somewhat carefully interpreted, since there is persistence in my instrumental variable of about 0.35 quarterly. The lead-lag effects of

stock market returns on GDP become much clearer when investigating the impacts on aggregate consumption and investment, to which I now turn.

Table 4.4 is exactly the same as the previous table, but replaces GDP growth with equivalent measures for Consumption and Aggregate Investment growth. Panel A looks at Consumption. Column (1) reports the naive OLS estimate to be 0.075, and the 2SLS estimate to be about one-third higher at 0.103. Using a similar calculation as for GDP, this implies that exogenous stock market fluctuations drive fluctuations in aggregate consumption of about 0.9% or about one third of the mean growth rate on an annualized basis. Columns (3) and (4) report the effects for Durable Consumption, which is more discretionary and therefore more volatile. Here the naive estimate is about 0.45 while the 2SLS estimate is 0.68, implying pass-through of shocks to market prices to Durable Consumption of about two-thirds, an economically large effect amounting to 6.1% in a typical year, the same as average non-durable consumption growth. Vendors of durable consumption goods are significantly exposed to the stock market! For Non-Durables, the 2SLS estimate is smaller at 0.12 but still 50% higher than the naive estimate amounting to an annualized effect of 1.1% or about 47% of mean non-durable consumption growth, comparable to aggregate GDP. I find no evidence for effects at longer lead times for either type of consumption, the only significant effects are in the same quarter as the price shock.

Panel B reports results for aggregate investment. Here the effects for 2SLS and OLS are very different. The OLS coefficient is 0.39 versus a 2SLS coefficient of 0.71, but operating at a lead time of one quarter (I find no evidence of contemporaneous effects of stock market price shocks on aggregate investment). This is the largest effect I find in my results: A coefficient of 0.71 implies that the annualized impact of a one standard deviation stock market price shock is about 0.71 x 9% = 6.4% or about 1.4 times the mean of aggregate investment growth. To a first approximation then, stock market price shocks cause major fluctuations in aggregate investment, and may well be the major source of the

economic cycle, about 80% larger than implied by naive estimates which fail to account for endogeneity.

Decomposing investment into Residential and the rest, the effect is clearly much stronger through the Residential investment channel. The coefficient for Residential is 0.52 (OLS 0.38) versus 0.28 for Non-Residential (and only significant at the 10% level, versus 0.24 for OLS). These coefficients imply that one standard deviation shocks to the level of the stock market drive fluctuations in aggregate Residential Invesment of about 4.7% annually, or roughly 2.5 times mean Residential Investment growth, versus 'only' about 2.5% or about 50% of average growth for Non-Residential. Both channels are therefore hugely important absorbers of stock market price shocks, but Residential Investment responds with an enormous multiplier. Given I can plausibly claim these results are causal, net inflow to the stock market is the primary driver of the investment cycle in the United States, not the other way around.

In the rest of this section I investigate the components of Non-Residential Investment in more detail: I present results for my instrument where significant on, in particular, Equipment, Structures, and Intellectual Property. In particular I'm interested in exploring asymmetries in causal effects through price increases versus falls. Table 4.5 presents these results. In panel A the overall effect of stock market shocks on equipment purchases are not significant (unlike the OLS estimate in column (1)). However, this masks an important difference between upturns and downturns in the stock market: In column (5) I break my instrument into above- and below-average realizations in a reduced-form setting. The coefficient on my instrument is very large and significant, but is more than undone in downturns, with a coefficient on the interaction of my instrument and a below-average dummy of -15.861. This implies that one way in which firms respond to negative price shocks is through the cancellation or postponement of equipment investment. For Structures, with their longer lead times, I find no such effect: The effect of stock market shocks is present, but not more pronounced in downturns, perhaps because these projects are harder

to cancel or postpone. Finally, I find in Panel C, the effect on IP investment is very similar to that of Equipment, with a coefficient on the interaction term of -9.372, significant at 1%. These asymmetries can mask a benign overall picture of the impact of stock market shocks on certain kinds of investment. In stock market recessions, investment recessions appear to follow where logistically feasible, i.e. where the form of investment is flexible enough.

To summarise: Properly analyzed, stock market price shocks have large, first-order effects on GDP through consumption and investment, particularly through Durable Consumption and Residential Investment, and through more discretionary types of Non-Residential Investment in stock market falls, notably Equipment and IP. The implied economic magnitudes are large, consistent with idiosyncratic flows driving the investment, durable consumption and ultimately economic cycles.

4.4 Effects through the financing channel

In this section I explore further how corporate finance intermediates the impact of stock market price shocks on aggregate investment. I focus on seasoned equity offerings (SEOs), debt issuance and retirement (Debt), and Retained Earnings, the three traditional sources of corporate finance. My results in this section are of interest not only to macro-finance but also to corporate finance specialists who wish to understand the relative role of different types of finance as for example in the Myers & Majluf (1984) Pecking Order Theory. As predicted by that theory, I find that firms respond in aggregate to stock market shocks by reducing retained earnings and increasing seasoned equity offerings. I find no effect on debt.

Table 4.1 present summary statistics on my three financing variables of interest: ΔSEO_t, $\Delta Debt_t$, and $\Delta RetainEarn_t$. ΔSEO_t is the quarterly percentage growth rate of seasoned equity issuance in dollars by US listed firms, $\Delta Debt_t$ is correspondingly net debt issuance and $\Delta RetainEarn_t$ is net quar-

terly percentage growth in retianed earnings. Seasoned equity offerings growth averages 14% per quarter in nominal dollars, with a standard deviation of 58.4% (i.e. very volatile). Net new debt issuance averages 1.5% with a standard deviation of 1.2% (i.e. not volatile at all). Retained earnings dollar growth averages 2.4% per quarter with a standard deviation of 14%. The respective median growth rates are 4.4%, 1.5%, and 1.9%. The percentile distributions are very dispersed for SEO growth (a 25th percentile of -29.5% versus a 75th of 44.8%), not dispersed for debt (0.6% versus 2.2%) and somewhat for retained earnings growth (-5.8% versus 6.9%). These are consistent with SEOs being the major responder to stock market shocks, with retained earnings responding somewhat in the reverse direction, and debt mostly unresponsive. Serial correlations of each are respectively -0.42, 0.27 and -0.14.

Table 4.6 presents my results on how price shocks affect financing. Column (2) gives a statistically significant coefficient of 3.1 versus an OLS coefficient of 2.6. Similarly to the results on GDP, these effects are contemporaneous to the same quarter and somewhat, though not hugely, larger for the correct estimator. If most price changes are driven by exogenous flows, then in aggregate it appears firms treat them as such and respond accordingly. Again, the magnitudes are large: A one standard deviation shock to stock market prices from exogenous flows induces a 13.95% increase in seasoned equity fundraising in a typical quarter, almost exactly the same as the average growth rate. (To annualize, multiply by eight.) Presumably these are offerings by firms which have share issuance programs in place and are able to respond swiftly to price increases which their managers perceive to be exogenous, i.e. these are opportunistic share raises. In turn these equity raises are partially used to fund investment and partly, as discussed in the next paragraph, to retire retained earnings. These results are consistent with Baker & Wurgler (2000).

In contrast to equity, debt raises show no sensitivity to equity price shocks at all at the aggregate market level. (Note the contrast with Schmickler & Tremacoldi-Rossi (2022), who find that dividend flow-induced price increases

fund increased investment at the firm level through the debt channel.) Coefficients are basically zero. Retained earnings, on the other hand, respond in the opposite direction, with the 2SLS coefficient of -0.549 significant at the 10% level and much larger than the insignificant OLS estimate. This coefficient implies about a 2.5% fall in retained earning growth in response to a one standard deviation exogenous shock to the stock market, about minus one times the average growth rate. While much more muted (and statisically less significant) than the SEO response, this is consistent with a degree of substitution between retained earnings and new equity finance as sources of finance for investment, at least at the aggregate level.

In summary, firms at the aggregate level respond to stock market price shocks by ramping up equity issuance aggressively, show no response through a debt channel, and reduce retained earnings growth somewhat, consistent with some of the new equity being used to fund payouts to shareholders.

4.5 An alternative specification using dividend flows

I try an alternative test to check the robustness of my main result about GDP growth, based on the work of Hartzmark & Solomon (2022). Hartzmark & Solomon (2022) show that buying pressure from dividend payments, annonced weeks earlier, generates predictable positive returns on the value-weighted stock market index, just as Gabaix & Koijen (2021)'s GIV constructed from flows does. The difference is that dividend payments made weeks after they were announced are patently obviously exogenous with respect to real economic activity, whereas the flow-based GIV requires some justification.

I first replicate the result in table 2 of Hartzmark & Solomon (2022). I construct daily dividend payment yields, defined as the change in payments of dividends on the aggregate index D_t from the preus day D_{t-1} divided by the level of the index on day $t-1$. Summary statistics for this variable, dy_t, are

reported in table 4.1.

Table 4.7 replicates the results of Hartzmark & Solomon (2022) by regressing the daily return on the CRSP index, excluding dividends, on same day dividend yield dy_t in column (1). The coefficient is 0.524 with a standard error of 0.23, a statistically significant effect of entirely predictable payments on next-day returns. Column (2) instead uses previous day dividend yield and finds a slightly smaller coefficient of 0.493 with a standard error of 0.206. Note however that the R^2 of both regressions is extremely small, at 0.0004 and 0.0003 respectively. Thus although this instrument for price changes is easily seen as exogenous, it is a weak instrument. Harzmark et al. also show that just regressing returns on dummies for days with high dividend payments generates a very comparable result.

Because the instrument is weak, unlike flow-GIV, I cannot aggregate the fitted values of returns from daily regressions into lower-frequency instrument returns and generate the same results as in my previous table on GDP. When I do so the coefficient on GDP is positive, but not significant. Instead, motivated by these results, I sort all trading days in a quarter into days with above-average dividend payments and aggregate the returns on these days into the cumulative (excluding dividend) return on high dividend days, Δp_t^H, and the returns from the remaining low-dividend payment days, Δp_t. In table 4.8 I regress ΔGDP_t on these two sets of returns separately. Since the only difference between the returns adding up to Δp_t^H and those adding up to Δp_t^L are that the former are generated on higher-dividend payment days than the latter, any differential coefficient in this regression must be due to dividend payment-induced higher returns causing higher GDP growth. Because the number of trading days aggregated is the same for each variable, under the null of zero pass-through of price pressure to the real economy, the coefficents should be the same on each of Δp_t^H and Δp_t^L.

In table 4.8 column (1), I find the coefficient on Δp_t^H is a highly statistically significant 0.152, almost identical to the coefficient on instrument stock market

price changes using flow-GIV of 0.145. By contrast, the coefficient on Δp_t^L is 0.056 significant at the 10% level.

To test formally that the coefficient on Δp_t^H is higher than that on Δp_t^L, I rearrange my regression. Thus in column (1) I ran:

$$\Delta gdp_t = \alpha + \beta^H \Delta p_t^H + \beta^L \Delta p_t^L.$$

In column (2) I run:

$$\Delta gdp_t = \alpha + \delta^1(\Delta p_t^H + \Delta p_t^L) + \delta^2(\Delta p_t^H - \Delta p_t^L).$$

Under the null that $\beta^H \leq \beta^L$, $\delta^2 \leq 0$, whereas $\beta^H > \beta^L$ if and only if $\delta^2 > 0$. Column (2) reports an estimate of δ^2 of 0.048, significantly different from zero at the 10% level with a standard error of 0.028.

I also run both these regressions using Δp_{t-1}^H and Δp_{t-1}^L, that is lagged one quarter with respect to Δgdp_t. Although the coefficient estimates and their differences have similar signs to those in columns (1) and (2), they are not significant, rpboably because this instrument is much weaker than flow-GIV.

I conclude that a clearly exogenous though much weaker alternative approach to identifying effects of stock market changes on GDP growth gives very similar and statistically significant results to those of my main approach reported in table 4.3. My results are not dependent on my particular choice of instrument.

4.6 Conclusion

In this paper, I use Gabaix and Koijen's flow-based Granular Instrumental Variable to estimate the effect of stock market volatility on real economic activity. I present a motivation for a share-weighted average of idiosyncratic flows by investment sector into the aggregate stock market and confirm that it is indeed a strong instrument for stock market returns. I then show that instrumented stock market returns have a large impact on GDP growth, as well as aggregate

consumption and investment growth. The impact of returns is on all categories of consumption and investment, but most pronounced for Durable Consumption and Residential Investment, although the asymmetric impact of returns on Equipment and IP (both categories of Non-Residential Investment) is also remarkable. My results are consistent with flows into the stock market driving economic activity, as opposed to anticipated activity being the primary driver of stock market fluctuations.

I also investigate the financing channels through which these returns are intermediated by firms and the financial sector into investment. I find that firms show a strong increase in equity raising in response to stock price increases, as well as a reduction in retained earnings growth. These effects, like the effects on consumption and GDP, are largely contemporaneous, although the effect on investment is lagged one quarter.

To check whether my results are driven by my choice of instrument, I also test a second instrument, dividend payment flow, following the evidence of Harzmark and Solomon. Because this is a weak instrument, although unarguably exogenous to returns, I cannot do so using 2SLS. Instead I sort daily returns into high- and low- dividend payment days and test wheher high- dividend payment day returns (half the trading days in each quarter) have a greater impact on GDP growth than low dividend-payment days, and confirm that not only do high payment days have a significantly higher impact, but that the estimated impact is almost identical to that estimated using my main GIV instrument. I conclude that the dividend payment test is strong corroboration of my main result.

If stock market returns, driven by flows, are the primary driver of real economic activity, several policy implications follow. First, the Federal Reserve is correct to pay attention to the impact of its policies on the stock market. Second, it suggests that fiscal or monetary policy could be of first order importance in mitigating economic cycles, and could be financed largely or even entirely from long-run profits from policymakers standing ready to offset negative stock

market flows. Such policies would of course have secondary implications such as moral hazard, so careful consideration would be required.

In further research, I hope to investigate at a deeper level the effects I have found at more granular levels of the economy such as the industry and firm level. I could also consider alternative instruments other than dividend payment flows as a secondary test of my main results.

4.7 Tables

Table 4.1: Summary Statistics

This table presents summary statistics for constructed granular instrumental variables and quarterly price change (Panel A), macroeconomic variables (Panel B), corporate finance variables (Panel C), and variables for robustness check (Panel D). Z and Z^{pc} respectively are the granular instrumental variables before and after removing the first principal component of flows. Δp in Panel A is value-weighted CRSP index return excluding dividend (vwretx). All variables in Panel B are the annualised seasonally adjusted real growth rates. ΔSEO is the percentage change in aggregate seasoned equity offering in the US. $\Delta Debt$ is the percentage change in debt level of nonfinancial corporate business in the US. $\Delta RetainEarn$ is the percentage change in the retained earning of financial corporate business in the US. dy is the aggregate daily dividend yield on the dividend payment date, Δp in panel D is the daily vwretx on the dividend payment date. All variables in panel A, B and C are quarterly from 1993Q1 to 2019Q1. Variables in panel D are daily from 15 Jan 1926 to 30 Dec 2022.

	Mean	STD	Min	P25	Median	P75	Max	AR(1)
A: GIV variables (*N=105*)								
Z	-0.001	0.006	-0.021	-0.004	0.000	0.003	0.012	0.354
Z^{pc}	-0.001	0.005	-0.016	-0.004	0.000	0.003	0.010	0.371
Δp	0.014	0.082	-0.246	-0.017	0.025	0.057	0.197	0.009
B: Macroeconomic variables (*N=105*)								
Δgdp	0.025	0.023	-0.085	0.015	0.025	0.038	0.075	0.377
$\Delta Consumption$	0.028	0.019	-0.036	0.016	0.027	0.039	0.072	0.504
$\Delta Durable$	0.061	0.085	-0.263	0.025	0.055	0.105	0.377	-0.015
$\Delta nonDurable$	0.023	0.024	-0.057	0.011	0.026	0.037	0.084	0.166
$\Delta Investment$	0.046	0.119	-0.388	-0.022	0.043	0.115	0.424	0.293
$\Delta Residential$	0.019	0.127	-0.336	-0.030	0.045	0.100	0.265	0.548
$\Delta nonResidential$	0.050	0.074	-0.269	0.018	0.059	0.098	0.201	0.653
C: Corporate finance variables (*N=105*)								
ΔSEO	0.142	0.584	-0.811	-0.295	0.044	0.448	2.069	-0.422
$\Delta Debt$	0.015	0.012	-0.009	0.006	0.015	0.022	0.049	0.269
$\Delta RetainEarn$	0.024	0.140	-0.326	-0.058	0.019	0.080	0.551	-0.139
D: Robustness test (*N=22058*)								
dy	5.61E-05	1.05E-04	1.70E-08	6.64E-06	1.87E-05	6.26E-05	4.16E-03	6.51E-02
Δp	3.49E-04	1.12E-02	-8.99E-02	-4.31E-03	6.86E-04	5.46E-03	1.15E-01	-3.30E-02

Table 4.2: GIV and Stock Price

This table presents the results of OLS regressions where the dependent variable is the value-weighted CRSP index return excluding divididend (Δp_t). Z and Z^{pc} respectively are the granular instrumental variables before and after removing the first principal component of idiosyncratic flows PC_t.

	Δp_t		
	(1)	(2)	(3)
Z_t^{PC}	7.169***		
	(1.910)		
Z_t		8.104***	7.619***
		(1.774)	(1.316)
PC_t			2.682***
			(0.521)
N	105	105	105
R^2	0.216	0.301	0.412
F	14.089	20.873	40.276

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.3: GDP

This table presents the regression results where the dependent variable is real GDP growth rate Δgdp_t. Column (1) reports OLS regression on vwretx (Δp_t). Column (2) reports the second stage of the 2SLS regression, where $\Delta \hat{p}_t^{2SLS}$ is the fitted value of the first stage regression, instrumented by GIV Z_t. Column (3) reports OLS regression on lagged vwretx (Δp_{t-1}). Column (4) reports the second stage of the 2SLS regression, where $\Delta \hat{p}_{t-1}^{2SLS}$ is the fitted value of the first stage regression, instrumented by GIV Z_{t-1}. The controls are 1-4 lags of the dependent variables, percentage change in the CPI and the 3 month T-bill rate.

	Δgdp_t			
	(1)	(2)	(3)	(4)
Δp_t	0.118***			
	(0.036)			
$\Delta \hat{p}_t^{2SLS}$		0.145**		
		(0.059)		
Δp_{t-1}			0.042	
			(0.033)	
$\Delta \hat{p}_{t-1}^{2SLS}$				0.115**
				(0.058)
Control	Y	Y	Y	Y
N	101	101	101	101
R^2	0.384	0.375	0.237	0.186

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.5: Nonresidential Investment

	A: $\Delta Equipment_t$				
	(1)	(2)	(3)	(4)	(5)
Δp_{t-1}	0.358**				
	(0.144)				
$\Delta \hat{p}_{t-1}^{2SLS}$		0.326			
		(0.297)			
Q_{t-1}			0.013		
			(0.046)		
Z_{t-1}				2.955	14.556***
				(2.837)	(3.673)
$I_{\{Z_{t-1}>\bar{Z}\}}$					-0.082**
					(0.032)
$Z_{t-1} * I_{\{Z_{t-1}>\bar{Z}\}}$					-15.861***
					(5.197)
Control	Y	Y	Y	Y	Y
N	101	101	101	101	101
R^2	0.412	0.412	0.350	0.368	0.433

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Continued on next page

Table 4.5 – continued from previous page

	B: $\Delta Structures_t$				
	(1)	(2)	(3)	(4)	(5)
Δp_{t-1}	0.276**				
	(0.134)				
$\Delta \hat{p}_{t-1}^{2SLS}$		0.516**			
		(0.234)			
Q_{t-1}			0.074		
			(0.059)		
Z_{t-1}				4.942**	10.192**
				(2.202)	(4.026)
$I_{\{Z_{t-1}>\bar{Z}\}}$					-0.036
					(0.030)
$Z_{t-1} * I_{\{Z_{t-1}>\bar{Z}\}}$					-7.642
					(7.465)
Control	Y	Y	Y	Y	Y
N	101	101	101	101	101
R^2	0.255	0.233	0.237	0.232	0.276

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Continued on next page

Table 4.5 – continued from previous page

	C: $\Delta IntellectProperty_t$				
	(1)	(2)	(3)	(4)	(5)
Δp_{t-1}	0.123**				
	(0.058)				
$\Delta \hat{p}_{t-1}^{2SLS}$		0.178			
		(0.115)			
Q_{t-1}			0.033		
			(0.042)		
Z_{t-1}				1.560	7.581***
				(1.159)	(1.565)
$I_{\{Z_{t-1}>\bar{Z}\}}$					-0.037***
					(0.010)
$Z_{t-1} * I_{\{Z_{t-1}>\bar{Z}\}}$					-9.372***
					(2.839)
Control	Y	Y	Y	Y	Y
N	101	101	101	101	101
R^2	0.283	0.277	0.259	0.273	0.361

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.4: Consumption and Investment

This table presents the regression results where the dependent variable is the real Consumption growth rate (Panel A) and real Investment growth rate (Panel B). In Panel A, Column (1) and Column (2) respectively report the OLS and 2SLS regresssion on Δp_t, where the dependent variable is aggregate consumption growth $\Delta Consumption_t$. Column (3) and (4) respectively report the OLS and 2SLS regresssion on Δp_t, where the dependent variable is durable consumption growth $\Delta Durable_t$. Column (5) and (6) respectively report the OLS and 2SLS regresssion on Δp_t, where the dependent variable is nondurable consumption growth $\Delta nonDurable_t$. In Panel B, Column (1) and Column (2) respectively report the OLS and 2SLS regresssion on Δp_{t-1}, where the dependent variable is aggregate investment growth $\Delta Investment_t$. Column (3) and (4) respectively report the OLS and 2SLS regresssion on Δp_{t-1}, where the dependent variable is the residential investment growth $\Delta Residential_t$. Column (5) and (6) respectively report the OLS and 2SLS regresssion on Δp_{t-1}, where the dependent variable is nonresidential investment growth $\Delta nonResidential_t$. The controls are 1-4 lags of the dependent variables, percentage change in the CPI and the 3 month T-bill rate.

Panel A: Consumption						
	$\Delta Consumption_t$		$\Delta Durable_t$		$\Delta nonDurable_t$	
	(1)	(2)	(3)	(4)	(5)	(6)
Δp_t	0.075***		0.446***		0.082***	
	(0.019)		(0.115)		(0.029)	
$\Delta \hat{p}_t^{2SLS}$		0.103***		0.678***		0.121***
		(0.027)		(0.176)		(0.043)
Control	Y	Y	Y	Y	Y	Y
N	101	101	101	101	101	101
R^2	0.509	0.497	0.232	0.189	0.275	0.258
Panel B: Investment						
	$\Delta Investment_t$		$\Delta Residential_t$		$\Delta nonResidential_t$	
	(1)	(2)	(3)	(4)	(5)	(6)
Δp_{t-1}	0.393*		0.381**		0.243***	
	(0.235)		(0.164)		(0.080)	
$\Delta \hat{p}_{t-1}^{2SLS}$		0.713**		0.522**		0.283*
		(0.337)		(0.230)		(0.164)
Control	Y	Y	Y	Y	Y	Y
N	101	101	101	101	101	101
R^2	0.197	0.153	0.455	0.447	0.522	0.520

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.6: Financing Channels

The dependent variable is the percentage change in seasoned equity offering ΔSEO_t (Column (1) and Column (2)), the percentage change in debt level of nonfinancial corporate business $\Delta Debt_t$ (Column (3) and Column (4)), or the percentage change in retained earnings of nonfinancial corporate business ($\Delta RetainEarn_t$). The odd columns report the OLS results and even columns report the 2SLS results. The controls are real GDP growth, percentage change in the CPI and the 3 month T-bill rate.

	ΔSEO_t		$\Delta Debt_t$		$\Delta RetainEarn_t$	
	(1)	(2)	(3)	(4)	(5)	(6)
Δp_t	2.628***		0.012		-0.049	
	(0.649)		(0.015)		(0.182)	
$\Delta \hat{p}_t^{2SLS}$		3.132***		0.001		-0.549*
		(1.036)		(0.033)		(0.293)
Control	Y	Y	Y	Y	Y	Y
N	100	100	105	105	103	103
R^2	0.172	0.168	0.035	0.030	0.184	0.117

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.7: Dividend Payout and Change in Stock Price

The dependent variable is the value-weighted CRSP index return excluding dividend (Δp_t). dy_t is the dividend yield at time t and dy_{t-1} is the lagged dy_t.

	Δp_t	
	(1)	(2)
dy_t	0.524**	
	(0.227)	
dy_{t-1}		0.493**
		(0.206)
N	22058	22057
R^2	0.0004	0.0003
F	5.344	5.751

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.8: Dividend Payment-Driven Returns

This table presents the results of OLS regressions where the dependent variable is the real GDP growth rate Δgdp_t. Δp^H (Δp^L) is the cumulative returns of days in a quarter with above (below)-average dividend payments.

	Δgdp_t			
	(1)	(2)	(3)	(4)
Δp_t^H	0.152***			
	(0.048)			
Δp_t^L	0.056*			
	(0.029)			
$\Delta p_t^H + \Delta p_t^L$		0.104***		
		(0.030)		
$\Delta p_t^H - \Delta p_t^L$		0.048*		
		(0.026)		
Δp_{t-1}^H			0.042	
			(0.037)	
Δp_{t-1}^L			0.022	
			(0.052)	
$\Delta p_{t-1}^H + \Delta p_{t-1}^L$				0.032
				(0.032)
$\Delta p_{t-1}^H - \Delta p_{t-1}^L$				0.010
				(0.032)
Control	Y	Y	Y	Y
N	101	101	101	101
R^2	0.396	0.396	0.231	0.231

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Chapter 5

Conclusion

In my first thesis chapter, I develop a mean-variance model of capital market equilibrium with heterogeneous beliefs. The model shows that with either very low or very high disagreement, the standard CAPM holds because leverage constraint and short-sale constraint either do not bind or bind to a similarly large degree. When disagreement is moderate, leverage constraint is generally tighter than short-sales constraint. Consequently, less risky assets will be underpriced relative to the standard CAPM, while riskier assets will be overpriced. We confirm our theoretical predictions using the daily trading volume in the stock market as our measure of disagreement.

In the second chapter, I study an endowment economy with heterogeneous agents and two complementary consumption goods, one of which is indivisible, together with a discrete wealth distribution. The indivisibility gives rise to Friedman-Savage convexity in indirect utility. Agents care about relative, not just absolute, wealth, and about the magnitude of wealth differences as well as rankings. While agents dislike small risks, they like large ones. Poorer agents exhibit a preference for lottery-like assets with negative expected returns and also invest a smaller share of wealth than richer agents in assets with positive expected returns.

In the last chapter, I use Gabaix and Koijen (2021) to construct a Granular Instrumental Variable based on idiosyncratic flows from different investment sectors as an instrument for aggregate stock market returns. I use these instru-

mented returns to investigate the impact of stock market returns on the real economy. I find that stock market returns have a large impact on aggregate GDP, consumption and investment growth, primarily, but not only, through Durable Consumption and Residential Investment. Motivated by the results of Harzmark and Solomon (2021), I use high- and low- dividend payment day returns, aggregated to the quarterly level, as an alternative identification strategy and find an almost identical impact of stock returns on GDP growth.

Bibliography

Anderson, E. W., Ghysels, E. & Juergens, J. L. (2005), 'Do Heterogeneous Beliefs Matter for Asset Pricing ?', *The Review of Financial Studies* **18**(3), 875–924.

Atiase, R. K. (1991), 'Volume of Trading and the Dispersion in Financial Anaysts' Earnings Forecasts', *The Accounting Review* **66**(2), 389–401.

Atmaz, A. & Basak, S. (2018), 'Belief Dispersion in the Stock Market', *Journal of Finance* **73**(3), 1225–1279.

Baker, M., Bradley, B. & Wurgler, J. (2011), 'Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly', *Financial Analysts Journal* **67**(1), 40–54.

Baker, M. & Wurgler, J. (2000), 'The equity share in new issues and aggregate stock returns', *Journal of Finance* **55**(5), 2219–2257.

Banerjee, S. & Kremer, I. (2010), 'Disagreement and Learning : Dynamic Patterns of Trade', *Journal of Finance* **65**(4), 1269–1302.

Bansal, R. & Shaliastovich, I. (2010), 'Confidence risk and asset prices', *American Economic Review* **100**(2), 537–541.

Barinov, A. (2013), 'Analyst disagreement and aggregate volatility risk', *Journal of Financial and Quantitative Analysis* **48**(6), 1877–1900.

Basak, S., Makarov, D., Shapiro, A. & Subrahmanyam, M. G. (2018), 'Security design with status concerns', *Available at SSRN 2556344* .

Becker, G. S., Murphy, K. M. & Werning, I. (2005), 'The equilibrium distribution of income and the market for status', *Journal of Political Economy* **113**(2), 282–310.

Bessembinder, H., Chan, K. & Seguin, P. J. (1996), 'An empirical examination of information, differences of opinion, and trading activity', *Journal of Financial Economics* **40**(1), 105–134.

Black, F. (1972), 'Capital Market Equilibrium with Restricted Borrowing', *The Journal of Business* **45**(3), 444–455.

Campbell, J. Y. & Cochrane, J. H. (1999), 'By force of habit: A consumption-based explanation of aggregate stock market behavior', *Journal of Political Economy* **107**(2), 205–251.

Campbell, J. Y., Grossman, S. J. & Wang, J. (1993), 'Trading Volume and Serial Correlation in Stock Returns', *The Quarterly Journal of Economics* **108**(4), 905–939.
URL: *https://academic.oup.com/qje/article-lookup/doi/10.2307/2118454*

Chan, Y. L. & Kogan, L. (2002), 'Catching up with the joneses: Heterogeneous preferences and the dynamics of asset prices', *Journal of Political Economy* **110**(6), 1255–1285.

Chetty, R. & Szeidl, A. (2007), 'Consumption commitments and risk preferences', *The Quarterly Journal of Economics* **122**(2), 831–877.

Cieslak, A. & Vissing-Jorgensen, A. (2021), 'The economics of the fed put', *The Review of Financial Studies* **34**(9), 4045–4089.

Cole, H. L., Mailath, G. J. & Postlewaite, A. (1992), 'Social norms, savings behavior, and growth', *Journal of Political economy* **100**(6), 1092–1125.

David, A. (2008), 'Heterogeneous beliefs, speculation, and the equity premium', *Journal of Finance* **63**(1), 41–83.

DeMarzo, P. M., Kaniel, R. & Kremer, I. (2008), 'Relative wealth concerns and financial bubbles', *The Review of Financial Studies* **21**(1), 19–50.

Edwards, S. (2019), 'On latin american populism, and its echoes around the world', *Journal of Economic Perspectives* **33**(4), 76–99.

Eeckhoudt, L. & Kimball, M. (1992), Background risk, prudence, and the demand for insurance, *in* 'Contributions to insurance economics', Springer, pp. 239–254.

Fischer, S. & Merton, R. C. (1984), Macroeconomics and finance: The role of the stock market, Technical report, National Bureau of Economic Research.

Frazzini, A. & Pedersen, L. H. (2014), 'Betting against beta', *Journal of Financial Economics* **111**(1), 1–25.
URL: *http://dx.doi.org/10.1016/j.jfineco.2013.10.005*

Friedman, M. & Savage, L. J. (1948), 'The utility analysis of choices involving risk', *Journal of Political Economy* **56**(4), 279–304.

Gabaix, X. & Koijen, R. S. (2021), 'In search of the origins of financial fluctuations: The inelastic markets hypothesis'.

Garratt, R. & Marshall, J. M. (1994), 'Public finance of private goods: The case of college education', *Journal of Political Economy* **102**(3), 566–582.

Goetzmann, W. N. & Massa, M. (2005), 'Dispersion of opinion and stock returns', *Journal of Financial Markets* **8**(3), 325–350.

Grossman, S. J. & Stiglitz, J. E. (1980), 'On the impossibility of informationally efficient markets', *The American economic review* **70**(3), 393–408.

Hartzmark, S. M. & Solomon, D. H. (2022), Predictable price pressure, Technical report.

Hong, H. & Sraer, D. A. (2016), 'Speculative Betas', *Journal of Finance* **71**(5), 2095–2144.

Jarrow, R. (1980), 'Heterogeneous Expectations , Restrictions on Short Sales , and Equilibrium Asset Prices', *Journal of Finance* **35**(5), 1105–1113.

Keynes, J. M. (1936), *The General Theory of Employment, Interest and Money*, MacMillan.

Kimball, M. S. (1990), 'Precautionary saving in the small and in the large', *Econometrica* **58**(1), 53–73.

Lintner, J. (1969), 'The Aggregation of Investor's Diverse Judgments and Preferences in Purely Competitive Security Markets', *Journal of Financial and Quantitative Analysis* **4**(4), 347–400.

Lou, D. (2012), 'A flow-based explanation for return predictability', *The Review of Financial Studies* **25**(12), 3457–3489.

Morck, R., Shleifer, A. & Vishny, R. W. (1990), 'Do managerial objectives drive bad acquisitions?', *Journal of Finance* **45**(1), 31–48.

Myers, S. C. & Majluf, N. S. (1984), 'Corporate financing and investment decisions when firms have information that investors do not have', *Journal of Financial Economics* **13**(2), 187–221.

Ng, Y.-K. (1965), 'Why do people buy lottery tickets? choices involving risk and the indivisibility of expenditure', *Journal of Political Economy* **73**(5), 530–535.

Piazzesi, M., Schneider, M. & Tuzel, S. (2007), 'Housing, consumption and asset pricing', *Journal of Financial Economics* **83**(3), 531–569.

Piketty, T. (2014), *Capital in the twenty-first century*, Cambridge, MA: Harvard University Press.

Pollet, J. M. & Wilson, M. (2008), 'How does size affect mutual fund behavior?', *The Journal of Finance* **63**(6), 2941–2969.

Rabin, M. (2000), 'Risk aversion and expected-utility theory: a calibration theorem', *Econometrica* **68**(5), 1281–1292.

Rosen, S. (1997), 'Manufactured inequality', *Journal of Labor Economics* **15**(2), 189–196.

Roussanov, N. (2010), 'Diversification and its discontents: Idiosyncratic and entrepreneurial risk in the quest for social status', *The Journal of Finance* **65**(5), 1755–1788.

Saez, E. (2021), Public economics and inequality: Uncovering our social nature, *in* 'AEA Papers and Proceedings', Vol. 111, pp. 1–26.

Savor, P. & Wilson, M. (2014), 'Asset pricing: A tale of two days', *Journal of Financial Economics* **113**(2), 171–201.

Schmickler, S. & Tremacoldi-Rossi, P. (2022), 'Spillover effects of payouts on asset prices and real investment', *Available at SSRN* .

Sharpe, W. F. (1964), 'Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk', *The Journal of Finance* **19**(3), 425–442.
URL: *http://doi.wiley.com/10.1111/jofi.12597*

Simsek, A. (2013), 'Belief Disagreements and Collateral Constraints', *Econometrica* **81**(1), 1–53.
URL: *http://doi.wiley.com/10.3982/ECTA9956*

Sirri, E. R. & Tufano, P. (1998), 'Costly search and mutual fund flows', *The journal of finance* **53**(5), 1589–1622.

Staiger, D. & Stock, J. H. (1997), 'Instrumental variables regression with weak instruments', *Econometrica* **65**(3), 557–586.

Verrecchia, R. E. (1979), 'A Proof of the Existence of "Consensus Beliefs"', *Journal of Finance* **34**(4), 957–963.

Wang, J. (1994), 'A model of competitive stock trading volume', *Journal of political Economy* **102**(1), 127–168.

Wardlaw, M. (2020), 'Measuring mutual fund flow pressure as shock to stock returns', *The Journal of Finance* **75**(6), 3221–3243.