University of California
Los Angeles

Essays in Finance

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Management

by

Mahyar Kargar

2019

Abstract of the Dissertation

Essays in Finance

by

Mahyar Kargar

Doctor of Philosophy in Management

University of California, Los Angeles, 2019

Professor Andrea Lynn Eisfeldt, Co-chair

Professor Mikhail Chernov, Co-chair

In the first chapter of this dissertation, I study the asset pricing implications of heterogeneity in the financial intermediary sector. To examine the impact of massive balance sheet adjustments within the intermediary sector during the Great Recession, I propose a dynamic model with heterogeneous intermediaries and financial frictions. The model implies that a significant fraction of risk premia variation can be attributed to the composition of the intermediary sector. Asset reallocations, comparable in magnitude to the one observed during the recent financial crisis, lead to a substantial increase in the risk premia and volatility. An empirical measure of the composition of the financial sector strongly forecasts future excess returns with significant additional predictive power beyond many established forecasting variables in the literature. Moreover, shocks to wealth distribution among intermediaries have strong explanatory power for the cross-section of assets and are priced with a positive price of risk.

In the second chapter of this dissertation (with William Mann), we estimate small marginal costs and large markups at private colleges in the United States. For identification, we exploit a tightening of credit standards in the PLUS loan program, which decreased enrollment, revenues, and expenditures at private colleges with low-income students. We estimate that markups represented more than half of charges for students disqualified by the change.

Markups were higher at for-profit schools, and in states with fewer public schools and lower education spending. Our results contrast prior estimates of small markups in higher education. We conclude that the positive relationship between federal student aid and rising tuition costs, the so-called Bennett Hypothesis, arises due to imperfect competition between colleges.

The dissertation of Mahyar Kargar is approved.

Pierre-Olivier Weill

William Giles Mann

Tyler Stewart Muir

Mikhail Chernov, Committee Co-chair

Andrea Lynn Eisfeldt, Committee Co-chair

University of California, Los Angeles

2019

To Danielle,

I could not ask for a better partner and friend, than who you have proven yourself to be.

Thank you for your sacrifices;

to my parents Simin and Javad, and my sister Mahsa,

for laying the best foundation, and providing the most love and support, I could ask for;

and to Dora, Angie, and Frank,

for your unconditional support.

Table of Contents

1 Heterogeneous Intermediary Asset Pricing 1
1.1 Motivating Evidence 8
1.2 Model 11
1.2.1 Endowment and Agents 11
1.2.2 Financial Markets and Budget Constraints 13
1.2.3 Financial Constraints 14
1.2.4 Agents' Optimization Problems 15
1.2.5 Equilibrium 15
1.3 Model Solution 16
1.3.1 Endogenous State Variables 17
1.3.2 Hamilton-Jacobi-Bellman Equations 20
1.3.3 Recursive Markov Equilibrium 21
1.3.4 Numerical Solution 22
1.4 Calibration and Parameters Choices 22
1.5 Model Results 23
1.5.1 Constrained versus Unconstrained Economy 24
1.5.2 Cyclical Properties of Intermediary Leverage 30
1.5.3 Heterogeneous versus Representative Intermediaries 31
1.5.4 Implications of Financial Sector's Balance Sheet Adjustments 33
1.5.5 Empirical Predictions of the Model 34
1.5.6 Reconciling Empirical Evidence in AEM and HKM 35
1.6 Empirical Implications of the Model 36
1.6.1 Measuring Heterogeneity in the Intermediary Sector 37

1.6.2 Intermediary Heterogeneity and Time-Series Predictability 38
1.6.3 Intermediary Heterogeneity and the Cross-Section of Asset Returns . 42
1.6.4 Additional Robustness Checks 47
1.7 Conclusion 47
1.8 Appendix 1: Proof of Propositions 49
1.9 Appendix 2: Numerical Procedure 53
1.10 Appendix 3: Additional Model Results 59
1.10.1 Heterogeneous vs. Representative Intermediaries 59
1.10.2 Three-Dimensional Plots 60
1.11 Appendix 4: Data Sources 60
1.12 Appendix 5: Robustness Checks for Empirical Results 62
1.12.1 Predictive Regressions 62
1.12.2 Cross-Sectional Asset Pricing Tests 64
1.13 Appendix 6: Tables and Figures 70
1.14 Bibliography 91

2 Student Loans, Marginal Costs, and Markups: Estimates from the PLUS Program 97
2.1 Expenditures and costs in higher education 101
2.2 Demand shock: The PLUS standards tightening 104
2.2.1 The PLUS program 104
2.2.2 The 2011 tightening of PLUS credit standards 105
2.3 Implementation 106
2.3.1 Data 106
2.3.2 Cross-sectional treatment intensity 107
2.3.3 Characteristics of sample schools 108

2.4 Impact of the standards tightening 110
2.4.1 Graphical analysis 110
2.4.2 Regression analysis 112
2.5 Estimating markups charged to PLUS recipients 114
2.5.1 Instrumental-variables estimates of costs and markups 114
2.5.2 Heterogeneity in estimated markups 116
2.6 Conclusion 121
2.7 Appendix 1: Aggregating and merging IPEDS and Title IV 122
2.8 Appendix 2: Tables and Figures 124
2.9 Bibliography 136

List of Figures

1.1 Leverage and change in assets of different financial intermediaries 79

1.2 Risk premia, the price of risk, valuation, and volatility 80

1.3 Optimal portfolios and the risk-free rate 81

1.4 Dynamics of the endogenous state variables 82

1.5 Cyclical properties of intermediary leverage 83

1.6 Asset reallocation within the financial sector 84

1.7 State variable diffusions 85

1.8 Price of risk 85

1.9 State variables x and y in the data 86

1.10 Heterogeneous and representative intermediaries 87

1.11 Realized versus predicted mean returns: Heterogeneous intermediary factor . 88

1.12 Equilibrium in the unconstrained economy 89

1.13 Equilibrium in the economy with time-varying margin constraints 90

2.1 Aggregate revenues and expenses for private colleges 131

2.2 Average revenues and expenses per student for private colleges 131

2.3 Total borrowing through subsidized Stafford, unsubsidized Stafford, and PLUS programs since 2010 132

2.4 Average borrowing through Stafford and PLUS programs since 2010 132

2.5 Number of PLUS loan recipients for treated and untreated school systems . . 133

2.6 Enrollment at treated and untreated schools 133

2.7 Total tuition and fee revenue at treated and untreated private schools 134

2.8 Tuition and fee revenue per undergraduate student at treated and untreated private schools 134

2.9 Total expenditures at treated and untreated private schools 135

2.10 Total expenditures at treated and untreated private schools on only the three largest categories 135

List of Tables

1.1 Parameter values for the endowment economy model 70

1.2 State variables statistics 70

1.3 Predictive regressions: y_t. 71

1.4 Predictive regressions: y_t and Controls. 72

1.5 Cross-sectional asset pricing tests 73

1.6 Predictive regressions: Excluding the Great Recession 74

1.7 Predictive regressions for the market excess return: Robustness 75

1.8 One-way sorted CRSP portfolios on exposures to the heterogeneous intermediary factor 76

1.9 Two-way sorted CRSP portfolios 76

1.10 The heterogeneous intermediary mimicking portfolio (HIMP): Comparing models 77

1.11 The heterogeneous intermediary mimicking portfolio (HIMP): Comparing models with alternative projections 78

2.1 Summary statistics 124

2.2 Effect of the credit standards tightening on PLUS loan usage and enrollment 125

2.3 Effect of the credit standards tightening on college revenue, pricing, and spending 125

2.4 Effect of the credit standards tightening on PLUS loan usage and enrollment: Alternative approaches to measuring exposure to the standards-tightening treatment 126

2.5 Effect of the credit standards tightening on college revenue, pricing, and spending: Alternative approaches to measuring exposure to the standards-tightening treatment 127

2.6 Effects of enrollment on college expenses and revenue: IV regression 128

2.7 Cross-sectional heterogeneity in the estimated markups 129

2.8 Heterogeneity with *state* fixed effects 130

Acknowledgments

I am incredibly grateful to my advisors Andrea Eisfeldt and Mike Chernov, as well as my committee members Tyler Muir, Pierre-Olivier Weill, and especially William Mann, for invaluable support and guidance. I thank Alexi Savov for valuable discussions and feedback in the early stages of the project that led to the first chapter of this dissertation. For helpful comments and discussions, I thank Darren Aiello, Andy Atkeson, Tony Bernardo, Bruce Carlin, Michael Ewens, Alex Fabisiak, Mark Garmaise, Salil Gadgil, Chady Gemayel, François Geerolf, Valentin Haddad, Barney Hartman-Glaser, Lars Lochstoer, Francis Longstaff, Stavros Panageas, Rob Richmond, Bill Schiesser, Andrés Schneider, Avanidhar Subrahmanyam, Clinton Tepper, Edward Van Wesep, Ivo Welch, Geoffery Zheng, Irina Zviadadze, and seminar participants at University of Colorado Leeds, University of Illinois Gies, Indiana University Kelley, OSU Fisher, Rice University Jones, the LA Finance Day 2016, the UBC Summer Finance 2016 Early Ideas session, UCLA Anderson's Finance Brown Bag and Student Seminar series, and UCLA Economics.

Vita

1997-2001	B.S. Electrical Engineering, Sharif University of Technology
2002-2003	M.S. Electrical Engineering, University of Southern California
2004-2009	Ph.D. Electrical and Computer Engineering, University of California, Irvine
2005-2008	Senior Integrated Circuit Design Engineer, ClariPhy Communications
2008-2009	Staff Integrated Circuit Design Engineer, Quellan Inc.
2009-2013	Principal Scientist, Broadcom Corporation
2010-2013	M.B.A Finance, UCLA Anderson
2013	Harold M. Williams Fellow, top 2% GPA of M.B.A class of 2013, UCLA Anderson
2013	J. Fred Weston Award, Academic Excellence in Finance, UCLA Anderson
2013	Summer Intern, Research Affiliates
2013-2017	Anderson Fellowship, UCLA Anderson
2013-2017	Ph.D. Fellow, Laurence and Lori Fink Center for Finance and Investments
2014	Student Travel Award, American Finance Association
2015-2019	Teaching Assistant, UCLA Anderson
2016	Student Travel Award, Macro Finance Society
2017-2018	Dissertation Year Fellowship, UCLA Graduate Division
2019	Xavier Drèze Award for the Most Outstanding Ph.D. Research Paper, UCLA Anderson

CHAPTER 1

Heterogeneous Intermediary Asset Pricing

The financial sector witnessed a massive restructuring of its balance sheet during the Great Recession: Over the period from the first quarter of 2008 to the fourth quarter of 2009, which includes the most dramatic episode of the crisis in the fall of 2008, (i) broker-dealers drastically reduced asset holdings by approximately $1.7 trillion (a 35% drop) while commercial banks increased total asset by nearly $1 trillion (a 7.5% rise), and (ii) broker-dealers reduced leverage by about 47% while bank holding companies increased leverage by approximately 72%. This evidence is at odds with canonical intermediary asset pricing models which feature a *representative* financial sector. This work makes two main contributions. First, I extend the existing frameworks to explain these massive asset flows. Second, I apply my model to study its empirical asset pricing implications for time-series predictability and the cross-section of assets.

To explain these large asset flows and resolve the puzzling evidence presented above, I present a dynamic asset pricing model with two key features: (i) intermediaries *heterogeneous* in their risk-bearing capacity, and (ii) state-dependent margin constraints. I show that the *composition* of the intermediary sector, captured by one of model's two state variables, has important asset pricing implications beyond the health of the overall financial sector previously considered in the literature. I quantify the importance of this heterogeneity for the level and variation of the risk premium.

Guided by my model, I present two main empirical results that transcend the specific 2008 crisis episode. First, the wealth share of broker-dealers in the financial sector, a measure of the composition of the intermediary sector, strongly forecasts future market excess returns with additional predictive power beyond many popular forecasting variables in the literature.

Second, this measure of heterogeneity has strong explanatory power for the cross-section of asset: Shocks to the relative wealth share of broker-dealers in the financial sector, explains the cross-section of equity and bond returns about as well or better than existing intermediary asset pricing models.

As a corollary, my model reconciles seemingly contradictory asset pricing evidence from recent empirical evaluations of *representative* intermediary-based models. In particular, Adrian, Etula, and Muir (2014) (henceforth, AEM) and He, Kelly, and Manela (2017) (henceforth, HKM) find opposite signs for the price of intermediary leverage shocks in the cross-section of assets (positive and negative, respectively). Importantly, AEM and HKM measure intermediary leverage in *different* parts of the financial sector: security broker-dealers, and bank holding companies, respectively. The economic mechanism of the model, presented below, implies opposite leverage dynamics for different parts of the financial sector, resolving this puzzling evidence.

The model features two main ingredients. First, I assume agents differ in their attitudes toward risk: Two financial intermediaries (labeled A and B) and a household sector (C agents) in order of increasing risk aversion. I think of A and B intermediaries as broker-dealers and banks, respectively. This is consistent with the evidence that more aggressive hedge fund and broker-dealers have higher leverage on their balance sheets compared to more passive commercial banks. In equilibrium, less risk averse intermediaries hold levered positions in the risky asset financed by borrowing from the more risk averse agents. Second, investors face financial frictions in the form of occasionally binding state-dependent margin constraints. These constraints cap the maximum funding that intermediaries can obtain, which in turn impacts their asset demand.

Although all agents face margin constraints, in equilibrium, only more aggressive A type intermediaries face binding constraints. This calibration assumption has empirical support in the data: Hedge funds and broker-dealers primarily rely on collateralized repo financing with haircuts, while the commercial banking sector has access to more stable funding sources,

such as insured deposits and discount window lending from a central bank.[1]

The primary economic mechanism of the model is as follows: Following a negative shock more aggressive A type intermediaries face binding margin constraints and are forced to reduce leverage by selling assets. To clear the risky asset market, the less aggressive B intermediaries take on a larger portion of the asset than they would in the absence of constraints. The risk premium must increase to compensate them for bearing more risk. Since constraints are more likely to bind in high marginal utility states where intermediary wealth shares are low, consistent with empirical evidence, the model generates opposite cyclical dynamics for leverage of different intermediaries: Procyclical for more aggressive and countercyclical for less aggressive intermediaries.[2]

Heterogeneity in intermediaries' risk appetite and margin constraints are *both* necessary to match and understand asset reallocations within the financial sector that are consistent with observed patterns. In a model with *representative* intermediaries, only the aggregate wealth share of the financial sector matters for asset prices. Thus, in the absence of heterogeneity among intermediaries, models are unable to match asset flows within the financial sector. Similarly, in a *frictionless* setting, an adverse shock reduces intermediaries' risk-bearing capacity and results in a fall in prices and a direct increase in leverage for both regardless of their degree of risk tolerance. Therefore, without margin constraints, both intermediaries would lever up and down at the same time, and the model is inconsistent with empirical evidence on balance sheet adjustments in the financial sector.

My model features occasionally binding, time-varying margin constraints: The level of margin required in the model is state-dependent; and (inversely) linked to endogenously determined return volatility resembling an approximate Value-at-Risk (VaR) rule. Since

[1]In the second quarter of 2018, approximately 36% of total financial assets and 50% of total liabilities for security broker-dealers are due to lending and borrowing in the repo market, respectively. For private depository institutions, approximately 73% of total liabilities are comprised of checkable, time and savings deposits. Source: Financial Accounts of the United States (Flow of Funds).

[2]As mentioned above, one can think of aggressive and passive intermediaries as broker-dealers (BDs) and bank holding companies (BHCs), respectively, consistent with the observation that BDs are more likely to face binding borrowing constraints in bad times than BHCs.

there is empirical evidence that return volatility is higher in bad times (Schwert (1989), for example), such an approach is consistent with findings of Gorton and Metrick (2012) that haircuts tend to rise in crises.

I show that the model's equilibrium dynamics can be described by two endogenous state variables: (i) total wealth share of the financial sector (i.e. A and B agents) in the economy, and (ii) the more aggressive intermediary's (A types) wealth share as a fraction of the total financial sector. The former is the main state variable in many recent representative intermediary-based models.[3] The unique feature of the model is the second state variable: It emphasizes that the composition of the financial sector is a key factor in determining asset prices. Models with a representative financial sector are silent about the composition of intermediaries and its asset pricing implications.

Risk tolerant intermediaries (A and B agents) have levered balance sheets by borrowing from more risk averse households. Leverage increases intermediaries' exposure to aggregate shocks: Positive shocks result in their wealth share to increase. Following a negative aggregate shock, the wealth of levered intermediaries falls faster than that of households, and hence their share of total wealth, the first state variable in the model, declines. Moreover, since more risk tolerant intermediaries (i.e., A agents) have higher leverage than more risk averse ones, they are more likely to face binding margin constraints in bad times when they become tighter. As such, their wealth share in the financial sector, the second state variable of my model, declines as well. A key takeaway from model's economic mechanism is that both state variables exhibit *procyclical* dynamics: In high marginal utility states, both wealth share of the aggregate financial sector and more aggressive intermediary's net worth share in the financial sector are low.

I then examine two quantitative implications of my model. I first use the model to quantify the asset pricing implications of massive financial flows between intermediaries observed in the 2008 crisis. Explaining and studying the impact of these massive asset reallocations are one of the main goals of this study. As mentioned above, during 2008–

[3]See He and Krishnamurthy (2013), Brunnermeier and Sannikov (2014), Gertler and Kiyotaki (2010), and Drechsler, Savov, and Schnabl (2018) for a few recent models.

2009, broker-dealers drastically reduced asset holdings and leverage (by \$1.7 trillion and 47%, respectively), while banks increased both (by \$1 trillion and 72%, respectively). My model implies that a dealer deleveraging episode comparable in magnitude to the one observed during the crisis leads to an approximately 55% increase in the risk premia and a 5% increase in volatility. These balance sheet adjustments have no impact on asset prices in existing models with representative intermediaries.

Second, I show that the composition of the financial sector is responsible for a significant fraction of risk premia variation beyond the wealth share of the aggregate financial sector. With an independent and identically distributed (i.i.d.) aggregate endowment, variation in risk premia is only due to the aggregate wealth share as well as the composition of the financial sector, model's two state variables. I show that approximately 20% of the variation in risk premia can be attributed to the wealth distribution among intermediaries, which is a measure of the composition of the financial sector. Thus, failing to account for heterogeneity among intermediaries can lead to missing a substantial portion of risk premia variation. This result implies a novel empirical prediction of the model: The composition of the financial sector should strongly forecast future excess returns.

Next, I study the empirical implications of the model for time-series predictability and the cross-section of assets. I show that in addition to wealth share of the entire financial sector, keeping track of the composition of intermediaries is also crucial for determining risk premia. I define an empirical proxy for measuring this composition: The ratio of the equity of security broker-dealers to sum of equities of broker-dealers and holding companies from the Financial Accounts of the United States (Flow of Funds).

Consistent with my model's prediction, the composition of the financial sector strongly and *negatively* predicts future excess return. Model's second state variable, which captures this wealth distribution in the financial sector, exhibits procyclical dynamics: Times when broker-dealers are relatively more impaired in the financial sector coincide with high marginal utility states where prices are low and future expected returns are high. I show that the composition of the financial sector, captured by the wealth share of broker-dealers in the intermediary sector, is a strong predictor of future excess returns leading to additional

predictive power beyond many established return forecasting variables in the literature.

Moreover, *shocks* to the composition of intermediaries, which I denote the heterogeneous intermediary factor (HIFac), are priced in the cross-section of asset returns with a *positive* price of risk: The HIFac alone exhibits strong explanatory power for the cross-section of equity and bond returns about as well or better than existing intermediary asset pricing models in AEM and HKM. I further document that including aggregate intermediary leverage as a second asset pricing factor increases cross-sectional fit by at least ten percentages points, depending on whether the factor is leverage shocks for broker-dealers or bank holding companies.

In addition, I reconcile seemingly contradictory evidence in AEM and HKM by proposing a unifying general equilibrium framework with heterogeneous intermediaries. Pricing kernels in AEM and HKM measure marginal utilities of *different* financial intermediaries: broker-dealers and bank holding companies, respectively. Given the economic mechanism presented above, different parts of the financial sector exhibit opposite leverage dynamics. Therefore, it does not seem surprising that the literature with a representative financial sector arrives at conflicting asset pricing results.

I provide further evidence that heterogeneity in the financial sector is an important risk factor. Stock portfolios sorted on their exposure to HIFac (shocks to dealers' wealth share in the financial sector) exhibit monotonically increasing excess returns: the highest-beta quintile has approximately 5% higher annualized excess return relative to the lowest-beta portfolio. Existing representative intermediary asset pricing models are unable to capture these results.

Finally, I construct a mimicking portfolio for the heterogeneous intermediary factor from my model. Mimicking portfolios for representative intermediary factors in AEM and HKM are unable to fully span the heterogeneous intermediary factor-mimicking portfolio (FMP): I find large and highly significant alphas when I regress my model's FMP on AEM's and HKM's FMPs both individually and in bivariate regressions. This result corroborates my earlier findings: the composition of the financial sector is an important source of risk and

has pricing information beyond representative intermediary asset pricing factors.

My paper extends macroeconomic models with a financial sector (e.g., He and Krishnamurthy (2012) and He and Krishnamurthy (2013), Brunnermeier and Sannikov (2014), and Gertler and Kiyotaki (2010)) to a framework with heterogeneous intermediaries. This literature builds on financial accelerator models of Bernanke and Gertler (1989), Kiyotaki and Moore (1997), and Bernanke, Gertler, and Gilchrist (1999) which emphasize the importance of financial frictions and leverage for persistence and amplification of aggregate shocks.[4] The literature traditionally modeled intermediaries as one representative sector. Such an approach does not allow for the heterogeneity in financial intermediaries, documented in the data, to play a role in equilibrium.[5]

A few recent papers focus on the importance of a heterogeneous financial sector. The paper closest to this work is Ma (2017). In independent, contemporaneous work, he shows that an SDF estimated from a model with intermediaries heterogeneous in the tightness of their constraints exhibits higher cross-sectional R^2 than AEM and HKM factors. In contrast to this work, Ma (2017) focuses exclusively on explaining the cross-sectional variation of asset returns. Moreover, unlike his model where financing constraints are always binding, I study occasionally binding state-dependent leverage constraints. I also fully characterize the whole dynamic system instead of simply a log-linearized representation around the steady state in Ma (2017). Coimbra and Rey (2017) develop a model with intermediaries heterogeneous in their Value-at-Risk constraints and limited liability resulting in risk-shifting. Gertler, Kiyotaki, and Prestipino (2016) extend Gertler and Kiyotaki (2010)'s framework by incorporating a shadow banking sector alongside retail banks and allow the possibility of runs. The latter papers do not study the asset pricing implications of heterogeneous intermediaries. I contribute to this literature by showing that the composition of the financial

[4]See Brunnermeier, Eisenbach, and Sannikov (2012) for a survey of macro-based models with financial frictions.

[5]A few other recent models with explicit roles for financial intermediaries include Danielsson, Shin, and Zigrand (2012), Adrian and Shin (2014), Adrian and Boyarchenko (2015), Moreira and Savov (2017), and Drechsler, Savov, and Schnabl (2018).

sector has strong predictive power for excess returns of many assets and it is also priced in the cross-section of equity and bond returns.

This chapter also contributes to the recent empirical intermediary asset pricing literature. As noted above, two recent papers, Adrian, Etula, and Muir (2014) and He, Kelly, and Manela (2017), evaluate the explanatory power of models where *representative* intermediaries face, respectively, debt and equity constraints for cross-sectional variations in expected returns. They find opposite signs for the estimated price of risk (and thus conflicting cyclical dynamics) for intermediary leverage. I reconcile these seemingly contradictory evidence in a unifying general equilibrium framework where the financial sector is modeled as two sectors heterogeneous in risk-bearing capacity facing margin constraints.

Finally, this work relates to the extensive literature on asset pricing implications of investor heterogeneity and portfolio constraints. Dumas (1989), Wang (1996), Bhamra and Uppal (2009), Bhamra and Uppal (2009), Longstaff and Wang (2012), Gârleanu and Panageas (2015), and Santos and Veronesi (2016) study equilibrium in *frictionless* economies with two heterogeneous agents and different preference assumptions. Basak and Cuoco (1998), Kogan, Makarov, and Uppal (2007), Gârleanu and Pedersen (2011), Danielsson, Shin, and Zigrand (2012), He and Krishnamurthy (2012) and He and Krishnamurthy (2013), Chabakauri (2013), and Rytchkov (2014) examine asset pricing implications of exogenous and endogenous portfolio constraints in economies populated by two heterogeneous agents with one or many assets where constrained agent have logarithmic or CRRA preferences. I use this general framework to study the asset pricing implications of large flows within the intermediary sector. I present a general equilibrium model in an economy populated by three heterogeneous agents with recursive preferences who face state-dependent borrowing constraints.

1.1 Motivating Evidence

Before presenting the theoretical framework, in this section, I provide motivating evidence on heterogeneity of the intermediary sector. Empirical evidence from asset reallocations within the financial sector recorded during the Great Recession seem puzzling through the lens of

representative intermediary-based models.[6] Depending on the type of frictions considered, models with a representative financial sector can only describe intermediaries who are either buyers or sellers of assets during a crisis, but not both, implying opposite cyclical dynamics for intermediary leverage: *countercyclical* in models with equity constraints (He and Krishnamurthy (2012) and He and Krishnamurthy (2013) and Brunnermeier and Sannikov (2014)) or *procyclical* if intermediary faces a debt constraint (Brunnermeier and Pedersen (2009) and Adrian and Shin (2014)). Intermediaries exhibit heterogeneous behavior in the cyclical properties of their leverage. Figure 1.1a presents time-series of leverage for different financial intermediaries: security brokers and dealers (BDs), and bank holding companies of New York Fed's primary dealers (BHCs), intermediaries recently studied in AEM, and HKM, respectively. Broker-dealer's (book) leverage is calculated from balance sheet data in Table L.130 of the Financial Accounts of the United States (Flow of Funds) from the Federal Reserve and is defined as the ratio total financial assets to total equity (total financial assets minus total liabilities). BHC leverage is defined as the ratio of total market assets (book debt plus market equity) to total market equity constructed for publicly-traded holding companies of the New York Fed's primary dealer counterparties using data from CRSP/Compustat and Datastream. Over the period from the first quarter of 2008 to the fourth quarter of 2009, which includes the Lehman bankruptcy in the fall of 2008, broker-dealers *reduced* leverage by approximately 47% (from 35 to 19) while holding companies *increased* leverage by approximately 72% (from 22 to 38) during the same period.[7]

We observe opposite cyclical leverage patterns for different financial intermediaries: BD leverage is *procyclical*, while BHC leverage is *countercyclical*. Over the sample period from

[6]Evidence of intermediary heterogeneity during the crisis has also been recently documented in the literature. He, Khang, and Krishnamurthy (2010) and Begenau, Bigio, and Majerovitz (2017) document flows of financial assets within the intermediary sector during the Great Recession and show that broker-dealers and hedge funds reduced leverage by selling securitized assets to commercial banks who have access to stable deposits. Ang, Gorovyy, and Inwegen (2011) document that hedge funds decreased leverage prior to the onset of the financial crisis while the leverage of banks and the financial sector continued to increase. Ben-David, Franzoni, and Moussawi (2012) provide additional evidence of hedge fund deleveraging during the crisis.

[7]According to He, Khang, and Krishnamurthy (2010), book leverage of commercial banks rose from 10 to between 20 and 32 over the period from 2007Q4 to 2009Q1.

1970Q1 to 2017Q4, shocks to broker-dealer leverage exhibit a *positive* correlation of 0.12 (t-stat of 1.82) to innovations in the real GDP, while BHC leverage shocks have a *negative* correlation of -0.19 (t-stat of -2.62). Correlations with GDP innovations become stronger post 2000 with coefficients of 0.37 (t-stat of 3.27) and -0.39 (t-stat of -3.56) for broker-dealer and holding company leverage, respectively.[8]

In Figure 1.1b, I plot the quarterly change in total financial assets for security broker-dealers and private depository institutions (commercial banks) from Tables L.130 and L.110 of the Flow of Funds, respectively. Over the period from the first quarter of 2008 to the fourth quarter of 2009, broker-dealers, who mainly depend on collateralized repo financing, massively *reduced* asset holdings by approximately \$1.7 trillion (from \$4.9 to \$3.2 trillion), while commercial banks, who have access to more stable deposit financing, *increased* total asset by nearly \$1 trillion (from \$13.4 to \$14.4 trillion). This evidence is consistent with findings of He, Khang, and Krishnamurthy (2010) who document that during the 2008 crisis, hedge funds and broker-dealers reduce holdings of securitized assets by approximately \$800 billion and commercial banks increase holdings of these assets by approximately \$550 billion.[9]

Models with representative intermediaries are unable to capture this heterogeneity within the financial sector and study its implications for asset prices and the real economy. In the next section, I present a general equilibrium model with heterogeneous intermediaries and financial frictions that is consistent with opposite cyclical dynamics of leverage within the financial sector. My model implies that a dealer deleveraging episode comparable to one observed during the recent financial crisis, leads to an approximately 55% increase in the

[8]As discussed in HKM, observed opposite cyclical properties for leverage of different intermediaries above are unlikely to be entirely attributed to the differences between book- and market-based values for calculating BD and BHC leverage, respectively. To see this, I calculate holding company *book* leverage by simply replacing market equity with book equity in the calculation above. I find that book and market BHC leverage are in fact strongly *positively* correlated over the sample period (1970Q1 to 2017Q4) with correlation coefficient of 0.64 (t-stat of 11.64). Also, both market and book leverage for BHCs are strongly *negatively* correlated with book leverage of broker-dealers over the sample period with correlation coefficients of -0.50 (t-stat of -7.89) and -0.24 (t-stat of -3.47), respectively.

[9]From the fourth quarter of 2007 to the first quarter of 2009, the Federal Reserve and the GSEs increased holdings of securitized assets by approximately \$350 billion. See He, Khang, and Krishnamurthy (2010) for more details.

risk premia and a 5% increase in endogenous volatility. I then study model's asset pricing implications for time-series predictability and the cross-section of expected returns.

1.2 Model

In this section I present a general equilibrium model featuring heterogeneous intermediaries and financial constraints. My model reconciles seemingly contradictory results for the sign of price of intermediary leverage shocks from recent empirical evaluations of representative intermediary-based models. The model nests key forces behind Brunnermeier and Pedersen (2009) and He and Krishnamurthy (2012) with two main ingredients: (i) agents differ in their attitudes toward risk, and (ii) different intermediaries face state-dependent leverage constraints with varying degree of tightness, while equity issuance is ruled out by assumption.

I consider an endowment economy in continuous time populated by a continuum of agents whose total mass is one. There are three types of agents: A, B, and C with recursive preferences and different levels of risk aversion. To ensure the existence of a non-degenerate stationary wealth distribution, I assume each agent faces an exogenous constant mortality rate $\kappa > 0$. New agents are born at the same rate κ per unit of time with a fraction $\bar{u}$ as type A, a fraction $\bar{v}$ as type B, and a fraction $1 - \bar{u} - \bar{v}$ as type C. So, the total population is kept constant (normalized to one). In aggregate, the newborns inherit the wealth of their deceased parents on a per capita basis. Gârleanu and Panageas (2015) show that under these conditions, the possibility of exit makes agents more impatient: Their effective time preference is increased by κ.[10]

1.2.1 Endowment and Agents

The aggregate endowment D_t evolves according to

$$\frac{dD_t}{D_t} = \mu_D \, dt + \sigma_D \, dZ_t, \tag{1.1}$$

[10]The only purpose of introducing the OLG framework and mortality risk is to make the model stationary.

where μ_D and σ_D are constant parameters and Z_t is a standard Brownian motion defined on a fixed probability space $(\Omega, \mathcal{F}, P)$ and a filtration $\{\mathcal{F}_t, t \geq 0\}$ of sub-σ-algebras of $\mathcal{F}$ satisfying the usual conditions, as defined by Protter (2004).[11] The shock dZ_t is the only source of uncertainty in the model representing a permanent shock to the aggregate dividend. I assume that the growth rate of the endowment is positive, $\mu_D - \sigma_D^2/2 > 0$. Without loss of generality I set $D_0 = 1$. Similar to He and Krishnamurthy (2013) and Brunnermeier and Sannikov (2014), I assume agents are unable to hedge the aggregate risk.[12]

To separate the effects of elasticity of intertemporal substitution (EIS) and risk aversion, I assume all agents have stochastic differential utility as in Duffie and Epstein (1992), the continuous-time analog of recursive preferences of Epstein and Zin (1989). In particular, an agent of type i has the lifetime utility $U_{i,t}$ at time t given by

$$U_{i,t} = \mathbb{E}_t \left[\int_t^\infty f_i \left(C_{i,s}, U_{i,s} \right) ds \right], \tag{1.2}$$

where

$$f_i(C_{i,t}, U_{i,t}) = \left(\frac{1-\gamma_i}{1-1/\psi_i} \right) U_{i,t} \left[\left(\frac{C_{i,t}}{[(1-\gamma_i)U_{i,t}]^{1/(1-\gamma_i)}} \right)^{1-1/\psi_i} - (\rho + \kappa) \right]. \tag{1.3}$$

Function f_i aggregates over current consumption $C_{i,t}$ and future utility $U_{i,t}$. Parameters γ_i and ψ_i denote agent i's coefficient of relative risk aversion and EIS, respectively. These preferences reduce to standard power utility when $\psi_i = 1/\gamma_i$. All agents are assumed to have a common subjective discount factor ρ increased by κ as mentioned above.

Agents are heterogeneous in their attitudes toward risk γ_i. A agents are the most risk tolerant and C agents are the most risk averse. B agents are more risk tolerant than C types: $\gamma_A < \gamma_B < \gamma_C$. I think of A agents representing shadow banks (broker-dealers, hedge funds, etc.) and B agents as traditional banks, and C agents representing the household sector.

[11] The filtration represents the resolution over time of information commonly available to investors.

[12] Di Tella (2017) provides a moral hazard framework where aggregate uncertainty shocks lead to balance sheet recessions even though agents can write complete contracts on the aggregate state of the economy.

In equilibrium A and B will have levered balance sheets by borrowing from C agents. The financial sector (A and B agents) face time-varying margin constraints, which I discuss later in detail.

1.2.2 Financial Markets and Budget Constraints

All agents can trade a risky asset in fixed supply (normalized to one) and an instantaneous (from t to $t+dt$) risk-free bond in zero net supply which pays the endogenously-determined interest rate r_t. The risky asset is a claim on the aggregate endowment $\{D_t\}$, so, the total return on the risky claim is

$$dR_t = \frac{dP_t + D_t\,dt}{P_t} \equiv \mu_t\,dt + \sigma_t\,dZ_t, \tag{1.4}$$

where P_t is the price of the risky claim, μ_t is its expected return, and σ_t is its volatility, all determined in equilibrium. I use the consumption good as the numeraire. I also denote the dividend-price ratio of the risky asset by $F_t = D_t/P_t$.

Let $W_{i,t}$ denote agent i's wealth and assume $W_{i,0} > 0$ for $i \in \{A, B, C\}$.[13] Let w_s^i be the share of agent i's wealth invested in the endowment claim. Then agent i's financial wealth evolves according to the following standard dynamic budget constraint

$$\frac{dW_{i,t}}{W_{i,t}} = \left(r_t + w_{s,t}^i\left(\mu_t - r_t\right) - c_{i,t}\right)dt + w_{s,t}^i\sigma_t dZ_t, \tag{1.5}$$

where $c_i \equiv C_i/W_i$ is agent i's consumption-wealth ratio. The agent earns the risk-free rate, earns the risk premium on the risky asset, and pays for consumption. The intermediary leverage is defined as the ratio of asset over equity. Thus, when portfolios weights w_s^A or w_s^B exceeds one, the intermediaries operate with leverage by raising debt from households C.

[13]Throughout the paper, I use terms net worth and equity interchangeably.

1.2.3 Financial Constraints

I assume agent face a occasionally binding state-dependent margin constraint: At each moment in time, borrowers are restricted on how much leverage they can use on their balance sheets. In other words, lenders impose margin requirements to protect themselves against losses caused by adverse price movements.[14]Margins are set to shield lenders against adverse price movements and are widely used in the financial sector to fund levered balance sheets. They have also been previously studied in the asset pricing literature (see Brunnermeier and Pedersen (2009), Gârleanu and Pedersen (2011), Chabakauri (2013), and Rytchkov (2014) for some recent examples).

The tightness of the margin constraint can be determined by the regulators (e.g. Federal Reserve Regulation T) or by security broker-dealers (e.g. overcollaterlization of repos by a hedge fund's prime brokerage).

At any time t, I assume margin constraints restricts agent i's portfolio weight w_s^i to be below a certain state-dependent threshold $\bar{\theta}_t$

$$w_{s,t}^i \leq \bar{\theta}_t, \tag{1.6}$$

where, $\bar{\theta}_t$ determines the form of margin constraints, which is linked to endogenously-determined equilibrium objects (e.g. volatility of risk asset returns). Since equilibrium objects also depend on the state of the economy, margin requirements are state-dependent as well.

In particular, I assume margin requirements depend on the volatility of the risky asset return σ_t, and have the following functional form

$$\bar{\theta}_t = \bar{m} \left(\frac{1}{\bar{m}\alpha\sigma_t} \right)^{\nu}, \tag{1.7}$$

[14]In this chapter I abstract from the question of why the intermediaries face dynamic margin constraint and do not model the contracting problem among agents. Adrian and Shin (2014) provide a microfoundation for the Value-at-Risk constraint using a moral hazard problem in a static partial equilibrium setting.

where ν, α, and $\bar{m}$ are parameters that determine the type and tightness of the constraint, respectively. When $\nu = 0$, agents face a *constant* margin requirement: $\bar{\theta} = \bar{m}$. When $\nu = 1$, equation (1.7) resembles a Value-at-Risk rule.[15] In the latter case, the level of margin constraint is endogenous since it is inversely linked to the return volatility, which is an equilibrium object.[16]

1.2.4 Agents' Optimization Problems

Since agents are identical within each type and have homothetic preferences, I consider the problem faced by a representative agent i for $i \in \{A, B, C\}$. Each agent solves a standard Merton (1973) dynamic portfolio choice problem subject to margin constraints: agent i starts with initial wealth $W_{i,0} > 0$, decides how much to consume as a fraction of her wealth, $c_{i,t}$, and what fraction of her net worth to invest in risky asset, $w^i_{s,t}$, in order to maximize her value function in (1.2), subject to the dynamic budget constraint (1.5) and endogenous margin constraints (1.6). So, agent i's problem is

$$V_{i,t} = \max_{(c_i \geq 0, w^i_s)} U_{i,t}$$

$$\text{s.t.:} \quad \text{dynamic budget constraint (1.5) and margin constraint (1.6)} \tag{1.8}$$

and a solvency constraint $W_{i,t} \geq 0$.

1.2.5 Equilibrium

The definition of the competitive equilibrium is standard and is given below.

---

[15] Value-at-Risk constraints aim at limiting downside risk and maintaining an equity cushion large enough so that the default probability is kept below some benchmark level. They are common for banks and other leveraged financial institutions and are embedded in Basel II and Basel III regulatory frameworks. See Danielsson, Shin, and Zigrand (2012) and Adrian and Shin (2014) for recent examples.

[16] In equilibrium, more risk averse C agents lend to levered intermediaries. Thus, given the my calibration, the constraint can potentially bind only for A and B types. As mentioned above, in my calibration, γ_A and γ_B are chosen such that the constraint only (occasionally) binds for A intermediaries.

Definition 1. *A competitive equilibrium is the set of aggregate stochastic processes adapted to the filtration generated by Z_t: the price of claim on the aggregate endowment P, and the risk-free interest rate r; and a set of stochastic processes for each agent i: net worth W_i, consumption C_i, and stock holdings w_s^i; such that:*

i. Given the aggregate stochastic processes (P_t, r_t), choices $\left(C_{i,t}, w_{s,t}^i\right)$ solve agent i's optimization problem in (1.8).

ii. Markets clear

$$C_{A,t} + C_{B,t} + C_{C,t} = D_t \quad \textit{(goods market)} \tag{1.9}$$

$$w_{s,t}^A W_{A,t} + w_{s,t}^B W_{B,t} + w_{s,t}^C W_{C,t} = P_t \quad \textit{(stock market)} \tag{1.10}$$

The bond market clears by Walras' law. Note that bond market clearing implies that the aggregate wealth in the economy is equal to the value of the endowment claim, i.e.

$$W_{A,t} + W_{B,t} + W_{C,t} = P_t \,.$$

1.3 Model Solution

In order to solve the model, I need to determine how prices, portfolio choices, and consumption processes for all agents depend on the historical paths of the aggregate shock Z_t. The equilibrium can be characterized in a recursive formulation where all equilibrium objects are functions of two endogenous state variables, defined below. The computation of equilibrium requires solving the Hamilton-Jacobi-Bellman (HJB) partial differential equations of A, B, and C agents simultaneously. Unfortunately, the system of nonlinear PDEs does not admit a closed-form solution and I have to rely on numerical techniques. In this section, I first define my model's two endogenous state variables and derive their dynamics. I then characterize agents' value functions and provide some intuition for their optimal portfolio and consumption policy functions. I define a recursive Markov equilibrium and finally briefly discuss the numerical algorithm used to solve the PDEs.

1.3.1 Endogenous State Variables

Because Epstein-Zin preferences are homothetic, the optimal control variables for an agent are all linear in her wealth. The linear property allows me to simplify the endogenous state space, from an infinite-dimensional into a two-dimensional space. More precisely, I only need to keep track of the share of aggregate wealth that belongs to types A and B (the financial sector), as well as, the wealth share of A agents in the financial sector. I can derive equilibrium conditions as functions of the following endogenous state variables:

$$x_t \equiv \frac{W_{A,t} + W_{B,t}}{P_t}, \qquad y_t \equiv \frac{W_{A,t}}{W_{A,t} + W_{B,t}}. \tag{1.11}$$

Since the risk-free asset is in zero net supply, the aggregate wealth in the economy is equal to the risky asset price P_t. The state variable x is the share of aggregate wealth that belongs to the financial sector (i.e. A and B agents), and y is the type A intermediaries' wealth share as a fraction of the total financial sector.[17]

The state variable x (total wealth share of the financial sector in the economy), is the key state variable in recent intermediary asset pricing models with a representative financial sector (see He and Krishnamurthy (2013), Brunnermeier and Sannikov (2014), and Gertler and Kiyotaki (2010), for example). If only intermediaries can invest in the risky asset, state variable x represents the equity capital ratio of the financial sector.[18] HKM show shocks to capital ratio of intermediaries price the cross-section of expected return with a positive price of risk: intermediary's marginal value of wealth rises when capital ratio x falls.

State variable y, on the other hand, captures the wealth distribution within the intermediary sector. It represents heterogeneity among intermediaries in the sense that it would not be present in models with a representative financial sector. Distribution of wealth among different intermediaries clearly plays no role in the models with a representative financial sector.

[17]Note that the definitions in equation (1.11) ensure that the domain of both state variables is $[0, 1]$.

[18]In this case, because riskless bonds are in zero net supply and the risky asset is assumed to be in unit supply, total assets of the intermediary sector is equal to the risky asset price P.

In contrast, in Section 1.6, I show that the distribution of wealth between broker-dealers and bank holding companies (proxies for A and B agents, respectively) can negatively forecasts future returns for many asset classes. I also demonstrate that shocks to y are a priced risk factor in the cross-section of equity and bond returns with a positive estimated price of risk.

I restrict my attention to a Markov equilibrium (defined below) in the state space $(x, y) \in [0, 1] \times [0, 1]$, where all processes are functions of (x_t, y_t) only. Proposition 1, characterizes the dynamics of the two endogenous state variables (x, y).

Proposition 1. *The laws of motion for endogenous state variables x and y are given by*

$$dx_t = \kappa\left(\bar{x} - x_t\right) dt + x_t\left(\mu_{x,t}\, dt + \sigma_{x,t}\, dZ_t\right), \tag{1.12}$$

$$dy_t = \kappa\left(\bar{y} - y_t\right) dt + y_t(1 - y_t)\left(\mu_{y,t}\, dt + \sigma_{y,t}\, dZ_t\right) \tag{1.13}$$

where $\bar{x} = \bar{u} + \bar{v}$ and $\bar{y} = \bar{u}/(\bar{u} + \bar{v})$.

i. The drifts of x and y are given by

$$\mu_x = \left[y w_s^A + (1 - y) w_s^B - 1\right](\mu - r - \sigma^2) - y c_A - (1 - y) c_B + F \tag{1.14}$$

$$\mu_y = \left(w_s^A - w_s^B\right)(\mu - r) - c_A + c_B - \left[y w_s^A + (1 - y) w_s^B\right]\left(w_s^A - w_s^B\right)\sigma^2 \tag{1.15}$$

ii. The diffusions of x and y are given by

$$\sigma_x = \left[y w_s^A + (1 - y) w_s^B - 1\right]\sigma \tag{1.16}$$

$$\sigma_y = \left(w_s^A - w_s^B\right)\sigma \tag{1.17}$$

Proof. See Appendix 1.8. □

Given the dividend-price ratio F, the return process for the endowment claim in equation (1.4) can be rewritten as

$$dR = \frac{d(D/F)}{D/F} + F\, dt = \mu\, dt + \sigma\, dZ,$$

where time subscripts are dropped for notational simplification.

Using Ito's lemma, the expected return and volatility of the risky asset will be

$$\mu = \mu_D + F - \frac{F_x}{F}\left[\kappa(\bar{x} - x) + x(\mu_x + \sigma_D\sigma_x)\right] - \frac{F_y}{F}\left[\kappa(\bar{y} - y) + y(1-y)(\mu_y + \sigma_D\sigma_y)\right]$$
$$+ \left[\left(\frac{F_x}{F}\right)^2 - \frac{1}{2}\frac{F_{xx}}{F}\right] x^2\sigma_x^2 + \left[\left(\frac{F_y}{F}\right)^2 - \frac{1}{2}\frac{F_{yy}}{F}\right] y^2(1-y)^2\sigma_y^2$$
$$+ \left[2\left(\frac{F_x}{F}\right)\left(\frac{F_y}{F}\right) - \frac{F_{xy}}{F}\right] xy(1-y)\sigma_x\sigma_y \tag{1.18}$$

$$\sigma = \sigma_D - \frac{F_x}{F}x\sigma_x - \frac{F_y}{F}y(1-y)\sigma_y \tag{1.19}$$

Note that from (1.19), a part of the risk from holding the risky asset is fundamental, $\sigma_D\, dZ_t$, and a part is *endogenous*, $\left(-\frac{F_x}{F}x\,\sigma_x - \frac{F_y}{F}y(1-y)\,\sigma_y\right) dZ_t$. Equation (1.19) also implies that the volatility of returns σ exceeds the fundamental volatility σ_D when price-dividend ratio, $1/F$, and the state variables x and y are procyclical, i.e. $F_x > 0, F_y > 0, \sigma_x > 0$, and $\sigma_y > 0$, which is the case in equilibrium.

The following proposition provides the boundary conditions that the state variable diffusions satisfy.

Proposition 2. *The diffusion for state variables (x_t, y_t) satisfy the following boundary conditions:*

$$\lim_{x\to 0} x\,\sigma_{x,t} = \lim_{x\to 1} x\,\sigma_{x,t} = 0, \quad \forall y \in [0,1]$$
$$\lim_{y\to 0} y(1-y)\,\sigma_{y,t} = \lim_{y\to 1} y(1-y)\,\sigma_{y,t} = 0, \quad \forall x \in [0,1]$$

Proof. See Appendix 1.8. □

These boundary conditions will be used later to solve agents' HJB equations discussed below.

1.3.2 Hamilton-Jacobi-Bellman Equations

The recursive formulation of agent i's optimization problem is given by the following HJB equation

$$0 = \max_{c_i, w_s^i} f_i(c_i W_i, V_i(W_i, x, y))dt + \mathbb{E}_t\left[dV_i(W_i, x, y)\right], \tag{1.20}$$

where V_i is agent i's value function. With homothetic preferences, the value functions have the power form. The Following proposition characterizes agents' value functions.

Proposition 3. *The value function of agent $i \in \{A, B, C\}$ has the form*

$$V_i(W, x, y) = \frac{W_i^{1-\gamma_i}}{1-\gamma_i} J_i(x, y)^{\frac{1-\gamma_i}{1-\psi_i}}, \tag{1.21}$$

where J_i is agent i's consumption-wealth ratio, $c_i = J_i$.

Furthermore, J_i solves the following second-order partial differential equation (PDE)

$$\begin{aligned}
\rho + \kappa =& \frac{1}{\psi_i} J_i + \left(1 - \frac{1}{\psi_i}\right)\left[r + w_s^i(\mu - r) - \frac{\gamma_i}{2}\left(w_s^i\right)^2 \sigma^2\right] \\
&- \frac{1}{\psi_i}\left\{\frac{J_{i,x}}{J_i}\left[\kappa(\bar{x} - x) + x\mu_x\right] + \frac{J_{i,y}}{J_i}\left[\kappa(\bar{y} - y) + y(1-y)\mu_y\right]\right. \\
&\left. + (1-\gamma_i)\left(\frac{J_{i,x}}{J_i} x\sigma_x + \frac{J_{i,y}}{J_i} y(1-y)\sigma_y\right) w_s^i \sigma\right\} \\
&- \frac{1}{2\psi_i}\left[\left(\frac{\psi_i - \gamma_i}{1 - \psi_i}\right)\left(\frac{J_{i,x}}{J_i} x\sigma_x + \frac{J_{i,y}}{J_i} y(1-y)\mu_y\right)^2 + \frac{J_{i,xx}}{J_i} x^2 \sigma_x^2\right. \\
&\left. + 2\frac{J_{i,xy}}{J_i} xy(1-y)\sigma_x\sigma_y + \frac{J_{i,yy}}{J_i} y^2(1-y)^2\sigma_y^2\right]
\end{aligned} \tag{1.22}$$

Proof. See Appendix 1.8. □

Functions J_i capture agent i's investment opportunity set. In particular, note that from (1.21) if $\dfrac{1-\gamma_i}{1-\psi_i} > 0$ (which holds in my calibration), marginal utility of wealth is increasing in J_i.

The first-order conditions of agent's recursive problem gives the optimal consumption

and portfolio choice

$$c_i = \frac{C_i}{W_i} = J_i \tag{1.23}$$

$$w_s^{i,*} = \frac{\mu - r}{\gamma_i \, \sigma^2} + \frac{1}{\gamma_i} \left(\frac{1 - \gamma_i}{1 - \psi_i} \right) \left(\frac{J_{i,x}}{J_i} x \, \frac{\sigma_x}{\sigma} + \frac{J_{i,y}}{J_i} y(1-y) \frac{\sigma_y}{\sigma} \right) \tag{1.24}$$

The optimal unconstrained portfolio $w_s^{i,*}$ is the standard ICAPM result of Merton (1973): the first term in (1.24) is the myopic demand of a one-period mean-variance investor and the second term is the hedging demand capturing the variations in the agent's investment opportunity set. The optimal consumption-wealth ratio in (1.23) comes from the standard envelope condition.

So, from the optimal portfolio in the absence of constraints in (1.24) and the margin constraint in equation (1.6), the optimal portfolio is

$$w_{s,t}^i = \min \left(\bar{\theta}_t, w_{s,t}^{i,*} \right), \tag{1.25}$$

where the leverage upper bound $\bar{\theta}_t$ is defined in (1.7).

1.3.3 Recursive Markov Equilibrium

I derive a Markov equilibrium in state variables x_t and y_t. That is, I look for an equilibrium where all equilibrium objects (prices, consumption, and portfolio choices) can be written as functions of these two state variables. Next I define the Markov equilibrium in state space (x, y).

Definition 2. *A Markov equilibrium in state variables (x_t, y_t) is the set of functions: marginal value of wealth $J_i(x, y)$, dividend-price ratio $F(x, y)$, real interest rate $r(x, y)$ and policy functions $c_i(x, y), w_s^i(x, y)$ for $i \in \{A, B, C\}$, and laws of motion for endogenous state variables $\mu_x(x, y), \mu_y(x, y)$ and $\sigma_x(x, y), \sigma_y(x, y)$ such that*

i. marginal value of wealth J_i solves agent i's HJB equation, and c_i and w_s^i are corresponding policy functions, taking F, r and laws of motion for x and y as given.

ii. Markets for consumption good and risky asset clear

$$xy\, c_A + x(1-y)\, c_B + (1-x)\, c_C = F \quad (\textit{goods market}) \tag{1.26}$$

$$xy\, w_s^A + x(1-y)\, w_s^B + (1-x)\, w_s^C = 1 \quad (\textit{stock market}) \tag{1.27}$$

iii. The laws of motion for x and y satisfy (1.14)–(1.17).

1.3.4 Numerical Solution

The computation of equilibrium requires solving the HJB equations of the three types of agents simultaneously. Functions $J_A(x,y)$, $J_B(x,y)$, and $J_C(x,y)$ can be found by solving a system of second-order partial differential equations (PDEs) in (x,y). To do so, all equilibrium objects (e.g. $F, \sigma, \mu, \sigma_x, \mu_x, \sigma_y, \mu_y$, etc.) need to be expressed in terms of functions J_i and their derivative. Unfortunately, the system of nonlinear differential equations does not admit a closed-form solution and I have to rely on numerical techniques. This is particularly challenging in the presence of model's two endogenous state variables. I use projection methods, specifically orthogonal collocation using Chebyshev polynomials (Judd (1992) and Judd (1998)), to solve for equilibrium. Unlike a log-linearized representation around the steady state, this method provides a global solution and a full characterization of the whole dynamic system. In Appendix 1.9, I explain the numerical procedure in detail.

1.4 Calibration and Parameters Choices

Table 1.1 lists the parameter choices used in calibrating the model. I calibrate parameters to quarterly data.

I choose the drift and diffusion of the aggregate endowment (μ_D and σ_D) to match the mean and volatility of aggregate U.S. consumption data. To match the ratio of broker-dealer leverage (A intermediaries) to that of banks (B intermediaries), I set the risk aversions of A, B, and C agents to 2.5, 5.5, and 15, respectively.[19] I chose a common value of EIS for all

[19]From the Flow of Funds data, in my sample period (1970Q1-2017Q4), the average leverage of broker-

agent types and set $\psi_i = 1.5$.[20] The values of γ_i and ψ_i imply agent i's preference for the early resolution of uncertainty and have been extensively used in the asset pricing literature to address a number of asset pricing puzzles.[21] A value of EIS greater than one implies a decline in asset prices when the effective risk aversion in the economy increases. I set exit rate κ to 0.0154 which implies that agents on average live for 65 years which is consistent with the calibration in Gârleanu and Panageas (2015). The subjective discount rate ρ is set to 0.001 to achieve reasonable values for the real interest rates of between 2–3% annually.

The parameter $\bar{m}$ in equation (1.7) determines the constant margin requirements (when $\nu = 0$), and I set $\bar{m} = 4$. When $\nu = 1$, equation (1.7) resembles a Value-at-Risk constraint. In this case, parameter α determines the tightness of the constraint. I set $\alpha = 10$ which is approximately equal to the one-month Value-at-Risk at the 99% level.

1.5 Model Results

In this section, I first present additional properties of the equilibrium with margin constraints and compare them with the unconstrained economy with the same fundamentals and degree of heterogeneity among agents. The economy with margin constraints simultaneously exhibits higher risk premium and lower risk-free rate and volatility, compared to the frictionless benchmark. Although some of these effects have been previously documented in the literature, my analysis extends these results to an economy with three agents (households and two heterogeneous intermediaries) and recursive preferences.[22] The equilibrium with three

dealers is approximately 2.2 times higher than that of banks, roughly what I get in the model when state variables are at their unconditional means.

[20]Bansal and Yaron (2004) and Bansal, Kiku, and Yaron (2012) both use EIS value of 1.5 in their calibrations.

[21]See, for example, Bansal and Yaron (2004), Hansen, Heaton, and Li (2008), and Bansal and Shaliastovich (2013) for resolution of equity premium, value premium, and uncovered interest rate parity puzzles, respectively.

[22]For example, Rytchkov (2014) adds margin constraints to an endowment economy with two heterogeneous agents and CRRA preferences (similar to the models in Longstaff and Wang (2012) and Bhamra and Uppal (2014)) to show binding constraints reduce return volatility and risk-free rate, but increase expected

heterogeneous agents and two endogenous state variables is considerably more challenging to solve numerically.

In Section 1.5.2, I show, consistent with empirical evidence, the model can generate different cyclical dynamics for different intermediaries (i.e. A and B agent). Implications of the heterogeneity in the financial sector (as captured by the wealth distribution among intermediaries) for variation is risk premia beyond are discussed in Section 1.5.3. Section 1.5.4 studies the impact of a dealer deleveraging episode comparable to the one observed during the recent financial crisis, for risk premia and volatility. Finally, in Section 1.5.6, I show how the model reconciles seemingly contradictory evidence for the sign of price of intermediary leverage shocks in AEM and HKM.

1.5.1 Constrained versus Unconstrained Economy

Unconstrained Benchmark

As a benchmark, I first consider an economy without margin constraints, that is $\bar{\theta}(x_t, y_t) \to \infty$. In the absence of constraints, investors face complete markets and their Euler equations hold with equality in equilibrium. Without constraint, the economy is very similar to Gârleanu and Panageas (2015) but with *three* heterogeneous agents. It can be shown that without heterogeneity in risk tolerance, the interest rate, consumption-wealth ratio of each agent (J_i's), price-dividend ratio, and the price of risk are constant.[23]

Figures 1.2 presents various equilibrium variables (price-dividend ratio $1/F$, volatility of the risky asset return σ, Sharpe ratio $(\mu - r)/\sigma$, and risk premium on the endowment claim $\mu - r$) as a function of the state variable x while the state variable y is fixed at 0.56 (its stochastic steady state). In both figures solid blue lines correspond to the frictionless economy. Along the horizontal axis in each panel is the state variable x_t (the wealth share of types A and B), which ranges from 0 to 1.

returns and Sharpe ratio.

[23]These are very general results. For more details, see Gârleanu and Panageas (2015), Rytchkov (2014), and Longstaff and Wang (2012), for examples.

The top right panel of Figure 1.2 shows the volatility of returns. Even though fundamental volatility is constant ($\sigma_D = 3.5\%$ in my calibration), return volatility is time varying and it *exceeds* fundamental volatility in a hump-shaped pattern. As mentioned above and shown in equation (1.19), a part of the risk from holding the risky asset is fundamental and a part is endogenous. From equation (1.11), wealth shares of agents A, B, and C in the aggregate economy are equal to $x\,y, x(1-y)$ and $1-x$, respectively. Thus, when $(x, y) = (1, 1)$, $(x, y) = (1, 0)$, or $(x = 0, \forall y)$, the economy is populated by one type of agent (A, B, and C, respectively) and the volatility of the endowment claim coincides with the fundamental volatility σ_D. This can be validated from three-dimensional plots in Figure 1.13 in the Appendix. The Sharpe ratio and expected excess return $\mu - r$ both show countercyclical behavior as expected: higher risk premium and price of risk during distressed states when the intermediary sector is undercapitalized (low-x states) and/or broker-dealers deleverage (when y is low) are low.

The bottom right panel of Figure 1.2 shows that the risk premium largely tracks the Sharpe ratio. Note that this is the risk premium on a claim to the aggregate endowment, which has a relatively low volatility (3.5% in the baseline calibration). By comparison, equity volatility is around 16%. Therefore, the equity premium implied by the model is about five to six times larger than that of the endowment claim, putting it in the range of standard estimates in the literature.

When the EIS exceeds one, the substitution effect dominates the income effect so that greater risk aversion reduces asset demand and valuations fall. In this case the rise in the risk premium exceeds the fall in the real rate. In contrast, if the EIS is less than one, greater risk aversion counter-intuitively causes the valuations of risky assets to increase.

Figure 1.3 shows optimal portfolio weights in the risky asset for all three agents. For the most risk tolerant type A investors it always exceeds one and for the most risk averse type C agents it is always less than one. B's optimal portfolio is greater than one for most of the state space. Thus, without constraints, financial intermediaries (type A and B agents) borrow from type C investors (households) to take a levered positions in the endowment claim. As the wealth share of the financial sector get bigger and A agents have more relative

wealth in the financial sector, they borrow from traditional banks (B agents) as well.

Importantly, in the absence of constraints, the optimal leverage of A and B investors are both *countercyclical*: They are *higher* in bad states when investment opportunities are more attractive, which is counterfactual. This means when margin constraints are imposed to limit leverage, they are more likely to bind in high marginal utility states.

The relationship between portfolio weights and the wealth shares of the financial sector (x) and wealth share of broker-dealers in the financial sector (y) in Figure 1.3 is the result of market clearing for the risky asset in equation (1.27). When x is close to zero or both x and y are close to one, a single type of agent (type C in the first case and type A in the second case) dominates the economy, which reduces the opportunity for risk sharing. In the absence of constraints, agents of the dominant type must hold all their wealth in the risky asset, whereas agents of the vanishing type can be satisfied with only a small amount of borrowing or lending. Thus, when x is near zero, households (C agents) set prices and intermediaries (A and B types) take high leverage.

Equilibrium with Dynamic Margin Constraints

To study the impact of financial constraints on the equilibrium, I solve the model with the same fundamentals and heterogeneity as before, but now I assume investors face margin requirements in the form given in equation (1.7). As noted earlier, margin constraints are occasionally binding and state-dependent. Although all agents face the margin constraints in their optimization problems, in my calibration, the wedge between risk aversions γ_A and γ_B is such that constraints occasionally bind only for the most risk tolerant A types and more risk averse agents (i.e. types B and C) do *not* face binding leverage constraints, in equilibrium.[24]

Consistent with Kogan, Makarov, and Uppal (2007) and Rytchkov (2014), the economy with margin constraints simultaneously exhibits higher risk premium and Sharpe Ratio and

[24]The equilibrium will be qualitatively similar if we assume the least risk averse A intermediaries face tighter margin constraints than more risk averse B types.

lower risk-free rate and volatility, compared to the frictionless benchmark.[25]

I focus on two cases: (i) a constant margin constraint with $\bar{\theta}_t = \bar{m}$ ($\nu = 0$ in equation 1.7), and (ii) the case where type A agents face endogenous time-varying constraints ($\nu = 1$) in the form $\bar{\theta}_t = 1/(\alpha\sigma_t)$, where parameter α determines the tightness of the constraint. In the second case, margin requirements are determined by a Value-at-Risk-type rule and the level of margins is endogenous because it is (inversely) related to the return volatility, an equilibrium object.

Figures 1.2 and 1.3 present various objects for equilibria with constant (dash-dotted purple line) and state-dependent Value-at-Risk-type (dashed red line) margin constraints. There are few important observations from these figures. First, the impact of both types of constraints are qualitatively similar and the only difference is the magnitude. With the choice of parameters presented in Table 1.1, the time-varying margins are more restrictive and the effects are stronger with $\nu = 1$ in equation (1.7).

The top left panel of Figure 1.2 shows the impact of margin constraints on the price-dividend ratio. In my calibration, margin constraints do not substantially decrease asset's valuation ratio when they bind, reducing the price-dividend ratio by less than 1% at steady state relative to the complete-market benchmark.

The top right panel of Figure 1.2 plots the return volatility σ. The impact of constraints on volatility is unambiguous: portfolio constraints *reduce* the volatility of the risky asset return relative to the unconstrained economy. Also the volatility decreases in states where the constraint actually binds, although the point at which the constraint starts to bind depends on the form and severity of the constraint. The intuition behind this effect is as follows. In equilibrium, less risk averse A and B investors (the financial sector) borrow from more risk averse households and operate with leverage. It is well established in the macro-finance literature that leverage makes returns more volatile than the fundamental volatility: levered

[25]Both papers study a two-agent economy with CRRA preferences, whereas my model has three heterogeneous agents with recursive preferences and two endogenous state variables which is considerably more challenging to solve.

balance sheets amplify an aggregate shock to dividends.[26] Binding margin constraints reduce dynamic risk sharing and leverage in equilibrium, thereby reducing the return volatility. In my calibration, binding dynamic margin constraints results in the reduction of the return volatility by approximately 6% at model's stochastic steady state.

The bottom panels of Figure 1.2 demonstrate that portfolio constraints *increase* the Sharpe ratio and risk premium of the endowment claim. This is again a general effect and does not depend on the form of the constraints. The intuition is straightforward: because the leverage of the risk tolerant agent (A type) is bounded in the part of the state space where the margin constraints bind, to clear the market, the more risk averse investors (B and C types) are forced to take on a larger portion of the risky asset that they would without constraints. To induce them to buy more, the risk premium should increase. Following negative risky asset returns, margin constraints bind and broker-dealers (A types) are forced to sell assets. As a result, to clear the market, the expected returns must increase enough to entice more risk averse agents to take on a larger supply of the risky asset than before the shock. Since banks (B types) are not facing binding constraints, as discussed above, they increase leverage following a negative shock. Thus the model can qualitatively match the empirical evidence on opposite cyclical patterns of intermediary leverage in the financial sector documented in Figure 1.1a. At model's stochastic steady state, binding margin time-varying constraints causes the Sharpe ratio and risk premium to rise by approximately 39% and 36%, respectively.

Figure 1.3 also shows the effect of margin constraints on optimal portfolio weights. As explained above, because A's leverage is countercyclical in the unconstrained economy, the margin constraints will bind in states where x and y are low. In the model with margin constraints, type A agents operate with leverage in all states of the economy, however when margin constraints bind, leverage is restricted by $\bar{\theta}_t$. In other words, the presence of binding leverage constraints makes broker-dealers' leverage *countercyclical*: they are forced to sell assets and delever in bad states of the economy where constraints bind. To clear the risky

[26] See Kiyotaki and Moore (1997), Bernanke, Gertler, and Gilchrist (1999), and Brunnermeier and Sannikov (2014), for example.

asset market, both B and C investors need to absorb this additional supply and increase their portfolio weights as we see from dashed and dotted lines in the top right and bottom left panels of Figure 1.3.

Finally, the middle right panel of Figure 1.3 also shows that margin constraints (regardless of the form) *reduce* the risk-free rate. This result is also intuitive. In the absence of constraints, A and B investors operate with leverage by borrowing from type Cs. The upper bound for leverage for type As reduces the demand for credit, thus lowering the risk-free rate. In my calibration, when margin constraint binds, the risk-free rate decreases by approximately 8% in the stochastic steady state $(x_{ss}, y_{ss}) = (0.36, 0.56)$.

Figure 1.4 illustrates the evolution of model's two endogenous state variables x and y in the constrained and frictionless economies. The drift of $x_t(y_t)$ is positive for low levels and becomes negative for high values of $x_t(y_t)$. The points where the drift crosses zero is the *stochastic steady state* of the endogenous state variable, the point of attraction of the system in the absence of shocks. Importantly, from the right panels in Figure 1.4, notice that the diffusion terms σ_x and σ_y are always positive. This implies that following a negative aggregate shock, both state variables decline, i.e. x and y both exhibit *procyclical* dynamics in the model. In Section 1.6.1 below, I verify this also holds in the data. As shown in Proposition 2, at the boundary points of the state space ($x = 0, x = 1, y = 0$, and $y = 1$), the diffusion of state variables x and y are zero. This can be verified from the top- and bottom-right panels of Figure 1.4.

The left panels of Figure 1.4 also illustrate that portfolio constraints of both types reduce the volatility of the state variables. Because the impact of the constraints on the sensitivity of the price-dividend ratio to the state variables is very small (as shown in the top left panel of Figure 1.2), a decrease in σ_x and σ_y translates into a decrease in the return volatility σ as presented in the top-right panel of Figure 1.2 (this follows directly from equation (1.19)).

1.5.2 Cyclical Properties of Intermediary Leverage

Countercyclical Holding Company and Financial Sector leverage, *Procyclical* leverage for Broker-Dealers

In this section, I show that the model is able to generate leverage patterns for different intermediaries that are consistent with the empirical evidence presented earlier. As mentioned above and illustrated in Figure 1.3, in the *absence* of margin constraints, broker-dealers and bank holding companies both exhibit *countercyclical* leverage: their optimal leverage is higher in bad states. However, when broker-dealers face state-dependent margin constraints inversely dependent on return volatility, their leverage exhibit an (almost) opposite cyclical behavior. Since return volatility is hump-shaped (see the top right panel of Figure 1.2), margin constraints cause A type's leverage to be $\cup$-shaped when constraints bind. When the constraints are sufficiently tight, shadow bank leverage is *procyclical*, consistent with the empirical evidence from broker-dealers leverage presented in Figure 1.1a (the solid blue line) and also documented in AEM.

Leverage of the financial sector (types A and B in the model), w_s^{FS}, is defined as the share of sector's wealth held in the risky assets:

$$w_s^{FS} = \frac{w_s^A W_A + w_s^B W_B}{W_A + W_B} = y w_s^A + (1-y) w_s^B \tag{1.28}$$

where state variable $y = W_A/(W_A + W_B)$ is the wealth share of A types in the financial sector as defined in equation (1.11). We see that the financial sector leverage is the weighted average of A and B types' optimal leverage with a time-varying weight equal to the state variable $y \in [0, 1]$: the wealth share of broker-dealers (A types) in the financial sector.

The left panel of Figure 1.5 presents financial sector's leverage in the unconstrained equilibrium and in the model with endogenous margin constraints. As discussed above, margin constraints reduce financial sector's leverage when they bind: binding constraints reduce A type's leverage causing the return volatility to decrease relative to the frictionless benchmark.

The right panel of Figure 1.5 presents intermediary leverage in the equilibrium with a Value-at-Risk-type state-dependent margin constraint. In the model with margin constraints, financial sector and bank holding companies exhibit *countercyclical* leverage, while broker-dealers could have procyclical leverage when constraints bind. This is again consistent with the evidence presented in Figure 1.1a (dashed red line for holding company leverage and solid blue line for broker-dealer leverage) and also recently documented in the empirical intermediary asset pricing literature in AEM and HKM.

1.5.3 Heterogeneous versus Representative Intermediaries

Since the aggregate endowment in equation (1.1) is i.i.d. with constant volatility, variation in risk premia is only due to wealth distributions and intermediation frictions captured by state variables x and y. State variable x, representing the wealth share of the total financial sector in the economy, is the main determinant of time-varying risk premia in representative intermediary models. In my model with a heterogeneous financial sector, however, the wealth distribution among intermediaries (captured by state variable y) also contributes to the variation in risk premia. In this section, I answer the following question: What fraction of the variation in risk premia can be attributed to the state variable y, a measure of the composition of the financial sector?

To answer this question and investigate the role of heterogeneity in the financial sector, I compare the main results of the three-sector model with the ones from an economy with identical fundamentals but *two* heterogeneous agents instead: a household sector (C types identical to the main model) and a *representative* intermediary sector (I), where I agents face endogenous margin constraints as in the original model (equation 1.7 with $\nu = 1$).

Most of the parameters in the two-agent representative intermediary model are identical to the ones in the main model listed in Table 1.1: Household's risk aversion $\gamma_C = 15$, EIS for household and the representative intermediary $\psi_C = \psi_I = 1.2$, rate of time preference $\rho = .001$, growth rate and volatility of the aggregate endowment $\mu_D = .022$, $\sigma_D = .035$, and agents birth and death rates $\kappa = .0154$ (exactly as in the original three-sector model). I set

risk aversion of the representative intermediary sector to $\gamma_I = 3.3$: the wealth-weighted risk aversion of the financial sector (A and B investors) in the main model (with heterogeneous intermediaries) evaluated at the stochastic steady state for the two state variables.[27] Population share of the representative intermediary sector is set to $\bar{x} = .12$: sum of population shares of two intermediary sectors in the main model, $\bar{u}$ and $\bar{v}$ from Table 1.1. Note that with a representative intermediary, there is only *one* endogenous state variable x: the wealth share of the intermediary sector.[28]

I simulate representative- and heterogeneous-intermediary models for 3,000 quarters 20,000 times and examine the distribution of risk premium volatility.[29] As expected, the model with heterogeneous intermediaries exhibits more variation in risk premia than the one with a representative financial sector. Since the aggregate intermediary sectors in both models are (almost) identical, any excess variation in risk premia in the heterogeneous intermediary model has to be due to state variable y. In my calibration, approximately 20% of the variation in risk premia can be attributed to heterogeneity in the financial sector (state variable y).[30] Therefore, failing to account for heterogeneity among intermediaries can lead to missing a substantial portion of the variation in risk premia.

---

[27]The wealth-weighted risk aversion of the intermediary sector in main three-agent model is $\gamma_I = \left(\frac{y}{\gamma_A} + \frac{1-y}{\gamma_B}\right)^{-1}$. Using values for $\gamma_A = 2.5$ and $\gamma_B = 5.5$ from Table 1.1, at the stochastic steady state $y_{ss} = 0.56$, we get $\gamma_{I,ss} = 3.3$.

[28]The wealth share of the financial sector is the key state variable in existing models with a representative intermediary sector (see He and Krishnamurthy (2013), Brunnermeier and Sannikov (2014), Di Tella (2017), and Drechsler, Savov, and Schnabl (2018), for example.)

[29]Figure 1.10 in Appendix 1.10 shows these distributions.

[30]Notice that the horizontal axis in Figure 1.10 is the standard deviation of the risk premium for the endowment claim, which has relatively low volatility (3.5% in my calibration) relative to the market (approximately 16%). Therefore, equity premium volatility implied by the model is about five to six times larger than that of the endowment claim (approximately 1% for the heterogeneous intermediary model, for example).

1.5.4 Implications of Financial Sector's Balance Sheet Adjustments

As discussed in Section 1.1 and documented in Figures 1.1a and 1.1b, during the height of the financial crisis, broker-dealers substantially delevered (reduced leverage by approximately 47% relative the previous quarter), while during the same period, holding companies increased leverage by 72%. In this section, I measure the impact of this balance sheet adjustments within intermediaries on risk premia and endogenous risk.

When the least risk averse A types face binding margin constraints, they are forced to reduce leverage by selling assets. To clear the risky asset market, the more risk averse agents (B and C types) need to take on a larger portion of the asset than they would in the absence of constraints. In order to entice them to buy more, the risk premium must increase.

I perform the following exercise: I try to match the aforementioned increase and decrease in leverage of broker-dealers and holding companies, respectively, by tightening the margin constraint faced by A types in baseline calibration in Table 1.1. This is consistent with empirical evidence that the contraction in repo market financing during the recent financial crisis hit broker-dealers (represented by A agents in the model) particularly hard and forced them to deleverage (see Gorton and Metrick (2012) and He, Khang, and Krishnamurthy (2010), for example). As noted earlier in Section 1.5.1, tighter margin constraints leads to a decrease in total leverage and lower volatility. In order to achieve an increase in volatility consistent with the empirical evidence, I also make households relatively more risk averse than intermediaries.

Figure 1.6 presents results of this exercise. Tighter margin constraints is implemented by an increase in the parameter α. I also increases risk aversion of households relative to the intermediary sector (lower γ_I/γ_C) to obtain an increase in volatility. The top left panel shows that at model's stochastic steady state, increasing parameter α by 50% (consistent with the rise in repo-haircuts index during the 2008 crisis from Gorton and Metrick (2012)'s Fig. 4) and reducing γ_I/γ_C by 26% (from 0.61 to 0.45), results in approximately 48% decline in A types' leverage. This deleveraging is very close to what broker-dealers experienced during the crisis.

To clear the risky asset market, the risk premium must increase enough to entice more risk averse agents to take on a larger supply of the asset than before the shock. More risk averse holding companies increase leverage as a result of dealers' deleveraging (the top right and middle left panels): B type leverage rises by 79% and households' holding of the risky asset remain relatively unchanged (see the middle left panel of Figure 1.6). As the middle right and bottom left panels of Figure 1.6 show, this dealer deleveraging results in an approximately 55% increase in the risk premium and Sharpe ratio a 5% rise in endogenous volatility.

More importantly, asset reallocations between intermediaries do not impact the aggregate wealth share of the financial sector, and thus do no affect risk premia and volatility in a model featuring representative intermediaries. My model with a heterogeneous financial sector captures the implications of these balance sheet adjustments for equilibrium objects.

1.5.5 Empirical Predictions of the Model

In this section I present two main empirical predictions of the model which I test in the data: (1) The composition of the financial sector, captured in the state variable y, negatively predicts returns, and (2) shocks to this measure of composition is priced in the cross-section of assets with a *positive* price of risk

State variables load *Positively* on aggregate shocks. Figure 1.7 presents the diffusions of state variables x and y (σ_x and σ_y in equation 1.16 and 1.17) in the equilibrium with state-dependent margin constraint. From equation 1.16, σ_x always remains positive. Since, in my calibration, σ_y is also positive in the entire state space, both state variables are *procyclical*: following a negative dZ shock, both state variables go down.

The empirical prediction of the model is that state variables *negatively* predict future excess returns. In particular, times when the more aggressive intermediaries are relatively more impaired in the financial sector (i.e., when state variable y is low), coincide with high marginal utility states when the risk premia is high.

Price of Risk is *decreasing* in x and y. The left (right) panel of Figure 1.8 plots the Sharpe ratio of the risky asset as a function of state variable x (y) for three different values of the other state variable: Its unconditional mean, and its unconditional mean ± 3 standard deviations. We see that the Sharpe ratio is *decreasing* in both sate variables x and y. In Assets that pay

The empirical prediction of the model is that shocks to state variables are priced in the cross-section of expected returns with a *positive* price of risk. In particular, asset that pay well when the more aggressive intermediaries are relatively more impaired in the financial sector (i.e., when state variable y is low), are valuable hedges and demand *lower* excess returns.

1.5.6 Reconciling Empirical Evidence in AEM and HKM

In this section, I argue that my model featuring heterogeneous intermediaries and leverage constraints can reconcile the conflicting cross-sectional asset pricing evidence, recently documented in AEM and HKM.[31] As mentioned earlier, AEM and HKM find opposite signs for the price of intermediary leverage shocks in the cross-section of asset returns and thus conflicting cyclical dynamics of the intermediary leverage.[32]

Importantly, in reaching these seemingly contradictory results, AEM and HKM measure intermediary leverage in different parts of the financial sector: Broker-dealers and bank holding companies, respectively. AEM use shocks to book leverage of broker-dealers to construct an intermediary stochastic discount factor (SDF) and show that it prices equity

[31] He, Kelly, and Manela (2017) present a simple, one-period model in their appendix which was originally suggested by Alexi Savov in a conference discussion of the paper. The model can reconcile the contradiction between HKM's results and the ones documented in AEM. This static framework, however, is unable to capture implications of balance sheet adjustments within the financial sector for risk premium, the price of risk, and volatility discussed above.

[32] As mentioned above, the data for security brokers-dealers are from Table L.130 of the Financial Accounts of the United States (Flow of Funds). The underlying source for this data comes from FOCUS and FOGS quarterly reports filed with the SEC by these broker-dealers *in isolation* from other parts of their holding companies which are not publicly available. Data for publicly-traded holding companies of primary dealers are from CRSP/Compustat and Datastream. For a more detailed description of data sources, see Appendix 1.11.

and bond portfolios with a *positive* price of risk implying *procyclical* intermediary leverage. HKM, on the other hand, find that shocks to leverage of bank holding companies for the New York Fed's primary dealers price the cross-section of returns for many asset classes with a *negative* price of risk. In contrast to AEM, HKM's negative price of risk suggests that intermediary leverage is *countercyclical*.

A direct implication of the opposite flows and leverage dynamics within the financial sector in my model with heterogeneous intermediaries is that measuring the price of risk for shocks to leverage of *different* intermediaries will result in opposite signs. Thus, these seemingly contradictory asset pricing results can be reconciled in a model where the intermediary sector is modeled as two heterogeneous sectors facing financial constraints. Moreover, in Section 1.6, I will show that the composition of the financial sector plays an important role for time-series predictability and also has significant explanatory power for the cross-section of expected returns.[33]

1.6 Empirical Implications of the Model

In this section, I study empirical implications of the model discussed in Section 1.5.5 and show that the composition of the intermediary sector, measured by the wealth share of broker-dealers in the financial sector, matters for time-series return predictability and is also priced in the cross-section of expected returns. I focus on asset pricing implications of the model with time-varying margin constraints ($\nu = 1$ in equation 1.7). I show that consistent with model's predictions, an empirical measure of the state variable y, which measures the composition of the financial sector, has strong forecasting power for returns on various asset classes beyond factors already established in the literature to predict returns. Moreover, innovation in this wealth distribution prices the cross-section of equity and bond portfolios.

[33]It is important to note that the intermediary leverage is endogenous, and the fact that shocks to leverage are priced does not necessarily mean that intermediaries are the marginal investors. It could well mean that leverage is proxy for aggregate risk aversion. See Santos and Veronesi (2016) and Haddad and Muir (2018) for more details.

1.6.1 Measuring Heterogeneity in the Intermediary Sector

The price of risk in my model is time-varying and depends on wealth share of the financial sector, state variable x, as well as, broker-dealers' wealth share in the intermediary sector, state variable y in the model. Since financial sector's wealth share, x, is the key state variable in many existing models with a *representative* intermediary sector, in this section, I emphasize the importance of wealth distribution within the intermediary sector, captured by state variable y, for forecasting future returns.[34]

Since risk premium is decreasing in y (see Figure 1.8 in the Appendix), an asset that pays well when y is low is less risky. Thus, my model predicts that a higher wealth share for BDs in the financial sector forecasts higher prices and thus, *negatively* predicts future returns.

In the data, I compute wealth share of the financial sector as the ratio of their market equity of to the total market value of firms in the CRSP universe:

$$x_t^{data} = \frac{\text{Market capitalization of the financial sector}_t}{\text{Total market capitalization of the CRSP universe}_t}. \tag{1.29}$$

I use monthly equity data from CRSP to compute x^{data}. The financial sector is identified as firms in the CRSP universe for whom the first two digits of the header standard industry classification (SIC) code equals 60 through 67.[35]

Since I don't have access to market data for broker-dealers, model's second state variable, y (wealth share of broker-dealers in the financial sector), is computed as the ratio of BD's book equity to the sum of BHC and BD book equities from the Flow of Funds Tables L.130

[34] Adrian, Moench, and Shin (2014) study return predictability in representative intermediary models and show book leverage of broker-dealers negatively forecasts future equity and bond returns. He, Kelly, and Manela (2017) also run time-series predictive regressions and show the squared reciprocal of capital ratio for bank holding companies of NY Fed's primary dealers positively predicts future returns for many asset classes.

[35] This definition of the financial sector has been commonly used in the literature. See Giglio, Kelly, and Pruitt (2016) and Acharya, Pedersen, Philippon, and Richardson (2017), for example.

and L.131:

$$y_t^{data} = \frac{\text{Book equity of BDs}_t}{\text{Book equity of BHCs}_t + \text{Book equity of BDs}_t}, \tag{1.30}$$

where equity is computed by subtracting total liabilities (excluding miscellaneous liabilities) from total financial assets.[36] For a detailed description of the data, see Appendix 1.11.

Table 1.2 reports the mean, standard deviation, and autocorrelation of the state variables in the data, both in levels and innovations. The factors are autocorrelated in levels but not in changes.

Figure 1.9 shows the time-series of x^{data} and y^{data} using the CRSP and Flow of Funds data confirming the dramatic growth of the sector from 1980 to the onset of the recent financial crisis. Consistent with the model, x^{data} and y^{data} are both *procyclical*: innovations in x^{data} and y^{data} are both *positively* correlated with the innovations in the real GDP with correlation coefficients of 0.24 (t-stat of 2.05) and 0.21 (t-stat of 2.99), respectively.[37]

1.6.2 Intermediary Heterogeneity and Time-Series Predictability

The risk premium in my model is time-varying due to its association with wealth share of the financial sector, state variable x, as well as, broker-dealers' wealth share in the intermediary sector, state variable y in the model. As a result, expected returns are time-varying in the model and are predictable using lagged state variables as predictors. Since financial sector's wealth share, x, is the key state variable in existing models with a *representative* intermediary sector, in this section, I emphasize the importance of wealth distribution within

[36] In unreported results, using 49 Fama-French industry definitions, I identify publicly-traded broker-dealers as all US firms in the CRSP universe with standard industry classification (SIC) codes 6211 (Security brokers, dealers & flotation companies) or 6221 (commodity contracts brokers & dealers). I then equivalently define state variable y using *market* data as $y_t^{data,mkt} = \frac{\text{Market cap of dealers}_t}{\text{Market cap of the financial sector}_t}$. During the sample period (1971Q1-2010Q4), time series of $y^{data,mkt}$ is highly positively correlated with y^{data} computed from book values (in equation 1.30) with correlation coefficient of 0.55 (t-stat of 8.88). Prior to 2010, book and market series are even more strongly positively correlated (correlation coefficient of 0.68 with t-stat of 11.74). Post 2010, however, the two series become negatively correlated with coefficient of -0.5 (t-stat of -2.92).

[37] It is worth pointing out that the decline in y^{data} post 2009 is because two of the largest broker-dealers (Goldman Sachs and Morgan Stanley) became bank holding companies in 2009Q1. Two other broker dealers were also acquired by bank holding companies: J.P. Morgan purchased Bear Sterns and Merrill Lynch became part of Bank of America.

the intermediary sector, captured by state variable y, for forecasting future returns.[38]

Due to the presence of a single aggregate dividend shock, the two state variables are very positively correlated in the model. As a result, one might be concerned about multicollinearity when both x and y are included as forecasting variables in predictive regressions. However, the model with occasionally binding margin constraints exhibits highly nonlinear dynamics. In particular, running predictive regressions *unconditionally* using very long sample of simulated data in the model (I used 300-year long sample and 20,000 simulations), results in negative and significant coefficients on *both* state variables x and y.

As discussed above, state variable y is procyclical (see Figure 1.7 in Appendix 1.10.2): Times when the more aggressive intermediary is relatively more impaired in the financial sector, i.e. when y is low, coincide with high marginal utility states where the risk premia is high. As such, my model predicts that a higher wealth share for BDs in the financial sector forecasts higher prices (lower returns) thus, *negatively* predicting future returns. To test this hypothesis, I regress one-year-ahead holding period excess return (from quarter $t+1$ to $t+4$) for asset i on the lagged wealth share of BDs in the financial sector, y_{data} defined in equation (1.30), and controls:

$$R^i_{t+1\to t+4} - r^f_t = \gamma^i_0 + \gamma^i_y\, y_t + \gamma^i_{\text{Ctrl}}\, \text{Ctrl}_t + \varepsilon^i_{t+1\to t+4} \tag{1.31}$$

where $R^i - r^f$ is the average excess return on asset i, and Ctrl represents the vector of control variables that are known in the literature to forecast returns. I use the following control variables: wealth share of the aggregate financial sector (x from the model), fluctuations in the aggregate consumption-wealth ratio (*cay* variable) defined in Lettau and Ludvigson (2001), real price-dividend (PD) and cyclically adjusted price-earnings (CAPE) ratios from Robert Shiller's website, and variance risk premium (VRP) from Bollerslev, Tauchen, and

[38] Adrian, Moench, and Shin (2014) study return predictability in representative intermediary models and show book leverage of broker-dealers negatively forecasts future equity and bond returns. He, Kelly, and Manela (2017) also run time-series predictive regressions and show the squared reciprocal of capital ratio for bank holding companies of NY Fed's primary dealers positively predicts future returns for many asset classes.

Zhou (2009).

The model predicts a *negative* and significant coefficients γ_y^i: Times when wealth share of broker-dealers in the financial sector is high are associated with low marginal utility states where asset prices are high and future expected returns are low. As test assets, I use value- and equally-weighted CRSP portfolios, mean excess return of 25 size/book-to-market and 10 momentum portfolios from Ken French's data library, as well as, an equally-weighted portfolio of assets within each non-equity class studied in HKM available form Asaf Manela's website.[39] The sample is quarterly starts in 1970Q1 and ends in 2017Q4 for equity portfolios and in 2012Q4 for non-equity assets (limited by data availability).

Table 1.3 presents results of the univariate predictive regressions in (1.31), with state variable y as the only predictor, for different test assets mentioned above. In Column (1), where the dependent variable is the market excess return, I can directly verify model's prediction that state variable y should negatively predict future aggregate risk premia: An increase in the measure of the composition of the financial sector (state variable y) of 1 percentage point in deviation from its mean decreases the expected excess return by 1.16 percentage points (per quarter). Consistent with model predictions, we observe negative and significant $\hat{\gamma}_y$ for Market, size/book-to-market, momentum, sovereign bonds, and options portfolios. For most asset classes $\hat{\gamma}_y$ is negative, as expected: this measure of the composition of the financial sector *negatively* forecast future returns.[40]

To examine whether forecasting relationships are stable over time, and an investor could have profited from observing the predictor variable y, I follow Goyal and Welch (2008) and Campbell and Thompson (2008) to evaluate the out-of-sample performance of the predictive

[39]Non-equity assets in HKM are mostly from previous studies. See Appendix 1.11 and He, Kelly, and Manela (2017) for more details on test assets.

[40]The predictor variable y_t is very persistent with AR(1) coefficient around 0.96 in quarterly data. I verify (in unreported regressions) that the absolute value of the regression coefficient $\hat{\gamma}_y$ and the R^2 both rise with the forecast horizon (see Cochrane (2005)'s Chapter 20 for more details). The estimates in Tables 1.3-1.7 are corrected for the Stambaugh (1999) bias. Moreover, if I use growth rate of variable y (I used one- and five-year growth rates) as predictors, I still observe negative and significant coefficients γ_y.

regressions. I compute an out-of-sample R^2 statistic (R^2_{OOS}) as:

$$R^2_{\text{OOS}} = 1 - \frac{\sum_{t=1}^{T} (r_t - \hat{r}_t)^2}{\sum_{t=1}^{T} (r_t - \bar{r}_t)^2},$$

where $\hat{r}_t$ is the fitted value from a predictive regression estimated through period $t-1$, and $\bar{r}_t$ is the historical average return estimated through period $t-1$. The R^2_{OOS} for market excess return, Column (1) of Table 1.3, is approximately 4%. To assess the economic significance of return predictability, I use Campbell and Thompson (2008)'s simple metric: The increase in expected returns of a one-period mean-variance (MV) investor from observing the predictor variable y. A quarterly out-of-sample R^2 of 4% leads to an increase in expected returns of approximately 3% per year for a MV agents with a risk-aversion coefficient of 5.

Table 1.4 provides results of the predictive regression in (1.31) adding several control variables mentioned above to regressions in Table 1.3.[41] In Column (1), as a benchmark, I report the forecasting regression for the market risk premium using only aforementioned control variables as predictors. In Column (2), I add dealer wealth share in the financial sector, y, as an additional predictor. Comparing Columns (1) and (2), it is particularly important to point out that the composition of the financial sector, captured in wealth distribution y, leads an additional 15 percentage points predictive power for future market excess returns over variables already known in the literature to forecast returns: The R^2 of the predictive regression on the market excess return goes from 0.28 to 0.43 when y is included in the regression in Column (2) of Table 1.4.[42] It similarly reports negative and significant $\hat{\gamma}_y$ for Market, size/book-to-market, momentum, sovereign bonds, and options portfolios.

In Appendix 1.12, I provide additional robustness checks for the time-series predictability regressions above. I first show that the predictive regression results are robust to excluding

[41]The sample is now shorter and starts in 1990Q1 due to data availability for Bollerslev, Tauchen, and Zhou (2009)'s variance risk premium, one of the control variables used in return forecasting regression in equation (1.31).

[42]In one-quarter ahead predictive regressions, the incremental R^2 increases by 5 percentages points (from 0.22 to 0.27) when intermediary composition variable y is added to the regression.

the Great Recession (years 2007 to 2009) from the sample. So, it is not just the financial crisis that drives this predictability results. I also rerun the forecasting regression in Table 1.4 by adding AEM's broker-dealer leverage ratio, HKM's intermediary capital ratio for different asset classes. Adding these additional predictors, however, does not change the sign and significance of the coefficient $\hat{\gamma}_y$.

In summary, in this section, I provided strong empirical evidence that the composition of the financial sector, captured by state variable y, matters for prices, beyond the health of the aggregate financial sector: It has strong predictive power for excess return on many assets beyond variables from *representative* intermediary asset pricing models, as well as, the ones already known in the literature to predict return.

1.6.3 Intermediary Heterogeneity and the Cross-Section of Asset Returns

As mentioned above, the model with heterogeneous financial intermediaries can reconcile seemingly contradictory evidence for the sign of estimated price of risk for intermediary leverage shocks documented in AEM and HKM. In this section, I explore the implications of a heterogeneous financial sector for the cross-section of returns.

As shown in Figure 1.2 (and also in the bottom right panel of Figure 1.13 in Appendix 1.10.2), risk premium on the endowment claim is *decreasing* in both state variables x and y.[43] This suggests assets that pay poorly when: (i) the financial sector is less capitalized (i.e. when x is low), and/or (ii) wealth share of broker-dealers in the financial sector is small (i.e. when y is low), are riskier and should command higher expected returns. I emphasize that point (ii) can only be made in a model with heterogeneous financial intermediaries.

1.6.3.1 Cross-sectional Asset Pricing Tests

Similar to He, Kelly, and Manela (2017), I construct the growth rate to dealers' wealth share in the financial sector, denoted y_t^{Δ}, as follows. I estimate a shock to dealer wealth share in

[43] This is true even in the absence of margin constraints as shown in the bottom right panel of Figure 1.12.

levels, ς_t, as an AR(1) innovation in the regression: $y_t = \phi_0 + \phi y_{t-1} + \eta_t$. I then convert these innovations to a growth rate by dividing them by the lagged wealth share:

$$\text{HIFac} = y_t^{\Delta} = \frac{\eta_t}{y_{t-1}} \tag{1.32}$$

I call this wealth share growth rate the *heterogeneous-intermediary factor* (HIFac) and use it to perform cross-sectional asset pricing tests. For each asset i, I first estimate betas from time-series regressions of portfolio excess returns on the risk factors:

$$R_{i,t}^e = a_i + \beta_{i,f}' \mathbf{f}_t + \vartheta_{i,t}, \quad i = 1, \ldots, N, \tag{1.33}$$

where $\mathbf{f}$ represents the $K \times 1$ vector of risk factors. I consider four cases: (1) $\mathbf{f}_t = \text{HIFac}_t$, (2) $\mathbf{f}_t = [\text{HIFac}_t \quad \text{MktRF}_t]'$, (3) $\mathbf{f}_t = [\text{HIFac}_t \quad \text{AEM}_t]'$, and (4) $\mathbf{f}_t = [\text{HIFac}_t \quad \text{HKM}_t \quad \text{MktRF}_t]'$, where AEM is the broker-dealer leverage factor from Adrian, Etula, and Muir (2014), HKM is the intermediary capital risk factor from He, Kelly, and Manela (2017), and MktRF represents the market risk premium. For comparison, I also report the pricing performance of AEM and HKM factors.

Next, in order to estimate factor risk prices, λ_f, I run a cross-sectional regression of average excess returns on the estimated risk exposures $\hat{\beta}_{i,f}$:

$$\mathbb{E}\left[R_{i,t}^e\right] = \alpha_i + \hat{\beta}_{i,f}' \lambda_f + \zeta_i, \quad i = 1, \ldots, N, \tag{1.34}$$

As mentioned above, the model predicts a *positive* and significant sign for the estimated price of risk λ_{HiFac}.

I test the ability of the heterogeneous intermediary factor in pricing the cross-section of 55 equity and bond portfolios: The test assets are 25 size and book-to-market and 10 momentum portfolios from Ken French's website, 10 maturity-sorted US government bond portfolios from CRSP's Fama Bond dataset with maturities up to five years in six month intervals, and 10 US corporate bond portfolios sorted on yield spreads from Nozawa (2017) obtained from Asaf Manela's website. I choose equity and bond portfolios as test assets due

to the availability of longer time-series than others such as options and CDS. Since I use many test assets beyond size and book-to-market portfolios, my model avoids the typical criticisms of asset pricing tests discussed in Lewellen, Nagel, and Shanken (2010).

Table 1.5 presents the main asset pricing results. Below estimated risk prices I report Shanken (1992) t-statistics that corrects for estimation error in betas and cross-correlations. I also report Fama and MacBeth (1973) t-statistics by running period-by-period cross-sectional regressions and computing standard errors of the time-series average of λs. I report cross-sectional R^2 and the mean absolute pricing error (MAPE), calculated as $\frac{1}{N}\sum|\zeta|$ where N is the number of test assets, as measures of model fit. I also report a $\chi^2(N-K)$ statistic (K is te number of factors) that tests if the pricing errors are jointly zero.

Column 1 of Table 1.5 reports the results of heterogeneous intermediary factor as a single pricing factor. The estimated price of risk is positive, which means assets that pay well in states with a low broker-dealer wealth share in the financial sector (i.e. assets with low betas on y_t) are valuable hedges and have lower expected returns in equilibrium. This risk price estimate confirms the procyclicality of broker-dealer wealth share y_t documented in Figure 1.9. The adjusted R^2 is 61% while the total MAPE is only 1.86%. The single-factor model can explains 62% of the variation in average returns in these cross-sections, with an average absolute pricing error around 1.8% per annum. Figure 1.11 in the Appendix, visually shows the HIFac's pricing performance in the cross-section of equity and bond returns by plotting realized versus predicted returns.

For robustness and comparison with recent empirical work, in Columns 2–6, I add additional pricing factors. In Column 2, I include market risk premium, MktRF, as an additional factor. However, the price of risk for MktRF is not statistically significant and in terms of almost all test statistics, the two-factor model is nearly identical to the single-factor model in Column 1. The market adds essentially no explanatory power to the intermediary heterogeneity factor.

In Columns 3 and 5, for reference, I present performances of the pricing factors in AEM and HKM, respectively. AEM use a leverage factor defined as the seasonally adjusted growth

rate in broker–dealer book leverage level from Flow of Funds. A shown in Column 3, for the test assets mentioned above, HIFac outperforms AEM with 38% lower MAPE (1.83% versus AEM's 2.96%) and 56% higher cross-sectional R^2 (61% compared to 39% in AEM).[44]

In HKM, the pricing factors are the market risk premium (MktRF) and shocks to intermediary capital ratio defined as the ratio of total market equity to total market assets (book debt plus market equity) for New York Fed's primary dealer holding companies. A shown in Column 5, my model with a single pricing factor performs almost as well as HKM's two factor model with nearly identical MAPE (1.83% vs. 1.89% for HKM) and cross-sectional R^2 (61% vs. HKM's 63%).

In Column 4, I add leverage factor from AEM to evaluate a model with two pricing factors: HIFac and AEM. Addition of AEM's leverage factor does not make price of HIFac risk insignificant or change its sign. This even raises the cross-sectional R^2 to 72%. Finally, in Column 6, I add two pricing factors from HKM: MktRF and shocks to intermediary capital ratio. Again, λ_{HIFac} remains positive and significant. Note that since HIFac is positively corrected with both HKM and AEM factors, it is not surprising that λ_{HIFac} has weaker statistical significance in the presence of these additional factors.[45]

In summary, The results in Table 1.5 demonstrates that heterogeneity in the financial sectors has explanatory power for the cross-section of expected returns even in the presence of representative intermediary asset pricing factors presented in AEM and HKM.

---

[44]The pricing performance of AEM reported in their Table III and shown in Figure 1 of their paper, is substantially better than the one reported in Column 3 of Table 1.5 (their reported MAPE is only 1.31% and $R^2 = 0.77$). This difference can stem from two possible sources: (i) In 2015, the Federal Reserve substantially revised and updated Flow of Funds historical data for security broker-dealers, changing the way assets and liabilities were counted. They specifically changed their handling of using gross vs net repo. For more detail, see Z.1 Financial Accounts Technical Q&As. (ii) The test assets used in this chapter are different from AEM's. I have the same 35 equity portfolios (25 size/book-to-market and 10 momentum portfolios) but use both Treasury and corporate bonds from Nozawa (2017), while AEM only use 6 Treasury bonds sorted by maturity from CRSP.

[45]HIFac has positive correlation of 13% and 9% with AEM's leverage and HKM's capital risk factors, respectively.

1.6.3.2 Sorted Portfolios on Exposures to Heterogeneous Intermediary Factor

The positive price of risk associated with shocks to wealth share of dealers in the financial sector means assets that pay more in states with a low dealer wealth share (i.e. assets with low betas on y_t shocks) are viewed as hedges thus have lower expected returns in equilibrium.

To empirically verify the positive price of risk for innovations in the wealth share of dealers in the financial sector, I sort stocks based on their exposures to these shocks and form portfolios by quintiles on a 10-year trailing window. I consider all common stocks (share codes 10 and 11) in the CRSP universe from Amex, NASDAQ, and NYSE (exchange codes 1,2, and 3). For every stock i at quarter t, I regress its quarterly excess return on constant and innovations in the heterogeneous intermediary factor (HIFac), defined in equation (1.32):

$$R^e_{i,t} = \alpha_i + \beta_{i,\text{HIFac}}\,\text{HIFac}_t + \xi_{i,t} \tag{1.35}$$

The coefficient $\beta_{i,\text{HIFac}}$ measures the exposure of firm i's stock to the factor's innovations. I then sort stocks into quintiles every quarter according to their $\beta_{i,\text{HIFac}}$.

Consistent with model's implications, when sorted on β_{HIFac}, average risk premia are increasing from the portfolio of low-beta stocks to the high-beta quintile. In the Appendix, I report the average returns of the beta-sorted portfolios in Table 1.8, along with return volatilities, average book-to-market ratio, average market cap, and alphas from CAPM and Fama-French three-factor models. Excess returns are monotonically increasing from quintile one to five and the top portfolio earns an approximately 5% premium over the lowest quintile. In Appendix 1.12, I further verify the results above are robust to double-sorting with asset pricing factors from recent models with representative intermediaries.

This exercise demonstrates that the heterogeneity in the financial sector is an important risk factor and has pricing information above and beyond representative intermediary asset pricing factors in AEM and HKM: even within portfolios sorted based on AEM or HKM factor betas, I see a monotonic progression in returns from low- to high-HIFac beta portfolios.

1.6.4 Additional Robustness Checks

In Appendix 1.12, I project the heterogeneous intermediary factor (HIFac) onto the space of traded returns to form a factor-mimicking portfolio that mimics the HIFac. To further verify that this heterogeneity an important source of risk, I evaluate the heterogeneous intermediary factor-mimicking portfolio (HIMP) relative to the mimicking portfolios for representative intermediary factors in AEM and HKM. I show that the mimicking portfolios for these representative intermediary factors cannot fully span the HIMP and there is more to be captured by the heterogeneity within the financial sector. This exercise confirms my earlier results: the heterogeneity in the financial sector is an important risk factor and has pricing information above and beyond representative intermediary asset pricing factors in AEM and HKM.

1.7 Conclusion

This chapter studies the asset pricing implications of heterogeneity among financial intermediaries. Evidence on large balance sheet adjustments within the intermediary sector during the Great Recession is at odds with existing models featuring representative intermediaries. To explain and study the implications of these massive asset reallocations within the financial sector, in this chapter, I present a model with two main ingredients: intermediaries heterogeneous in their aggressiveness, and occasionally binding leverage constraints. Heterogeneity in intermediaries' risk appetite and margin constraints are both necessary in matching and understanding reallocations within the financial sector.

My model implies that a dealer deleveraging episode, comparable in magnitudes to the one observed during the recent financial crisis, leads to an approximately 55% increase in the risk premia and a 5% increase in endogenous volatility. In contrast, since balance sheet adjustments among different intermediaries does not affect the wealth share of the aggregate financial sector, in models with representative intermediaries these asset reallocations have no impact on asset prices and risk premia.

I show that the composition of the financial sector accounts for a substantial portion of the variation in risk premia in the model. With an independent and identically distributed aggregate endowment, the variation in expected returns is entirely due to intermediation frictions captured by both the health of the overall financial sector as well as its composition. I show that the wealth distribution among intermediaries, a measure of the composition of the financial sector, accounts for approximately 20% incremental variation in risk premia over a representative intermediary model.

The model also generates opposite cyclical dynamics for leverage of the two intermediary sectors, reconciling empirical evidence that has previously seemed contradictory through the lens of representative intermediary asset pricing models of AEM and HKM. To construct the SDF, AEM and HKM measure marginal utilities of different financial intermediaries: security broker-dealers, and bank holding companies, respectively. Given the economic mechanism of my model, it does not seem surprising that they arrive at conflicting asset pricing results.

I examine the empirical implications of the model for time-series predictability and the cross-section of returns. Consistent with the model, wealth share of broker-dealers in the financial sector, a measure of heterogeneity among intermediaries, strongly and negatively forecasts future excess returns on many assets. In particular, it leads to additional predictive power for market risk premium beyond many established forecasting variables in the literature. I also show that using only shocks to the relative wealth share of broker-dealers in the financial sector, explains the cross-section of equity and bond returns about as well or better than existing intermediary asset pricing models. I further document that including aggregate intermediary leverage as a second asset pricing factor increases cross-sectional fit by at least 10 percentages points, depending on whether the factor is the leverage of broker-dealers or bank holding companies.

1.8 Appendix 1: Proof of Propositions

Proof of Proposition 1

Proof. State variable x is the wealth share of financial sector: $x_t = \frac{W^A+W^B}{W}$ and state variable y is wealth share of type A agents in the financial sector: $y_t = \frac{W^A}{W^A+W^B}$, where W, W^A, and W^B are the aggregate wealth, wealth of A, and B agents, respectively.

Dynamics of x_t: From equation (1.5), W^A has the following law of motion

$$\frac{dW^A}{W^A} = \left(r + w_s^A(\mu - r) - c_A\right)\, dt + w_s^A \sigma\, dZ,$$

and W^B has the similar law of motion.

The law of motion for the numerator, $W^A + W^B$, will be

$$\begin{aligned}\frac{d\left(W^A + W^B\right)}{W^A + W^B} &= \left[r + \left(yw_s^A + (1-y)w_s^B\right)(\mu - r) - \left(yc_A + (1-y)c_B\right)\right] dt \\ &\quad + \left(yw_s^A + (1-y)w_s^B\right)\sigma dZ_t\end{aligned}$$

Define wealth share of agents A and B as $u \equiv W^A/W = xy$, and $v \equiv W^B/W = x(1-y)$, respectively.[46] Since the aggregate wealth is $W = W^A + W^B + W^C$, the law of motion for the denominator is

$$\begin{aligned}\frac{dW}{W} &= \left[r + \underbrace{\left(xyw_s^A + x(1-y)w_s^B + (1-x)w_s^C\right)}_{=1 \text{ (by stock market-clearing)}}(\mu - r) - \underbrace{\left(xyc_A + x(1-y)c_B + (1-x)c_C\right)}_{=F \text{ (by goods market-clearing)}}\right] dt \\ &\quad + \left[\underbrace{xyw_s^A + x(1-y)w_s^B + (1-x)w_s^C}_{=1}\right]\sigma\, dZ_t \\ &= \left[r + (\mu - r) - F\right] dt + \sigma\, dZ_t\end{aligned}$$

[46] Agent C's wealth share will then be $1 - u - v$.

From Ito's lemma for ratio of two stochastic processes,

$$\frac{dx}{x} = \kappa(\bar{x} - x)dt + \left[\left(yw_s^A + (1-y)w_s^B - 1\right)(\mu - r - \sigma^2) - yc_A - (1-y)c_B + F\right]dt$$
$$+ \left(yw_s^A + (1-y)w_s^B - 1\right)\sigma\, dZ_t$$

Thus from the dynamics of x in equation (1.12) we have

$$\mu_x = \left(yw_s^A + (1-y)w_s^B - 1\right)(\mu - r - \sigma^2) - yc_A - (1-y)c_B + F$$
$$\sigma_x = \left(yw_s^A + (1-y)w_s^B - 1\right)\sigma$$

Dynamics of y_t: The numerator of y is W^A, and its denominator is $\left(W^A + W^B\right)$ which its law of motion is calculated above. So from Ito's lemma for a ratio, we get

$$\frac{dy}{y} = \kappa(\bar{y} - y)dt + (1-y)\left[\left(w_s^A - w_s^B\right)(\mu - r) - c_A + c_B - \left(yw_s^A + (1-y)w_s^B\right)\left(w_s^A - w_s^B\right)\sigma^2\right]dt$$
$$+ (1-y)\left(w_s^A - w_s^B\right)\sigma\, dZ_t$$

Thus from the dynamics of y in equation (1.13) we have

$$\mu_y = \left(w_s^A - w_s^B\right)(\mu - r) - c_A + c_B - \left(yw_s^A + (1-y)w_s^B\right)\left(w_s^A - w_s^B\right)\sigma^2$$
$$\sigma_y = \left(w_s^A - w_s^B\right)\sigma$$

□

Proof of Proposition 2

Proof. Since σ_x and σ_y are finite, we trivially get

$$\lim_{x\to 0} x\,\sigma_x = 0,\ \forall y \quad \text{and} \quad \lim_{y\to 0} y(1-y)\,\sigma_y = \lim_{y\to 1} y(1-y)\,\sigma_y = 0,\ \forall x.$$

We only need to show $\lim_{x\to 1} x\,\sigma_x = 0\ \forall y$. The market clearing condition for the risky asset

when $x \to 1$ becomes $yw_s^A + (1-y)w_s^B = 1$.

So, from the expression for σ_x in equation (1.16), we have:

$$x\,\sigma_x = x\left[yw_s^A + (1-y)w_s^B - 1\right]\sigma$$

which goes to zero as $x \to 1$ for all y from the stock market-clearing. □

Proof of Proposition 3

Proof. We can write agent i's optimization problem in equation (1.8) as

$$0 = \max_{c_i, w_s^i} \left\{ f_i\left(c_{i,t}, V_{i,t}\right) dt + \mathbb{E}_t\left[dV_{i,t}\right] \right\}$$

Using Ito's lemma we have

$$\mathbb{E}_t[dV_i] = V_{i,W_i}\mathbb{E}_t[dW_i] + \frac{1}{2}V_{i,W_iW_i}\mathbb{E}_t\left[dW_i^2\right] + V_{i,J_i}\mathbb{E}_t[dJ_i] + \frac{1}{2}V_{i,J_iJ_i}\mathbb{E}_t\left[dJ_i^2\right] + V_{i,W_iJ_i}\mathbb{E}_t[dW_i dJ_i]$$

where V_{i,W_i} and V_{i,W_iW_i} are the first and second partial derivatives of V_i with respect to W_i (similarly for V_{i,J_i}, V_{i,J_iJ_i}, and V_{i,W_iJ_i}.) Also posit the following Ito process for marginal value of wealth J_i:

$$\frac{dJ_i}{J_i} = \mu_{J_{i,t}}dt + \sigma_{J_{i,t}}dZ,$$

with adapted processes $\mu_{J_{i,t}} = \mu_{J_i}(x_t, y_t)$ and $\sigma_{J_{i,t}} = \sigma_{J_i}(x_t, y_t)$. I will drop t subscripts for notational simplicity.

Using Ito's lemma, we can find the drift and diffusions μ_{J_i} and σ_{J_i}

$$\begin{aligned}\mu_{J_i} &= \frac{J_{i,x}}{J_i}\left[\kappa(\bar{x}-x) + x\,\mu_x\right] + \frac{J_{i,y}}{J_i}\left[\kappa(\bar{y}-y) + y(1-y)\mu_y\right] \\ &\quad + \frac{1}{2}\frac{J_{i,xx}}{J_i}x^2\sigma_x^2 + \frac{J_{i,xy}}{J_i}xy(1-y)\sigma_x\sigma_y + \frac{1}{2}\frac{J_{i,yy}}{J_i}y^2(1-y)^2\sigma_y^2 \qquad (1.36)\\ \sigma_{J_i} &= \frac{J_{i,x}}{J_i}x\,\sigma_x + \frac{J_{i,y}}{J_i}y(1-y)\sigma_y \qquad (1.37)\end{aligned}$$

Plugging in the felicity function $f(C,U)$ in (1.3) and the conjecture for value function V_i in (1.21) into the HJB equation above, using the budget constraint in (1.5) and the law of motion for J_i in (1.36) and (1.37), and dropping $W_i^{1-\gamma_i} J_i^{\frac{1-\gamma_i}{1-\psi_i}}$ and dt terms yields

$$
\begin{aligned}
0 = \max_{c_i, w_s^i} \; & \frac{1}{1-1/\psi_i}\left[\left(\frac{c_i}{J_i^{1/(1-\psi_i)}}\right)^{1-1/\psi_i} - (\rho+\kappa)\right] + \left[r - c_i + \kappa + w_s^i(\mu - r) - \frac{\gamma_i}{2}\left(w_s^i\right)^2\sigma^2\right] \\
& + \left(\frac{1}{1-\psi_i}\right)\left\{\frac{J_{i,x}}{J_i}\left[\kappa(\bar{x}-x) + x\mu_x\right] + \frac{J_{i,y}}{J_i}\left[\kappa(\bar{y}-y) + y(1-y)\mu_y\right]\right. \\
& \left. +(1-\gamma_i)\left(\frac{J_{i,x}}{J_i}x\sigma_x + \frac{J_{i,y}}{J_i}y(1-y)\sigma_y\right)w_s^i\sigma\right\} + \frac{1}{2}\left(\frac{1}{1-\psi_i}\right)\left[\left(\frac{\psi_i-\gamma_i}{1-\psi_i}\right)\left(\frac{J_{i,x}}{J_i}x\sigma_x + \frac{J_{i,y}}{J_i}y(1-y)\mu_y\right)^2\right. \\
& \left. + \frac{J_{i,xx}}{J_i}x^2\sigma_x^2 + 2\frac{J_{i,xy}}{J_i}xy(1-y)\sigma_x\sigma_y + \frac{J_{i,yy}}{J_i}y^2(1-y)^2\sigma_y^2\right] + \lambda_i\left(\bar{\theta}_t - w_s^i\right),
\end{aligned}
$$

where λ_i is proportional to the Lagrange multiplier on the time-varying margin constraint. The first-order condition for consumption-wealth ratio and portfolio share will lead to equations (1.23) and (1.25):

$$c_i = J_i \tag{1.38}$$

$$w_s^i = \frac{1}{\gamma_i}\left[\frac{\mu - r}{\sigma^2} + \left(\frac{1-\gamma_i}{1-\psi_i}\right)\left(\frac{J_{i,x}}{J_i}x\,\frac{\sigma_x}{\sigma} + \frac{J_{i,y}}{J_i}y(1-y)\frac{\sigma_y}{\sigma}\right)\right] - \frac{1}{\gamma_i\sigma^2}\lambda_i \tag{1.39}$$

When the margin constraint for agent i is slack, $\lambda_i = 0$ and we have

$$w_s^{i,*} = \frac{\mu - r}{\gamma_i\sigma^2} + \frac{1}{\gamma_i}\left(\frac{1-\gamma_i}{1-\psi_i}\right)\left(\frac{J_{i,x}}{J_i}x\,\frac{\sigma_x}{\sigma} + \frac{J_{i,y}}{J_i}y(1-y)\frac{\sigma_y}{\sigma}\right)$$

When the margin constraint for agent i is binding, λ_i is strictly positive and $w_s^{i,const} = \bar{\theta}_t$.

Plugging in the $w_s^{i,const}$ into (1.39), we get the expression for the multiplier on the time-varying margin constraint:

$$\lambda_i = (\mu - r) + \left(\frac{1-\gamma_i}{1-\psi_i}\right)\left(\frac{J_{i,x}}{J_i}x\sigma_x + \frac{J_{i,y}}{J_i}y(1-y)\sigma_y\right)\sigma - \gamma_i\sigma^2\bar{\theta}_t \tag{1.40}$$

□

1.9 Appendix 2: Numerical Procedure

The computation of equilibrium is reduced to solving three second-order PDEs for functions J_i for $i \in \{A, B, C\}$.[47] I use Chebyshev orthogonal collocation method to solve the model.[48] The HJB equation for agent i can be written as the following functional equation:

$$\mathcal{H}_i\left(J_i\right) = \mathbf{0}.$$

I express marginal value of wealth functions $J_A(x,y), J_B(x,y)$ and $J_C(x,y)$ as bivariate Chebyshev polynomials of order N (I use $N = 20$), that is, I approximate J_i with tensor product of Chebyshev polynomials of order N:

$$\widehat{J}_i(x,y) = \sum_{j=0}^{N}\sum_{k=0}^{N} a_{jk}^i \psi_j(\omega_j(x))\psi_k(\omega_k(y)),\ i \in \{A,B,C\}. \tag{1.41}$$

where ψ_j is the Chebyshev polynomial of degree $j = 0, 1, \ldots, N$, called the *basis function*, $\Psi_{jk}(x,y) = \psi_j(x)\psi_k(y)$ is a tensor product basis, $\{a_{jk}^i\}_{j,k=1}^N$ are unknown coefficients for agent i, and ω_j's are the Chebyshev nodes (collocation points) defined below.

I then plug in $\widehat{J}_i$ into the HJB equation for agent i to form the *residual equation*:

$$\mathcal{R}_i\left(\cdot \mid \boldsymbol{a}^i\right) = \mathcal{H}_i\left(\widehat{J}_i\right),$$

and find the vector of coefficients $\boldsymbol{a}^i$ that makes the residual equation as close to $\mathbf{0}$ as possible given some objective function $\rho\left(\mathcal{R}_i\left(\cdot \mid \boldsymbol{a}^i\right), \mathbf{0}\right)$:

$$\boldsymbol{a}^i = \arg\min_{\boldsymbol{a}^i}\ \rho\left(\mathcal{R}_i\left(\cdot \mid \boldsymbol{a}^i\right), \mathbf{0}\right)$$

[47] Duffie and Lions (1992) show existence and uniqueness of infinite-horizon stochastic differential utility by partial differential equation techniques in a Markov diffusion setting.

[48] For more details, see Judd (1992) and Judd (1998) and Computational Tools & Macroeconomic Applications, NBER Summer Institute 2011 Methods Lectures, Lawrence Christiano and Jesus Fernandez-Villaverde, Organizers.

The most common objective function is a *weighted residual* given some weight functions $\phi_j : \Omega \to \mathbb{R}^m$:

$$\rho\left(\mathcal{R}_i\left(\cdot \mid \boldsymbol{a}^i\right), \mathbf{0}\right) = \begin{cases} 0 & \text{if } \iint\limits_{\Omega\times\Omega} \phi_j(x)\phi_k(y)\mathcal{R}_i\left(\cdot \mid \boldsymbol{a}^i\right) dx\, dy = \mathbf{0}, \text{ for } j,k = 1,\ldots,N \\ 1 & \text{otherwise} \end{cases}$$

In the pseudo-spectral (or collocation) method, the weight functions are chosen as: $\phi_j(x) = \delta(x - x_i)$ where δ is the dirac delta function and x_i's are the collocation points. In the *orthogonal collocation* method, which I use to solve the model, the basis functions are a set of orthogonal Chebyshev polynomials and collocation points are given by the roots of the N^{th} polynomial.

Chebyshev polynomials of degree n can be easily defined recursively:

$$\begin{aligned} \psi_0(\omega) &= 1 \\ \psi_1(\omega) &= x \\ \psi_{n+1}(\omega) &= 2\omega\psi_n(\omega) - \psi_{n-1}(\omega) \end{aligned} \tag{1.42}$$

As mentioned above, the collocation points are the N zeros of the Chebyshev polynomial of order $N, (\psi_N(\omega_j) = 0)$, and are given by the following expression

$$\omega_j = \cos\left(\frac{2j-1}{2n}\pi\right),\ j = 1,\ldots,N.$$

These roots are clustered quadratically towards ± 1. Chebyshev polynomials are defined on $\omega_i \in [-1, 1]$. Since the state variables $x, y \in [0, 1]$ in my model, I use the linear transformation $x_j = (1 + \omega_j)/2$.[49]

I calculate the derivatives of these functions as well as the state variable dynamics, agents'

[49]For a general state space $x \in [x_L, x_H]$, we use a linear transformation $x_j = x_L + 0.5(x_H - x_L)(1 + \omega_j)$.

portfolio choice, risky asset return and volatility using the relevant equilibrium expressions. I then plug these quantities into the HJB equations (1.22) and project the resulting residuals onto the complete set of Chebyshev polynomials up to order N. I use the built-in MATLAB function `fsolve` to find the coefficients of J_i polynomials that make the projected residuals equal to zero. This results in a highly accurate solution for coefficients in the $\widehat{J_i}$ functions with errors in the order of 10^{-20}.

The numerical algorithm is summarized below.

1. From goods market-clearing conditions and differentiating it with respect to the state variable, we get expressions for dividend yield F and its derivatives with respect to x and y.

$$
\begin{aligned}
F &= xyJ_A + x(1-y)J_B + (1-x)J_C, \\
F_x &= yJ_A + (1-y)J_B - J_C + xyJ_{A,x} + x(1-y)J_{B,x} + (1-x)J_{C,x}, \\
F_y &= xJ_A - xJ_B + xyJ_{A,y} + x(1-y)J_{B,y} + (1-x)J_{C,y}, \\
F_{xx} &= 2yJ_{A,x} + 2(1-y)J_{B,x} - 2J_{C,x} + xyJ_{A,xx} + x(1-y)J_{B,xx} + (1-x)J_{C,xx}, \\
F_{yy} &= 2xJ_{A,y} - 2xJ_{B,y} + xyJ_{A,yy} + x(1-y)J_{B,yy} + (1-x)J_{C,yy}, \\
F_{xy} &= J_A - J_B + xJ_{A,x} - xJ_{B,x} + yJ_{A,y} + (1-y)J_{B,y} - J_{C,y} + xyJ_{A,xy} \\
&\quad + x(1-y)J_{B,xy} + (1-x)J_{C,xy},
\end{aligned}
$$

   where $J_{i,x}$ and $J_{i,xx}$ are the first and second partial derivative of J_i with respect to x, respectively, and similarly for $J_{i,y}$, $J_{i,yy}$ and $J_{i,xy}$.

2. Using market-clearing condition for the endowment claim, plugging in the expression for agent C's optimal portfolio choice w_s^C from (1.24), and substituting for $(\mu - r)/\sigma^2$ from the expression for $w_s^{A,*}$, we will get the first of the two equations that $w_s^{A,*}$ and

w_s^B have to satisfy:

$$
\begin{aligned}
1 &= xyw_s^{A,*} + x(1-y)w_s^B + (1-x)w_s^C \\
&= xyw_s^{A,*} + x(1-y)w_s^B \\
&+ (1-x)\frac{1}{\gamma_C}\left\{\frac{\mu-r}{\sigma^2} + \left(\frac{1-\gamma_C}{1-\psi_C}\right)\left[\frac{J_{C,x}}{J_C}x\left(yw_s^{A,*} + (1-y)w_s^B - 1\right) + \frac{J_{C,y}}{J_C}y(1-y)\left(w_s^{A,*} - w_s^B\right)\right]\right\} \\
&= xw_s^{A,*} + yw_s^B + (1-x)\frac{1}{\gamma_C}\left\{\gamma_A w_s^{A,*} - \left(\frac{1-\gamma_A}{1-\psi_A}\right)\left[\frac{J_{A,x}}{J_A}x\left(yw_s^{A,*} + (1-y)w_s^B - 1\right)\right.\right. \\
&\left.+\frac{J_{A,y}}{J_A}y(1-y)\left(w_s^{A,*} - w_s^B\right)\right] \\
&\left.+\left(\frac{1-\gamma_C}{1-\psi_C}\right)\left[\frac{J_{C,x}}{J_C}x\left(yw_s^{A,*} + (1-y)w_s^B - 1\right) + \frac{J_{C,y}}{J_C}y(1-y)\left(w_s^{A,*} - w_s^B\right)\right]\right\}
\end{aligned}
$$

To get the second equation, I plug in the expression for $(\mu - r)/\sigma^2$ from A's optimal portfolio $w_s^{A,*}$ in the expression for w_s^B:

$$
\begin{aligned}
w_s^B &= \frac{1}{\gamma_B}\left[\frac{\mu-r}{\sigma^2} + \left(\frac{1-\gamma_B}{1-\psi_B}\right)\left(\frac{J_{B,x}}{J_B}x\left(yw_s^A + (1-y)w_s^B - 1\right) + \frac{J_{B,y}}{J_B}y(1-y)(w_s^A - w_s^B)\right)\right] \\
&= \frac{1}{\gamma_B}\left[\gamma_A w_s^A - \left(\frac{1-\gamma_A}{1-\psi_A}\right)\left(\frac{J_{A,x}}{J_A}x\left(yw_s^A + (1-y)w_s^B - 1\right) + \frac{J_{A,y}}{J_A}y(1-y)(w_s^A - w_s^B)\right)\right. \\
&\left.+\left(\frac{1-\gamma_B}{1-\psi_B}\right)\left(\frac{J_{B,x}}{J_B}x\left(yw_s^A + (1-y)w_s^B - 1\right) + \frac{J_{B,y}}{J_B}y(1-y)(w_s^A - w_s^B)\right)\right]
\end{aligned}
$$

We can rewrite the systems of equation as

$$
\begin{aligned}
a_{11}w_s^{A,*} + a_{12}w_s^B &= b_1 \\
a_{21}w_s^{A,*} + a_{22}w_s^B &= b_2
\end{aligned}
$$

where

$$
\begin{aligned}
a_{11} &= xy + (1-x)\frac{1}{\gamma_C}\left[\gamma_A - \left(\frac{1-\gamma_A}{1-\psi_A}\right)\left(\frac{J_{A,x}}{J_A}xy + \frac{J_{A,y}}{J_A}y(1-y)\right)\right. \\
&\quad \left. + \left(\frac{1-\gamma_C}{1-\psi_C}\right)\left(\frac{J_{C,x}}{J_C}xy + \frac{J_{C,y}}{J_C}y(1-y)\right)\right], \\
a_{12} &= x(1-y) + (1-x)\frac{1}{\gamma_C}\left[-\left(\frac{1-\gamma_A}{1-\psi_A}\right)\left(\frac{J_{A,x}}{J_A}x(1-y) - \frac{J_{A,y}}{J_A}y(1-y)\right)\right. \\
&\quad \left. + \left(\frac{1-\gamma_C}{1-\psi_C}\right)\left(\frac{J_{C,x}}{J_x}x(1-y) - \frac{J_{C,y}}{J_C}y(1-y)\right)\right], \\
a_{21} &= \frac{1}{\gamma_B}\left[\gamma_A - \left(\frac{1-\gamma_A}{1-\psi_A}\right)\left(\frac{J_{A,x}}{J_A}xy + \frac{J_{A,y}}{J_A}y(1-y)\right) + \left(\frac{1-\gamma_B}{1-\psi_B}\right)\left(\frac{J_{B,x}}{J_B}xy + \frac{J_{B,y}}{J_B}y(1-y)\right)\right], \\
a_{22} &= -1 + \frac{1}{\gamma_B}\left[-\left(\frac{1-\gamma_A}{1-\psi_A}\right)\left(\frac{J_{A,x}}{J_A}x(1-y) + \frac{J_{A,y}}{J_A}y(1-y)\right)\right. \\
&\quad \left. + \left(\frac{1-\gamma_B}{1-\psi_B}\right)\left(\frac{J_{B,x}}{J_B}x(1-y) - \frac{J_{B,y}}{J_B}y(1-y)\right)\right], \\
b_1 &= 1 + (1-x)\frac{1}{\gamma_C}\left[-\left(\frac{1-\gamma_A}{1-\psi_A}\right)\frac{J_{A,x}}{J_A}x + \left(\frac{1-\gamma_C}{1-\psi_C}\right)\frac{J_{C,x}}{J_C}x\right], \\
b_2 &= \frac{1}{\gamma_B}\left[-\left(\frac{1-\gamma_A}{1-\psi_A}\right)\frac{J_{A,x}}{J_A}x + \left(\frac{1-\gamma_B}{1-\psi_B}\right)\frac{J_{B,x}}{J_B}x\right].
\end{aligned}
$$

The system of equations above can be solved easily to get $w_s^{A,*}$ and w_s^B.

3. Since the return volatility can be written as

$$\sigma = \frac{\sigma_D}{1 + \frac{F_x}{F}x\left[yw_s^A + (1-y)w_s^B - 1\right] + \frac{F_y}{F}y(1-y)\left(w_s^A - w_s^B\right)}, \tag{1.43}$$

when the margin constrains for agent A bind, from equation (1.7) with $\nu = 1$, we must have

$$w_s^{A,const} = \frac{1 - \frac{F_x}{F}x + \left(\frac{F_x}{F}x - \frac{F_y}{F}y\right)(1-y)w_s^B}{\alpha\sigma_D - \left[\frac{F_x}{F}x + \frac{F_y}{F}(1-y)\right]y} \tag{1.44}$$

So, we have $w_s^A \leq w_s^{A,const}$. Then from (1.25) we can find A and B's portfolio weights

in the risky asset

$$w_s^A = \min\left(w_s^{A,*}, w_s^{A,const}\right),$$

where $w_s^{A,const}$ is given in equation (1.44).

4. From stock market clearing, we can get C's optimal portfolio weight

$$w_s^C = \frac{1 - xy\, w_s^A - x(1-y)\, w_s^B}{1-x}$$

5. Using the expression for the return volatility in equation (1.19) and plugging in expressions for σ_x and σ_y from equations (1.16) and (1.17), the expression for return volatility is

$$\sigma = \frac{\sigma_D}{1 + \dfrac{F_x}{F} x\left[y w_s^A + (1-y) w_s^B - 1\right] + \dfrac{F_y}{F} y(1-y)\left(w_s^A - w_s^B\right)}.$$

6. Using the expression for σ above, state variable diffusions (σ_x and σ_y) can be found from equations (1.16) and (1.17):

$$\sigma_x = \left[y w_s^A + (1-y) w_s^B - 1\right]\sigma, \quad \text{and} \quad \sigma_y = \left(w_s^A - w_s^B\right)\sigma.$$

7. From the expression for w_s^C, σ, σ_x, and σ_y, the expected excess return (risk premium) on the risky asset is

$$\mu - r = \gamma_C w_s^C \sigma^2 - \left(\frac{1-\gamma_C}{1-\psi_C}\right)\left(\frac{J_{C,x}}{J_C} x\sigma_x + \frac{J_{C,y}}{J_C} y(1-y)\sigma_y\right)\sigma$$

8. Using the optimal consumption-wealth ratios $c_i = J_i$, we can then compute drifts of the state variables μ_x and μ_y as

$$\begin{aligned}
\mu_x &= \left[y w_s^A + (1-y) w_s^B - 1\right](\mu - r - \sigma^2) - y J_A - (1-y) J_B + F \\
\mu_y &= \left(w_s^A - w_s^B\right)(\mu - r) - J_A + J_B - \left[y w_s^A + (1-y) w_s^B\right]\left(w_s^A - w_s^B\right)\sigma^2.
\end{aligned}$$

9. From equation (1.18) the expected return on the risky asset can be calculated

$$\mu = \mu_D + F - \frac{F_x}{F}\left[\kappa(\bar{x}-x) + x(\mu_x + \sigma_D\sigma_x)\right] - \frac{F_y}{F}\left[\kappa(\bar{y}-y) + y(1-y)(\mu_y + \sigma_D\sigma_y)\right]$$
$$+ \left[\left(\frac{F_x}{F}\right)^2 - \frac{1}{2}\frac{F_{xx}}{F}\right]x^2\sigma_x^2 + \left[\left(\frac{F_y}{F}\right)^2 - \frac{1}{2}\frac{F_{yy}}{F}\right]y^2(1-y)^2\sigma_y^2 + \left[2\left(\frac{F_x}{F}\right)\left(\frac{F_y}{F}\right) - \frac{F_{xy}}{F}\right]xy(1-y)\sigma_x\sigma_y.$$

10. The real interest rate is

$$r = \mu - (\mu - r).$$

11. Plugging expressions above into agent i's HJB equations in (1.22), we get the residual functions for agent i:

$$0 = -(\rho+\kappa) + \frac{1}{\psi_i}J_i + \left(1 - \frac{1}{\psi_i}\right)\left[r + w_s^i(\mu - r) - \frac{\gamma_i}{2}\left(w_s^i\right)^2\sigma^2\right]$$
$$- \frac{1}{\psi_i}\left\{\frac{J_{i,x}}{J_i}\left[\kappa(\bar{x}-x) + x\mu_x\right] + \frac{J_{i,y}}{J_i}\left[\kappa(\bar{y}-y) + y(1-y)\mu_y\right] + (1-\gamma_i)\left(\frac{J_{i,x}}{J_i}x\sigma_x + \frac{J_{i,y}}{J_i}y(1-y)\sigma_y\right)w_s^i\sigma\right\}$$
$$- \frac{1}{2\psi_i}\left[\left(\frac{\psi_i - \gamma_i}{1-\psi_i}\right)\left(\frac{J_{i,x}}{J_i}x\sigma_x + \frac{J_{i,y}}{J_i}y(1-y)\mu_y\right)^2 + \frac{J_{i,xx}}{J_i}x^2\sigma_x^2 + 2\frac{J_{i,xy}}{J_i}xy(1-y)\sigma_x\sigma_y + \frac{J_{i,yy}}{J_i}y^2(1-y)^2\sigma_y^2\right].$$

1.10 Appendix 3: Additional Model Results

1.10.1 Heterogeneous vs. Representative Intermediaries

I simulate representative- and heterogeneous-intermediary models for 3,000 quarters 20,000 times and examine the distribution of risk premium volatility. Figure 1.10 in shows these distributions. As expected, the model with heterogeneous intermediaries exhibits more variation in risk premia than the one with a representative financial sector. Since the aggregate intermediary sectors in both models are (almost) identical, any excess variation in risk premia in the heterogeneous intermediary model has to be due to state variable y. In my calibration, approximately 20% of the variation in risk premia can be attributed to heterogeneity in the financial sector (state variable y). Therefore, failing to account for heterogeneity among intermediaries can lead to missing a substantial portion of the variation in risk premia.

1.10.2 Three-Dimensional Plots

Figure 1.12 plots various objects the unconstrained equilibrium, where $\bar{\theta}_t = \bar{m}$. All variables are functions of the two state variables in the model: x (wealth share of agents A and B, i.e. the financial sector) and y (wealth share of A agents in the financial sector). These are the same objects plotted in solid blue line in Figures 1.2 and 1.3 but in three dimensions.

Figure 1.13 presents various variables in the equilibrium with time-varying margin constraint in the endowment model with $\bar{\theta}_t = \frac{1}{\alpha\sigma_t}$ as functions of state variables (x_t, y_t). These are the same equilibrium objects plotted in dashed red line in Figures 1.2 and 1.3 but in three dimensions.

1.11 Appendix 4: Data Sources

Broker-Dealer and Holding Company Data

Balance sheet data for broker-dealers and bank holding companies are from Tables L.130 and L.131 of Financial Accounts of the United States (Flow of Funds) from Federal Reserves, respectively. As noted in the description of Table L.130,

> Security brokers and dealers are firms that buy and sell securities for a fee, hold an inventory of securities for resale, or do both. The firms that make up this sector are those that submit information to the Securities and Exchange Commission on one of two reporting forms, either the Financial and Operational Combined Uniform Single Report of Brokers and Dealers (FOCUS) or the Report on Finances and Operations of Government Securities Brokers and Dealers (FOGS). The major assets of the sector are collateral repayable from funding corporations in connection with securities borrowing (included in miscellaneous assets), debt securities and equities held for redistribution, customers' margin accounts, and security repurchase agreements (reverse repos). Sector operations are financed largely by net transactions with parent companies, customers' cash accounts, loans for purchasing and carrying securities from depository institutions, and security repurchase agreements.

Also from Table L.131's description for holding companies,

> ...the holding companies sector consists of all top-tiered bank holding companies, savings and loan holding companies, U.S. Intermediate Holding Companies (IHCs),

and securities holding companies (collectively "holding companies") that file the Federal Reserve's Form FR Y-9LP, Parent Company Only Financial Statements for Large Holding Companies, FR Y-9SP, Parent Company Only Financial Statements for Small Holding Companies, or FR 2320, Quarterly Savings and Loan Holding Company Report. Holding companies required to file FR Y-9LP include those with total consolidated assets of $1 billion or more or meet other criteria, such as having a material amount of debt or equity securities outstanding that are registered with the Securities and Exchange Commission, being engaged in significant nonbanking activity, or conducting off-balance-sheet activities either directly or through a nonbank subsidiary. Those holding companies required to file FR Y-9SP have total consolidated assets less than $1 billion. Form FR 2320 must be filed by top-tier savings and loan holding companies exempt from initially filing the Y-9LP or Y-9SP, because even though they own a savings and loan institution, that is not their primary line of business. Mutual stock companies that file the FR 2320 are excluded because they do not hold any assets or liabilities at the holding company level. The major assets of holding companies, other than small amounts of loans and securities, are net transactions with their subsidiaries; this includes equity investments in subsidiaries and associated banks and net balances due from subsidiaries and related depository institutions. The main source of funding for the sector is the issuance of corporate bonds and commercial paper.[50]

Test Assets

Test assets for time-series and cross-sectional asset pricing tests are from two sources: (i) equity portfolios (25 portfolios formed on size and book-to-market and 10 momentum portfolios) are from Ken French's Data Library, and (ii) non-equity assets are from HKM obtained from Asaf Manela's website and include 10 maturity-sorted US government and 10 corporate bond portfolios sorted on yield spreads, 6 sovereign bond portfolios based on a two-way sort on a bond's covariance with US equity market and bond's S&P rating, 54 portfolios of S&P 500 index options sorted on moneyness and maturity split by contract type (27 calls and 27 puts), 20 CDS portfolios sorted by spreads using single-name 5-year contracts, 23 commodity portfolios with at least 25 years of return data, and 12 foreign exchange currency portfolios, six sorted on interest rate differentials and six sorted on momentum. Except for Treasury bond portfolios which are from CRSP, non-equity test assets in HKM are from previous studies.

[50]The holding companies sector has a large increase in the level of assets and liabilities in the 2009:Q1 because a number of large financial institutions became bank holding companies. These companies (including Goldman Sachs, Morgan Stanley, American Express, CIT Group, GMAC, Discover Financial Services, and IB Finance) had not previously been included in the financial accounts.

Intermediary Asset Pricing Factors

AEM and HKM factors are from Tyler Muir's and Asaf Manela's websites, respectively. AEM leverage factor is defined as the seasonally adjusted growth rate in broker-dealer book leverage from Table L.130 of the Flow of Funds, where leverage is defined as total financial assets divided by total financial assets minus total volatility. The intermediary capital ratio in HKM is the ratio of total market equity to total market assets (book debt plus market equity) of primary dealer holding companies of the New York Fed. Shocks to capital ratio (HKM capital factor) are defined as AR(1) innovations in the capital ratio, scaled by the lagged capital ratio. Data for publicly-traded holding companies of primary dealers are from CRSP/Compustat and Datastream. Primary dealers are large and sophisticated institutions and serve as trading counterparties of the NY Fed in its implementation of monetary policy. For the current and historical list of primary dealers see this link.

1.12 Appendix 5: Robustness Checks for Empirical Results

1.12.1 Predictive Regressions

Exclude the Great Recession form the Sample

As a robustness check, I remove the Great Recession (years 2007 to 2009) form the sample and rerun the predictive regressions in equation (1.31) with the market excess return as the dependent variable. Table 1.6 presents the results. Consistent with the first two columns of Table 1.4 with the full sample, we also see negative and significant coefficient γ_y with additional predictive power over control variables in the sample excluding the 2008 financial crisis. Importantly, the predictive regression results are robust to excluding the Great Recession from the sample: It is not just the financial crisis that drives my predictability results.

Include Factors from Representative Intermediary-Based Models

As robustness, in Table 1.7, I further examine the predictive power of the composition of financial intermediaries (captured by state variable y) in the presence of factors from *representative* intermediary asset pricing models studied in AEM and HKM. That is, I include AEM and/or HKM factors in predictive regressions in equation (1.31):

$$R^i_{t+1\to t+4} - r^f_t = \gamma^i_0 + \gamma^i_y\, y_t + \gamma^i_{\text{Rep}}\, \text{Rep}_t + \gamma^i_{\text{Ctrl}}\, \text{Ctrl}_t + \varepsilon_{t+1\to t+4},$$

where Rep represents the vectors of representative intermediary factors: broker-dealer leverage from AEM and BHC capital ratio from HKM. Ctrl represents the vector of control variables that are known in the literature to forecast returns. I use the following control variables: wealth share of the aggregate financial sector (x from the model defined in 1.29), fluctuations in the aggregate consumption-wealth ratio (*cay* variable) defined in Lettau and Ludvigson (2001), real price-dividend (PD) and cyclically adjusted price-earnings (CAPE) ratios from Robert Shiller's website, and variance risk premium (VRP) from Bollerslev, Tauchen, and Zhou (2009). For reference, the first column repeats the regression in Column (1) of Table 1.4. In Column (2), HKM's intermediary capital ratio (CapRatio) is added as an additional predictor. We observe that the coefficient on CapRatio is not statistically significant and the R^2 is only slightly increased (from 0.43 to 0.45). Removing y in Column (3) substantially reduces R^2 by 13%, emphasizing the predictive power of my measure of intermediary heterogeneity beyond CapRatio. In Column (4), AEM's borker-dealer leverage (BDLev) is added as an additional predictor. Similarly, the coefficient on BDLev is not statistically significant and the R^2 is only slightly increased (from 0.43 to 0.47). Removing y in Column (5) however, does not substantially reduce R^2 (only by 2%). Finally, in the last column, all three predictors are included simultaneously: The coefficient on y remains negative and highly significant with a large R^2 of 0.47. The coefficient is also economically significant: a 1% decrease in wealth share of dealers in the financial sector predicts a 1.2% (quarterly, 4.8% annualized) increase in market risk premium over the next four quarters. As before, I include several control variables that are known in the literature to forecast returns:

Wealth share of the aggregate financial sector (x from the model defined in 1.29), fluctuations in the aggregate consumption-wealth ratio (*cay* variable) defined in Lettau and Ludvigson (2001), real price-dividend (PD) and cyclically adjusted price-earnings (CAPE) ratios from Robert Shiller's website, and variance risk premium (VRP) from Bollerslev, Tauchen, and Zhou (2009).

1.12.2 Cross-Sectional Asset Pricing Tests

HIFac's Pricing Performance

Figure 1.11 visually shows the HIFac's pricing performance: The top panel plots the annualized realized against the predicted excess returns for the 55 equity and bond portfolios when HIFac is the only pricing factor (Column (1) in Table 1.5). Most of the portfolios line up closely to the 45-degree line. The bottom panel is similar to the top panel when HIFac and AEM are used as pricing factors, corresponding to Column (4) in Table 1.5. The model slightly outperforms the one in panel (a) as shown in above.

One-Way Sorted CRSP Portfolios

To empirically verify the positive price of risk for innovations in the wealth share of dealers in the financial sector, I sort stocks based on their exposures to these shocks and form portfolios by quintiles on a 10-year trailing window. I consider all common stocks (share codes 10 and 11) in the CRSP universe from Amex, NASDAQ, and NYSE (exchange codes 1,2, and 3). For every stock i at quarter t, I regress its quarterly excess return on constant and innovations in the heterogeneous intermediary factor (HIFac), defined in equation (1.32):

$$R^e_{i,t} = \alpha_i + \beta_{i,\text{HIFac}}\,\text{HIFac}_t + \xi_{i,t}$$

The coefficient $\beta_{i,\text{HIFac}}$ measures the exposure of firm i's stock to the factor's innovations. I then sort stocks into quintiles every quarter according to their $\beta_{i,\text{HIFac}}$.

The average returns of the beta-sorted portfolios are reported in Table 1.8, along with

return volatilities, average book-to-market ratio, average market cap, and alphas from CAPM and Fama-French three-factor model. Consistent with model's implications, when sorted on β_{HIFac}, average risk premia are increasing from the portfolio of low-beta stocks to the high-beta quintile. Excess returns are monotonically increasing from quintile one to five and the top portfolio earns an approximately 5% premium over the lowest quintile.

Two-Way Sorted CRSP Portfolios

In this section, I verify the results above are robust to double-sorting with asset pricing factors from recent models with representative intermediaries. In this exercise, I independently double-sort CRSP stocks into three-by-three portfolios on their exposures to the heterogeneous intermediary factor (HIFac) and either AEM or HKM representative intermediary asset pricing factors. Table 1.9 reports returns for double-sorted portfolios on exposures to HIFac and AEM and HKM betas. The return spread on HIFac-beta-sorted portfolios is 4.34% and 3.14% per year among stocks with low exposures to the AEM leverage and HKM capital factors, respectively.

This exercise demonstrates that the heterogeneity in the financial sector is an important risk factor and has pricing information above and beyond representative intermediary asset pricing factors in AEM and HKM: even within portfolios sorted based on AEM or HKM factor betas, I see a monotonic progression in returns from low- to high-HIFac beta portfolios.

The Heterogeneous Intermediary Factor-Mimicking Portfolio

As emphasized above, the main argument of the paper is that the heterogeneity in the intermediary sector has important implication for asset prices. To conduct additional robustness tests, in this section, I project the heterogeneous intermediary factor (HIFac) onto the space of traded returns to form a factor-mimicking portfolio that mimics the HIFac. To further verify that this heterogeneity an important source of risk, I evaluate the heterogeneous intermediary factor-mimicking portfolio (HIMP) relative to the mimicking portfolios for representative intermediary factors in AEM and HKM. I show that the mimicking port-

folios for these representative intermediary factors cannot fully span the HIMP and there is more to be captured by the heterogeneity within the financial sector.

This approach also allows me to run tests using higher frequency data and longer time series. Moreover, since the mimicking portfolio is a traded excess return, I can evaluate the model by testing alphas in the time-series regression without the need to estimate the cross-section risk prices.

Construction of HIMP To construct mimicking portfolio of the heterogeneous intermediary factor (HIFac), I follow AEM and project this factor, onto the space of excess returns by running the following regression:

$$\text{HIFac}_t = a_{\text{HI}} + b'_{\text{HI}} \left[BL, BM, BH, SL, SM, SH, Mom, Bond\right]_t + \varrho_t, \tag{1.45}$$

where HIFac is the heterogeneous intermediary factor defined in equation (1.32), and BL, BM, BH, SL, SM are, respectively, the excess returns of the six Fama-French portfolios on size (Small and Big) and book-to-market (Low, Medium, and High), and Mom is the momentum factor, obtained from Ken French's data library. $Bond$ is the first principal component (PC) of excess returns on six Treasury bond portfolios sorted by maturity from CRSP. The heterogeneous intermediary mimicking portfolio (HIMP) is then given by

$$\text{HIMP}_t = \widetilde{b}'_{\text{HI}} \left[BL, BM, BH, SL, SM, SH, Mom, Bond\right]_t, \tag{1.46}$$

where $\widetilde{b}_{\text{HI}} = \frac{b'_{\text{HI}}}{\sum b_{\text{HI}}} = [-0.34, 0.20, -1.04, -0.09, 0.41, 1.64, 1.04, -0.83]$ positively loading on the momentum factor.

HIMP vs. Mimicking Portfolios for AEM and HKM Factors To further verify that my heterogeneous intermediary factor captures sources of risk beyond the factors from representative intermediary asset pricing models, in this section I evaluate the performance of HIMP with mimicking portfolios for AEM and HKM factors. I similarly construct mimicking portfolios for AEM's broker-dealer leverage and HKM's holding company capital factors

using quarterly data for the two factors from Tyler Muir's and Asaf Manela's websites, respectively.[51] The mimicking portfolio for the heterogeneous intermediary factor has Sharpe ratio of 0.45 over the sample period (1970Q1 to 2017Q3), much higher than Sharpe ratios for AEM and HKM factor-mimicking portfolios (0.21 and 0.27, respectively).

To evaluate the importance of heterogeneity in the financial sector above an beyond representative intermediary factors, I regress HIFac on mimicking portfolios for AEM and HKM factors in the following regression:

$$\text{HIMP}_t = \alpha_{\text{MP}} + \beta'_{\text{FMP}} \, \text{FMP}_t + \epsilon_t, \tag{1.47}$$

where FMP is either the mimicking portfolio for broker-dealer leverage factor from AEM (AEM_MP), or the mimicking portfolio for capital factor for primary dealers' holding companies from HKM (HKM_MP), or both AEM_MP and HKM_MP. Notice the mimicking portfolios are traded excess returns, thus I can evaluate the model by testing alphas in the time-series regression without the need to estimate the cross-section risk prices. If HIMP is fully "explained" by AEM_MP, HKM_MP, or both, I expect to see small and insignificant α_{MP} in the regression above. I find the opposite to be true, however.

Table 1.10 presents the results. In Columns 1 and 2, I run univariate regression where the dependent variables are AEM_MP and HKM_MP, respectively. In both cases the intercept, α_{MP} is statistically significant at 1% level and the R^2 of the regressions are relatively low at 0.14 and 0.32, respectively. In Column 3, I added value-weighted return from CRSP (MktRF) to HKM_MP as independent variables which leads to very similar results to Column 3. In Column 4, I add both AEM and HKM factor-mimicking portfolios as right-hand-side variables in equation (1.47). We observe a large and significant α_{MP} and relatively small R^2. Adding MktRF in Column 5 to the regression in Column 4, further strengthen the results.

As a robustness check, I build factor-mimicking portfolios by projecting them instead onto the Fama-French three factors, the momentum factor, and the first PC of bond portfolios,

[51] The loadings for AEM and HKM factor-mimicking portfolios are $\widetilde{b}_{\text{AEM}} = [-0.98, 0.50, -0.03, -0.26, 0.96, 0.05, 0.16, 0.59]$ and $\widetilde{b}_{\text{HKM}} = [0.30, 0.03, 0.58, -0.06, -0.16, 0.25, 0.09, -0.03]$.

and repeat the regressions in Table 1.10. I arrive at very similar results: time-series alphas are large and significantly different from zero with low R^2 in all regressions. See Table 1.11 in Appendix 1.12.

This exercise confirms my earlier results: the heterogeneity in the financial sector is an important risk factor and has pricing information above and beyond representative intermediary asset pricing factors in AEM and HKM.

Alternative Projections for Factor-Mimicking Portfolios (FMPs) In this section, I repeat the exercise in Section 1.12.2 with an alternative set of returns. I construct a mimicking portfolio for the heterogeneous intermediary factor (HIFac) by projecting it onto the space of excess returns running the following regression:

$$\text{HIFac}_t = a_{\text{HI}} + b'_{\text{HI}} \left[\text{MktRF}, SMB, HML, Mom, Bond\right]_t + \varrho_t, \tag{1.48}$$

where HIFac is the heterogeneous intermediary factor defined in equation (1.32), MktRF, SMB, and HML are the Fama-French three factors, Mom is the momentum factor, and $Bond$ is the first principal component (PC) of excess returns on six Treasury bond portfolios sorted by maturity from CRSP. The heterogeneous intermediary mimicking portfolio (HIMP) is then given by

$$\text{HIMP}_t = \widetilde{b}'_{\text{HI}} \left[\text{MktRF}, SMB, HML, Mom, Bond\right]_t, \tag{1.49}$$

where $\widetilde{b}_{\text{HI}} = \frac{b'_{\text{HI}}}{\sum b_{\text{HI}}} = [0.19, 0.51, 0.22, 0.29, -0.22]$ positively loading on the momentum factor.

I similarly construct mimicking portfolios for AEM's broker-dealer leverage and HKM's holding company capital factors using quarterly data for the two factors from Tyler Muir's and Asaf Manela's websites, respectively. The mimicking portfolio for the heterogeneous intermediary factor has Sharpe ratio of 0.43 over the sample period (1970Q1 to 2017Q3), much higher than Sharpe ratios for AEM and HKM factor-mimicking portfolios (0.24 and 0.28, respectively).

To evaluate the importance of heterogeneity in the financial sector above an beyond representative intermediary factors, I regress HIFac on mimicking portfolios for AEM and HKM factors in the following regression:

$$\text{HIMP}_t = \alpha_{\text{MP}} + \beta'_{\text{FMP}} \, \text{FMP}_t + \epsilon_t, \tag{1.50}$$

where FMP is either the mimicking portfolio for broker-dealer leverage factor from AEM (AEM_MP), or the mimicking portfolio for capital factor for primary dealers' holding companies from HKM (HKM_MP), or both AEM_MP and HKM_MP. If HIMP is fully "explained" by AEM_MP, HKM_MP, or both, I expect to see small and insignificant α_{MP} in the regression above. I find the opposite to be hold in the data, however. Table 1.11 presents the results. Time-series alphas are positive and significant at 1% level in all columns similar to the regressions in Table 1.10.

1.13 Appendix 6: Tables and Figures

Table 1.1. Parameter values for the endowment economy model

This tables reports parameter values used in calibrating the model. The parameters are calibrated to quarterly data.

	Description	Symbol	Value
Preference Parameters			
	Risk aversion of type A	γ_A	2.5
	Risk aversion of type B	γ_B	5.5
	Risk aversion of type C	γ_C	15
	EIS of type A	ψ_A	1.5
	EIS of type B	ψ_B	1.5
	EIS of type C	ψ_C	1.5
	Rate of time preference	ρ	0.001
Endowment and Demography			
	Endowment growth rate	μ_D	0.022
	Endowment volatility	σ_D	0.035
	Agents birth/death rate	κ	0.015
	Population share of type A	$\bar{u}$	0.05
	Population share of type B	$\bar{v}$	0.07
Margin Constraint			
	Constant leverage constraint	$\bar{m}$	4
	Tightness of the dynamic margin constraint	α	10

Table 1.2. State variables statistics

This table reports statistics for empirical proxies for model's two state variables in level and changes (Innov.). AC(j) represent j^{th} autocorrelation. Data is quarterly from 1970Q1 to 2017Q4.

	x^{data}		y^{data}	
	Level	Innov.	Level	Innov.
Mean	0.141	0.000	0.191	0.000
Standard Deviation	0.037	0.007	0.068	0.018
AC(1)	0.966	0.003	0.958	−0.185
AC(2)	0.933	0.046	0.930	0.087
AC(3)	0.899	0.105	0.896	−0.087
AC(4)	0.861	0.004	0.870	−0.075
AC(5)	0.825	−0.025	0.851	0.071

Table 1.3. Predictive regressions: y_t.

This table provides results for one-year ahead predictive regressions according to $R^i_{t+1\to t+4} - r^f_t = \gamma_0 + \gamma^i_y\, y_t + \varepsilon^i_{t+1\to t+4}$, using lagged equity share of broker-dealers in the financial sector, the empirical proxy for state variable y defined in equation (1.30) which captures the composition of the financial sector, as the predictor of interest. The dependent variables are excess holding period returns from quarter $t+1$ to quarter $t+4$ on the CRSP value-weighted portfolio (Mkt_{t+1}), mean excess return on 25 Fama-French size and book-to-market (FF25_{t+1}), 10 momentum (Mom_{t+1}) portfolios, 10 maturity-sorted US government and 10 US corporate bond portfolios sorted on yield spreads (US bonds_{t+1}), mean excess returns on six sovereign bonds (Sov. bonds_{t+1}), 54 portfolios of S&P 500 index options sorted on moneyness and maturity (Options_{t+1}), 20 CDS portfolios sorted by spreads (CDS_{t+1}), 23 commodity ($\text{Commod.}_{\cdot t+1}$), and 12 foreign exchange (FX_{t+1}) portfolios. Size/book-to-market and momentum portfolios and the risk-free rate data are from Ken French's website. Data on sovereign bonds, options, CDS, commodities, and FX portfolios are from He, Kelly, and Manela (2017). The sample quarterly from 1974Q2 to 2017Q2 for market, FF25 and momentum portfolios, and to 2012Q4 for HKM assets. Hodrick (1992) standard errors are reported in parentheses to adjust for the fact that overlapping quarterly observations are used to forecast annual returns.

	Dependent variable:								
	Mkt	FF25	Mom	US bonds	Sov. bonds	Options	CDS	Commod.	FX
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
y_t^{data}	-1.16^{***}	-0.98^{***}	-0.77^{**}	-0.11	-1.09^{**}	-1.57^{**}	-0.04	-0.40	0.37
	(0.38)	(0.28)	(0.33)	(0.14)	(0.46)	(0.61)	(0.13)	(0.45)	(0.31)
Const	0.39^{***}	0.34^{***}	0.27^{***}	0.09^{**}	0.39^{***}	0.46^{***}	0.03	0.14	-0.08
	(0.08)	(0.05)	(0.06)	(0.04)	(0.11)	(0.14)	(0.03)	(0.11)	(0.08)
Obs	173	173	173	152	62	100	44	102	132
R^2	0.10	0.11	0.08	0.01	0.15	0.12	0.005	0.01	0.05
Adj R^2	0.09	0.11	0.07	0.01	0.14	0.11	-0.02	0.003	0.05
Note:									*p<0.1; **p<0.05; ***p<0.01

Table 1.4. Predictive regressions: y_t and Controls.

This table provides results for one-year ahead predictive regressions according to $R^i_{t+1\to t+4} - r^f_t = \gamma_0 + \gamma^i_y\, y_t + \gamma^i_{\text{Ctrl}}\, \text{Ctrl}_t + \varepsilon^i_{t+1\to t+4}$, using lagged equity share of broker-dealers in the financial sector, the empirical proxy for state variable y defined in equation (1.30) which captures the composition of the financial sector, as the predictor of interest. Ctrl represents the vector of control variables that are known in the literature to forecast returns. I use the following control variables: wealth share of the aggregate financial sector (x from the model defined in 1.29), *cay* variable from Lettau and Ludvigson (2001), real price-dividend (PD) and cyclically adjusted price-earnings (CAPE) ratios from Robert Shiller's website, and variance risk premium (VRP) from Bollerslev, Tauchen, and Zhou (2009). The dependent variables are excess holding period returns from quarter $t+1$ to quarter $t+4$ on the CRSP value-weighted portfolio (Mkt_{t+1}), mean excess return on 25 Fama-French size and book-to-market (FF25_{t+1}), 10 momentum (Mom_{t+1}) portfolios, 10 maturity-sorted US government and 10 US corporate bond portfolios sorted on yield spreads (US bonds_{t+1}), mean excess returns on six sovereign bonds (Sov. bonds_{t+1}), 54 portfolios of S&P 500 index options sorted on moneyness and maturity (Options_{t+1}), 20 CDS portfolios sorted by spreads (CDS_{t+1}), 23 commodity (Commod._{t+1}), and 12 foreign exchange (FX_{t+1}) portfolios. Size/book-to-market and momentum portfolios and the risk-free rate data are from Ken French's website. Data on sovereign bonds, options, CDS, commodities, and FX portfolios are from He, Kelly, and Manela (2017). The sample quarterly from 1990Q1 to 2017Q3 for market, FF25 and momentum portfolios, and to 2012Q4 for HKM assets. Hodrick (1992) standard errors are reported in parentheses to adjust for the fact that overlapping quarterly observations are used to forecast annual returns.

	Dependent variable:									
	Mkt		FF25	FFmom	US bonds	Sov. bonds	Options	CDS	Commod.	FX
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
y_t		-1.77^{***}	-1.72^{***}	-1.80^{***}	0.43^{***}	-0.27	-2.14^{***}	-0.18	-1.15	0.44
		(0.46)	(0.51)	(0.45)	(0.18)	(0.64)	(0.74)	(0.16)	(0.74)	(0.40)
Const	0.04	-0.01	0.23	0.14	0.14^{***}	0.54^{***}	0.03	0.001	-0.06	-0.09
	(0.19)	(0.13)	(0.15)	(0.14)	(0.03)	(0.11)	(0.10)	(0.03)	(0.13)	(0.07)
Ctrls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs	109	109	109	109	88	62	84	44	88	76
R^2	0.31	0.46	0.31	0.44	0.35	0.26	0.57	0.38	0.16	0.41
Adj R^2	0.28	0.43	0.27	0.40	0.30	0.18	0.54	0.28	0.10	0.36
Note:										*p<0.1; **p<0.05; ***p<0.01

Table 1.5. Cross-sectional asset pricing tests

This table presents pricing results for the 25 size/book-to-market, 10 momentum, 10 maturity-sorted Treasury bond portfolios from CRSP with maturities in six month intervals up to five years, and 10 US corporate bond portfolios sorted on yield spreads from Nozawa (2017). The table reports the prices of risk and test diagnostics, including mean absolute pricing errors (MAPEs), and adjusted R^2s, and a χ^2 statistic and p-value that tests whether the pricing errors are jointly zero. Shanken (1992)-corrected and Fama and MacBeth (1973) t-statistics (t-Shanken and t-FM, respectively) are reported in parentheses. Heterogeneous intermediary factor (HIFac) is defined as the AR(1) innovations in the wealth share of dealers, scaled by their lagged wealth share according to equation (1.32). The AEM leverage factor (AEMLevFac) is defined as the seasonally-adjusted growth rate in broker-dealer leverage from Table L.130 of the Flow of Funds. HKM capital factor (HKMFac) is the shock to intermediary capital ratio in He, Kelly, and Manela (2017), defined as the ratio of total market equity to total market assets (book debt plus market equity) for bank holding companies of New York Fed's primary dealer counterparties. MktRF is the excess return on CRSP value-weighted portfolio from Ken French's website. The sample is quarterly from 1970Q1 to 2017Q4. Returns and risk premia are reported in percentage per year (quarterly percentages multiplied by four).

	(1)	(2)	(3)	(4)	(5)	(6)
HIFac	38.27*	57.35**		34.87*		52.47*
t-Shanken	(1.83)	(2.11)		(1.78)		(1.74)
t-FM	(2.24)	(3.11)		(2.02)		(2.82)
MktRF		3.86			6.93*	4.61
t-Shanken		(1.20)			(1.97)	(1.38)
t-FM		(1.27)			(2.20)	(1.52)
AEMLevFac			32.50***	21.66**		
t-Shanken			(2.68)	(2.42)		
t-FM			(3.73)	(3.14)		
HKMFac					12.55**	11.80**
t-Shanken					(2.57)	(2.12)
t-FM					(3.96)	(3.68)
Intercept	3.47***	4.14***	5.32**	3.01**	2.77*	3.99***
t-Shanken	(2.92)	(2.90)	(2.03)	(2.32)	(1.84)	(2.38)
t-FM	(3.62)	(4.32)	(2.68)	(3.11)	(2.90)	(4.19)
Observations	55	55	55	55	55	55
Adjusted R^2	0.61	0.61	0.39	0.72	0.63	0.69
MAPE, %	1.83	1.84	2.96	1.58	1.89	1.67
$\chi^2(N-K)$	195.39	133.53	151.78	167.17	121.35	95.34
p-value	0.00	0.00	0.00	0.00	0.00	0.00
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 1.6. Predictive regressions: Excluding the Great Recession

This table provides results for one-year ahead predictive regressions according to $R^i_{t+1\to t+4} - r^f_t = \gamma^i_0 + \gamma^i_y\, y_t + \gamma^i_{\text{Ctrl}}\, \text{Ctrl}_t + \varepsilon^i_{t+1\to t+4}$, using lagged equity share of broker-dealers in the financial sector, y^{data}_t defined in (1.30). Ctrl represents the vector of control variables that are known in the literature to forecast returns. I use the following control variables: wealth share of the aggregate financial sector (x from the model defined in 1.29), *cay* variable from Lettau and Ludvigson (2001), real price-dividend (PD) and cyclically adjusted price-earnings ratios (CAPE) from Robert Shiller's website, and variance risk premium (VRP) from Bollerslev, Tauchen, and Zhou (2009). The dependent variables are excess holding period returns from quarter $t+1$ to quarter $t+4$ on the CRSP value-weighted portfolio (Mkt_{t+1}), mean excess return on 25 Fama-French size and book-to-market (FF25_{t+1}), 10 momentum (Mom_{t+1}) portfolios, 10 maturity-sorted US government and 10 US corporate bond portfolios sorted on yield spreads (US bonds_{t+1}), mean excess returns on six sovereign bonds (Sov. bonds_{t+1}), 54 portfolios of S&P 500 index options sorted on moneyness and maturity (Options_{t+1}), 20 CDS portfolios sorted by spreads (CDS_{t+1}), 23 commodity (Commod._{t+1}), and 12 foreign exchange (FX_{t+1}) portfolios. Size/book-to-market and momentum portfolios and the risk-free rate data are from Ken French's website. Data on sovereign bonds, options, CDS, commodities, and FX portfolios are from He, Kelly, and Manela (2017). Broker-Dealer leverage is calculated using data from Table L.130 of Flow of Funds and is defined: Total Financial Assets/(Total Financial Assets − Total Liabilities). The sample is quarterly from 1974Q1 to 2017Q3 for market, FF25 and momentum portfolios, and to 2012Q4 for HKM assets. The Great Recession (years 2007–2009) is excluded form the sample. Regression coefficients on BDLev are multiplied by 100. Hodrick (1992) standard errors are reported in parentheses to adjust for the fact that overlapping quarterly observations are used to forecast annual returns.

	Dependent variable:	
	Mkt	
	(1)	(2)
y_t		-1.06^{***}
		(0.34)
Constant	-0.12	-0.13
	(0.14)	(0.10)
Controls	Yes	Yes
Observations	97	97
R^2	0.48	0.54
Adjusted R^2	0.45	0.51
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 1.7. Predictive regressions for the market excess return: Robustness

This table provides results for one-year ahead predictive regressions using lagged equity share of broker-dealers in the financial sector, the empirical proxy for state variable y defined in equation (1.30) which captures the composition of the financial sector, as well as intermediary equity capital ratio (from HKM) and leverage of broker-dealers (from AEM) as predictors of interest. I use the same control variables as in Table 1.4: wealth share of the aggregate financial sector (x from the model defined in 1.29), *cay* variable from Lettau and Ludvigson (2001), real price-dividend (PD) and cyclically adjusted price-earnings (CAPE) ratios from Robert Shiller's website, and variance risk premium (VRP) from Bollerslev, Tauchen, and Zhou (2009). The dependent variable is excess holding period returns from quarter $t+1$ to quarter $t+4$ on the CRSP value-weighted portfolio (Mkt_{t+1}). The sample quarterly from 1974Q1 to 2017Q3. Broker-Dealer leverage is calculated using data from Table L.130 of Flow of Funds and is defined: Total Financial Assets/(Total Financial Assets − Total Liabilities). The capital ratio for New York Fed's primary dealer holding companies are downloaded from Asaf Manela's website. Hodrick (1992) standard errors are reported in parentheses to adjust for the fact that overlapping quarterly observations are used to forecast annual returns.

	Dependent variable: Mkt_{t+1}					
	(1)	(2)	(3)	(4)	(5)	(6)
y_t	-1.80^{***}	-1.98^{***}		-0.97^{*}		-1.19^{***}
	(0.43)	(0.48)		(0.51)		(0.43)
CapRatio_t		3.31	1.22			1.68
		(3.58)	(4.58)			(2.25)
BDLev_t				-1.35	-1.98^{***}	-1.13^{*}
				(0.85)	(0.56)	(0.59)
Constant	0.01	0.18	0.10	0.05	0.09	0.13
	(0.13)	(0.26)	(0.38)	(0.15)	(0.17)	(0.22)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	109	109	109	109	109	109
R^2	0.46	0.49	0.32	0.51	0.48	0.51
Adjusted R^2	0.43	0.45	0.28	0.47	0.45	0.47
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 1.8. One-way sorted CRSP portfolios on exposures to the heterogeneous intermediary factor

This table reports average excess returns, alphas, volatility, average book-to-market ratio, and average market capitalization for portfolios formed on their exposure to shocks to dealer wealth share in the financial sector. Shocks to dealer wealth share (HIFac) are defined as AR(1) innovations in the wealth share, scaled by the lagged wealth share as shown in equation (1.32). Data is quarterly from 1970Q1 to 2017Q3. Returns, volatilities, and alphas are annualized.

	L				H	HML
	(1)	(2)	(3)	(4)	(5)	(6)
Average Excess Return (%)	11.66	11.53	12.81	14.34	16.65	4.98
Volatility (%)	19.69	19.08	21.76	26.41	35.53	26.72
β_{HIFac}	−0.20	0.33	0.69	1.19	2.21	2.41
t-stat	−0.89	1.50	2.77	4.07	5.89	10.67
α_{CAPM}	4.77	4.33	4.56	4.75	4.44	−0.32
t-stat	3.40	3.90	3.63	2.84	1.57	−0.10
α_{FF3}	3.89	2.88	3.36	3.98	4.02	0.14
t-stat	3.14	4.21	5.45	5.04	2.24	0.05
Average Market Cap ($bn)	5.28	3.66	2.40	1.97	0.89	–

Table 1.9. Two-way sorted CRSP portfolios

This table reports average excess returns for portfolios independently double-sorted on their exposure to shocks to dealer wealth share in the financial sector (HIFac) and beta to the AEM leverage factor (AEM LevFac), as well as, double-sorted portfolios on HIFac beta and HKM capital ratio factor (HKM CapFac) beta. Shocks to dealer wealth share (HIFac) are defined as AR(1) innovations in the wealth share, scaled by the lagged wealth share as shown in equation (1.32). AEM leverage and HKM capital factors are from Tyler Muir's and Asaf Manela's websites, respectively. Returns are annualized in percentage points. Data is quarterly from 1970Q1 to 2017Q3.

	HIFac				
AEM LevFac	(1)	(2)	(3)	(3)−(1)	t-stat
(1)	12.85	14.05	17.19	4.34	1.49
(2)	11.47	12.75	14.63	3.16	1.25
(3)	11.42	12.19	14.67	3.24	1.25
(3)−(1)	−1.43	−1.86	−2.54	–	–
t-stat	−0.65	−0.85	−0.91	–	–
	HIFac				
HKM CapFac	(1)	(2)	(3)	(3)−(1)	t-stat
(1)	10.05	11.39	13.19	3.14	1.20
(2)	11.91	12.48	14.82	2.91	1.28
(3)	14.75	15.17	17.74	2.99	1.13
(3)−(1)	4.70	3.78	4.55	–	–
t-stat	1.48	1.30	1.48	–	–

Table 1.10. The heterogeneous intermediary mimicking portfolio (HIMP): Comparing models

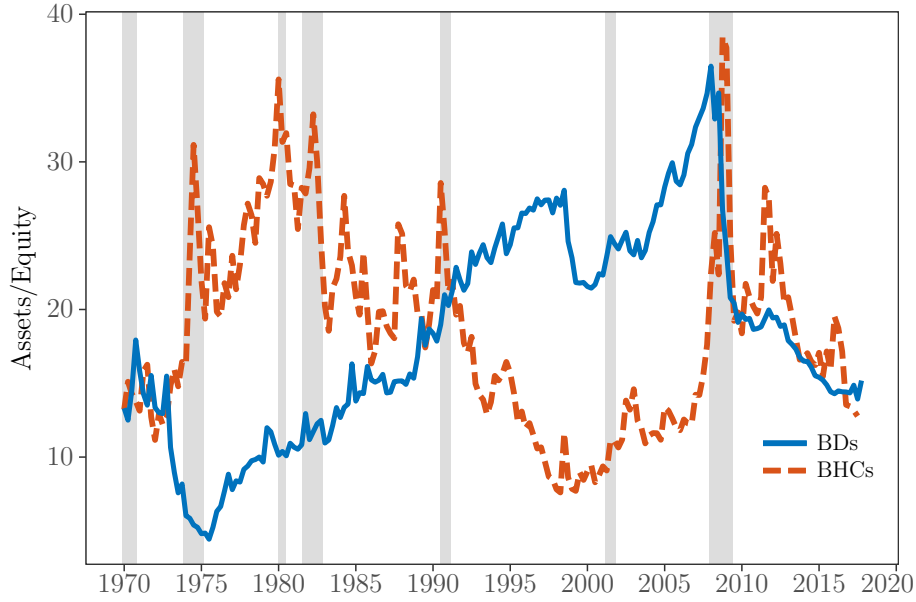
This table presents time-series regression results of heterogeneous intermediary mimicking portfolio (HIMP) on mimicking portfolios for the representative intermediary factors in AEM and HKM according to: $\text{HIMP}_t = \alpha_{\text{MP}} + \beta'_{\text{FMP}} \text{FMP}_t + \epsilon_t$, where FMP is either the mimicking portfolio for broker-dealer leverage factor from AEM (AEM_MP), or the mimicking portfolio for capital factor for primary dealers' holding companies from HKM (HKM_MP), or both AEM_MP and HKM_MP. The factor-mimicking portfolios are constructed by projecting the heterogeneous intermediary, AEM's leverage, and HKM's capital factors unto the space of equity and bond returns according to equations (1.45) and (1.46). The sample is quarterly from 1970Q1 to 2017Q3. Standard errors are in parentheses.

	Dependent variable: HIMP				
	(1)	(2)	(3)	(4)	(5)
α_{MP}	5.03***	4.03***	4.14***	3.74***	3.91***
	(0.94)	(0.85)	(0.85)	(0.83)	(0.83)
AEM MP	0.72***			0.37***	0.49***
	(0.13)			(0.12)	(0.13)
HKM MP		0.94***	0.68**	0.82***	0.15
		(0.10)	(0.27)	(0.10)	(0.30)
MktRF			0.27		0.65**
			(0.26)		(0.27)
Observations	191	191	191	191	191
R^2	0.14	0.32	0.33	0.35	0.37
Adjusted R^2	0.14	0.32	0.32	0.35	0.36
Note:				*p<0.1; **p<0.05; ***p<0.01	

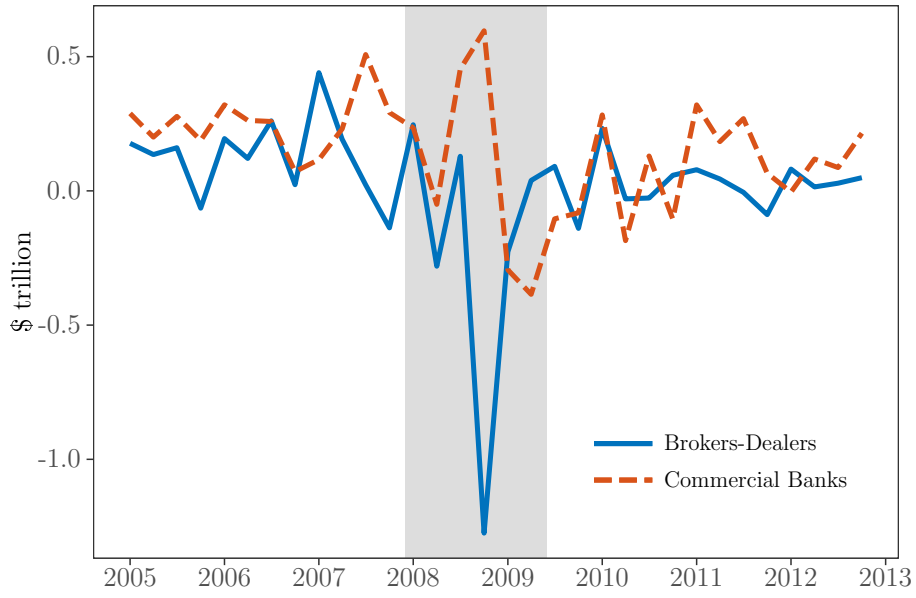
Table 1.11. The heterogeneous intermediary mimicking portfolio (HIMP): Comparing models with alternative projections

This table presents time-series regression results of heterogeneous intermediary mimicking portfolio (HIMP) on mimicking portfolios for the representative intermediary factors in AEM and HKM according to: $\text{HIMP}_t = \alpha_{\text{MP}} + \beta'_{\text{FMP}} \text{FMP}_t + \epsilon_t$, where FMP is either the mimicking portfolio for broker-dealer leverage factor from AEM (AEM_MP), or the mimicking portfolio for capital factor for primary dealers' holding companies from HKM (HKM_MP), or both AEM_MP and HKM_MP. The factor-mimicking portfolios are constructed by projecting the heterogeneous intermediary, AEM's leverage, and HKM's capital factors unto the space of equity and bond returns according to equations (1.45) and (1.46). The sample is quarterly from 1970Q1 to 2017Q3. Standard errors are in parentheses.

	Dependent variable: HIMP				
	(1)	(2)	(3)	(4)	(5)
α_{MP}	1.21***	0.96***	0.95***	0.88***	0.92***
	(0.25)	(0.22)	(0.23)	(0.22)	(0.22)
AEM MP	0.55***			0.23**	0.32***
	(0.09)			(0.09)	(0.11)
HKM MP		0.38***	0.39***	0.34***	0.13
		(0.04)	(0.11)	(0.04)	(0.14)
MktRF			−0.01		0.13
			(0.07)		(0.08)
Observations	191	191	191	191	191
R^2	0.16	0.35	0.35	0.38	0.38
Adjusted R^2	0.15	0.35	0.35	0.37	0.37
Note:			*p<0.1; **p<0.05; ***p<0.01		



(a) Intermediary leverage



(b) Change in total financial assets

Figure 1.1. Leverage and change in assets of different financial intermediaries

Panel (a) presents time-series of leverage for different financial intermediaries: security broker-dealers (BDs) and bank holding companies (BHCs). Leverage for broker-dealers (solid blue line) is defined as the ratio total financial assets to total equity (total financial assets minus total liabilities) from Table L.130 of the Flow of Funds. BHC leverage (dashed red line) is defined as the ratio of total market assets (book debt plus market equity) to total market equity constructed for publicly-traded holding companies of the NY Fed's primary dealer counterparties using CRSP/Compustat and Datastream, where market equity is outstanding shares times stock price and book debt is total assets minus common equity. Panel (b) presents quarterly change in total financial assets for BDs and Private Depository Institutions (DIs). BDs' (solid blue line) and DIs' (dashed red line) total financial assets are from Tables L.130 and L.110 of the Financial Accounts of the United States (Flow of Funds), respectively. Data is quarterly from 1970Q1 to 2017Q4. The vertical shaded bars indicate NBER recessions.

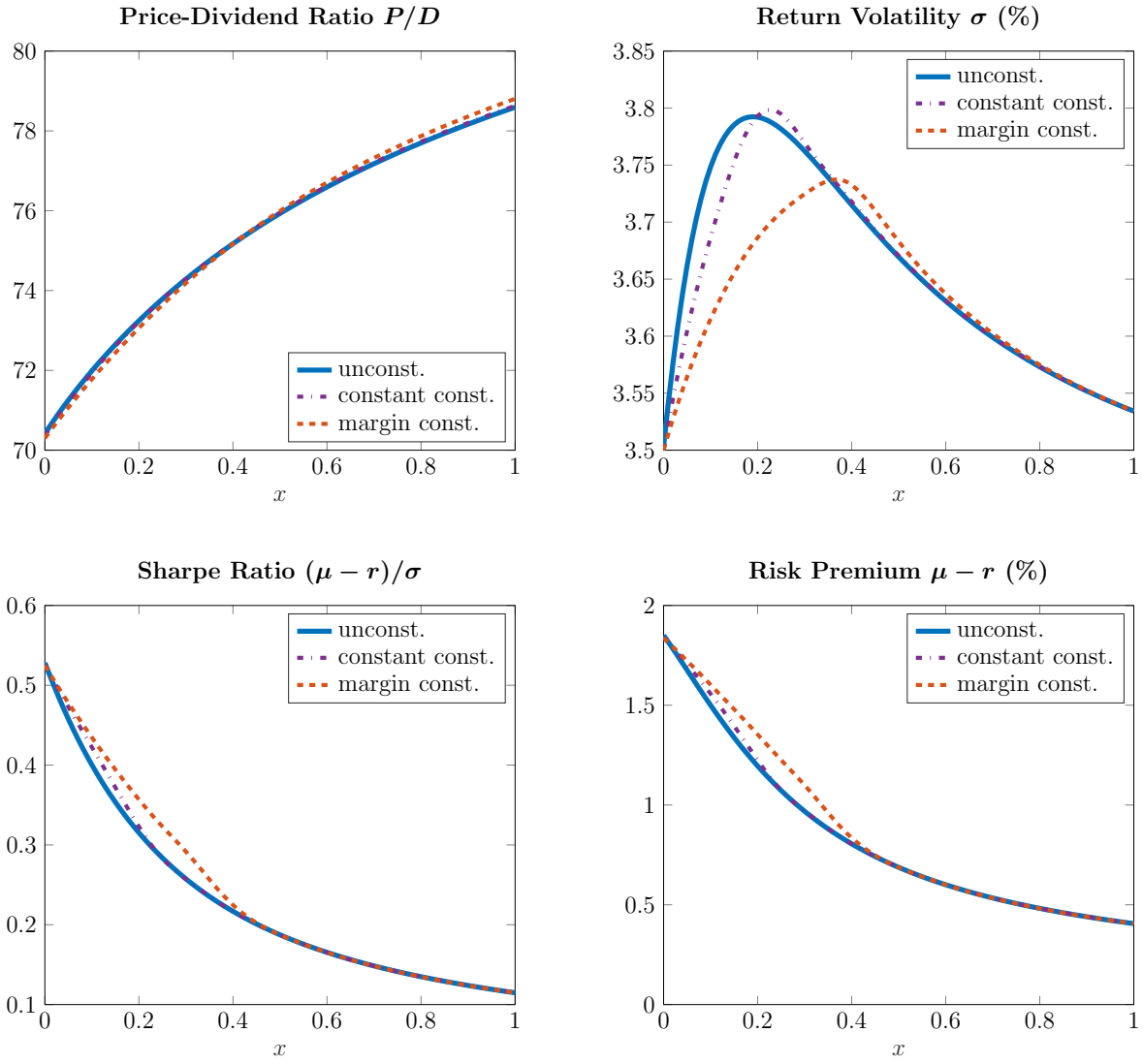


Figure 1.2. Risk premia, the price of risk, valuation, and volatility

This figure presents price-dividend ratio $1/F$, return volatility σ, Sharpe ratio and risk premium on the endowment claim in constrained and unconstrained equilibria as functions of state variable x (wealth share of the financial sector i.e. type A and B agents) under the benchmark parameters in Table 1.1. Each quantity is plotted against state variable x_t while the value of the second state variable y_t (wealth share of type A investors, i.e. broker-dealers, in the financial sector) is fixed at 0.56 (its value at the stochastic steady state). The solid blue line corresponds to the unconstrained economy, the dash-dotted purple line corresponds to the economy with a constant portfolio constraint ($\bar{\theta}_t = \bar{m}$), and the dashed red line corresponds to the the economy with a Value-at-Risk (VaR)-type margin constraint ($\bar{\theta}_t = \frac{1}{\alpha\sigma_t}$). Three-dimensional plots are provided in Appendix 1.10.2.

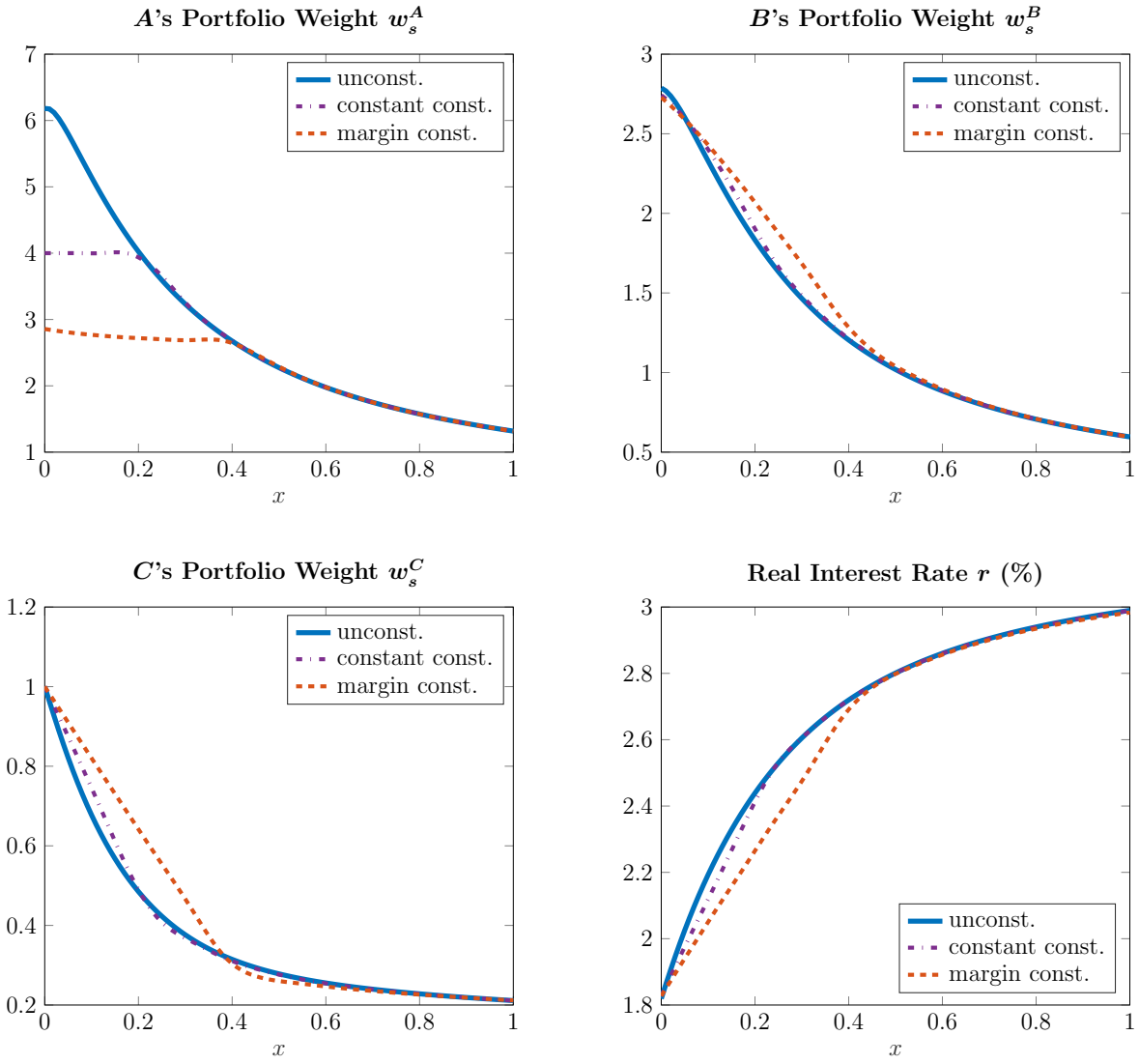


Figure 1.3. Optimal portfolios and the risk-free rate

This figure presents portfolio weights of each type of agent w_s^A, w_s^B, and w_s^C as well as the real interest rate r_t in constrained and unconstrained equilibria as functions of state variable x (wealth share of the financial sector i.e. A and B agents) under the benchmark parameters in Table 1.1. Each quantity is plotted against state variable x_t while the value of the second state variable y_t (wealth share of type A investors in the financial sector) is fixed at 0.56 (its value at the stochastic steady state). The solid blue line corresponds to the unconstrained economy, the dash-dotted purple line corresponds to the economy with a constant portfolio constraint ($\bar{\theta}_t = \bar{m}$), and the dashed red line corresponds to the the economy with a Value-at-Risk (VaR)-type margin constraint ($\bar{\theta}_t = \frac{1}{\alpha\sigma_t}$). Three-dimensional plots are provided in Appendix 1.10.2.

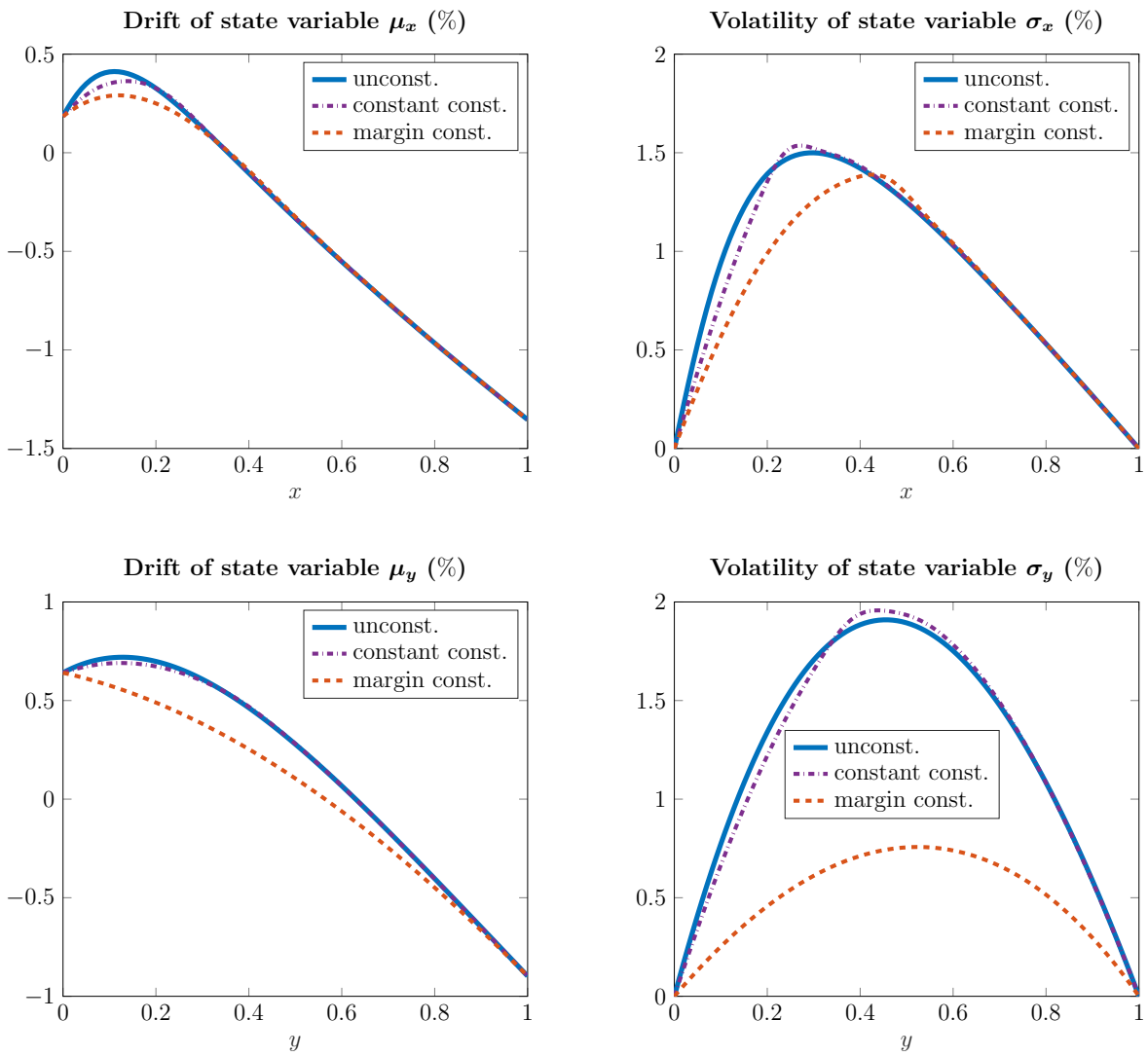


Figure 1.4. Dynamics of the endogenous state variables

This figure presents dynamics of the state variables x and y (wealth share of the financial sector i.e. A and B agents, and wealth share of type A investors in the financial sector, respectively) in constrained and unconstrained equilibria under the benchmark parameters in Table 1.1. Drift and volatility of state variable x (i.e. μ_x, σ_x) are plotted as functions of x while the value of the state variable y is fixed at 0.56. Drift and volatility of state variable y (i.e. μ_y, σ_y) are plotted as functions of y while the value of the state variable x is fixed at 0.25. The solid blue line corresponds to the unconstrained economy, the dash-dotted purple line corresponds to the economy with a constant portfolio constraint ($\bar{\theta}_t = \bar{m}$), and the dashed red line corresponds to the the economy with a Value-at-Risk (VaR)-type margin constraint ($\bar{\theta}_t = \frac{1}{\alpha\sigma_t}$). Three-dimensional plots are provided in Appendix 1.10.2.

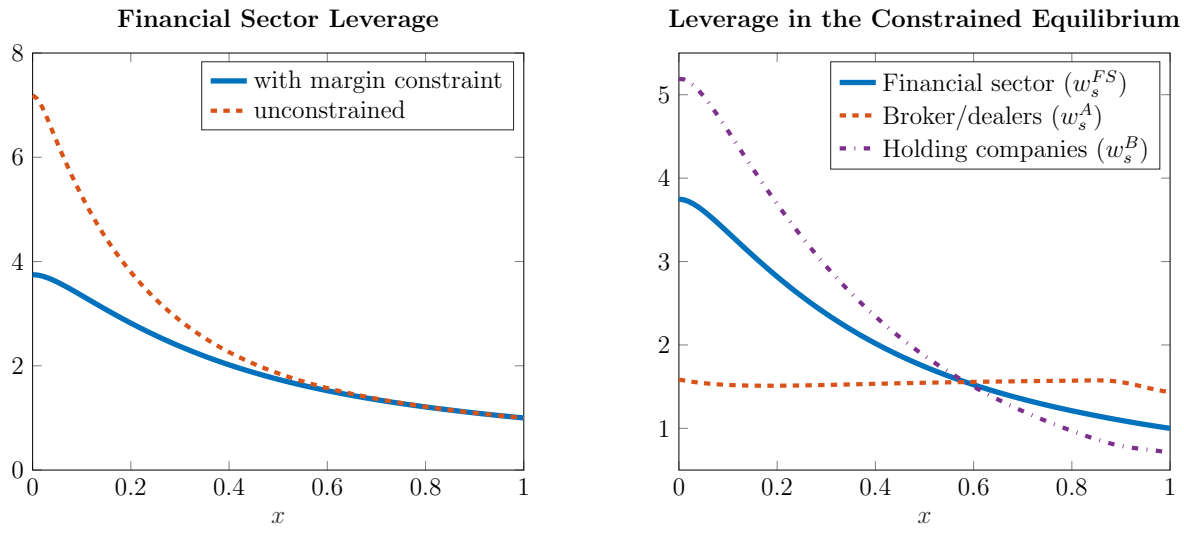


Figure 1.5. Cyclical properties of intermediary leverage

This figure presents optimal intermediary leverage in the unconstrained and constrained equilibria under parameters listed in Table 1.1. The left panel plots leverage of the financial sector (equation 1.28) in the unconstrained equilibrium (dashed red line) and the model with time-varying margin constraints, $\bar{\theta}_t = \frac{1}{\alpha\sigma_t}$ (solid blue line). The right panel plots intermediary leverage in the main model with endogenous margin constraints. The solid blue line corresponds to leverage of the financial sector (w_s^{FS}), the dashed red line presents broker-dealers' leverage (w_s^A), and the dash-dotted purple line corresponds to leverage of bank holding companies (w_s^B). Each quantity is plotted against state variable x_t (wealth share of the financial sector i.e. A and B agents) while the value of the second state variable y_t (wealth share of type A investors in the financial sector) is held fixed at 0.56 (its stochastic steady state value).

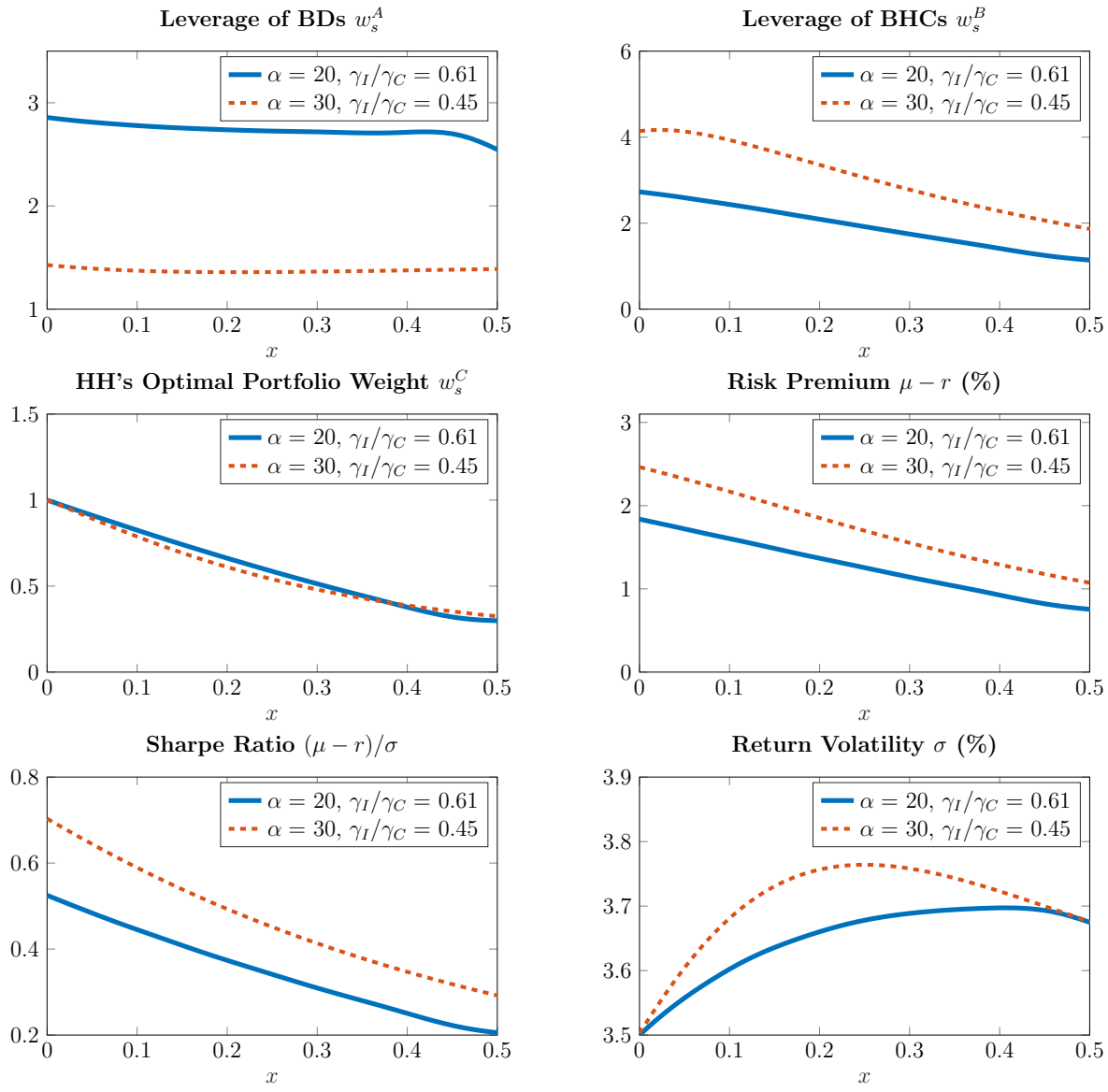


Figure 1.6. Asset reallocation within the financial sector

This figure presents portfolio weights for dealers, holding companies, and households (A, B, and C types, respectively), as well as, the risk premium and Sharpe ratio of the risky claim on the aggregate endowment, and the volatility of the risky asset return in the baseline model (solid blue line) and a model with tighter margin constraints and less risk averse financial sector (dashed red line). The changes in tightness of the margin constraint (parameter α) and relative risk aversion of the financial and household sectors (γ_I/γ_C) are such that leverage of A (B) types is reduced (increased) by approximately 47% (72%): changes documented during the Great Recession in Figure 1.1a. Each quantity is plotted against state variable x (wealth share of the financial sector i.e. A and B agents) while value of the state variable y (wealth share of dealers i.e. type A investors in the financial sector) is fixed at 0.56, its value at the stochastic steady state. Parameters for the baseline model are presented in Table 1.1.

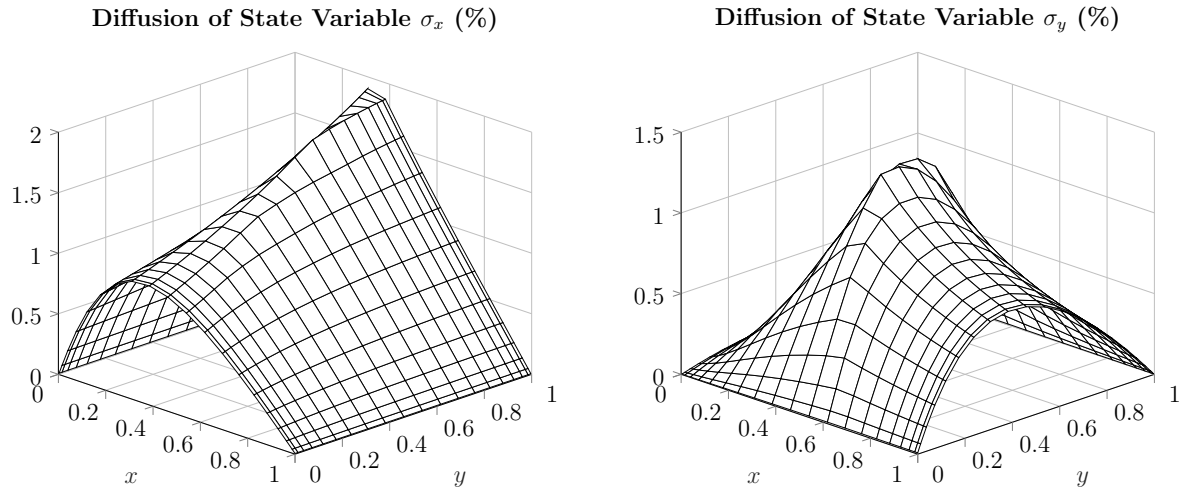


Figure 1.7. State variable diffusions

This figure presents the diffusions of state variables x and y (σ_x and σ_y, respectively) in the economy with time-varying margin constraints as functions of state variable x (wealth share of the financial sector i.e. type A and B agents) and y_t (wealth share of type A agents in the financial sector) under the benchmark parameters in Table 1.1.

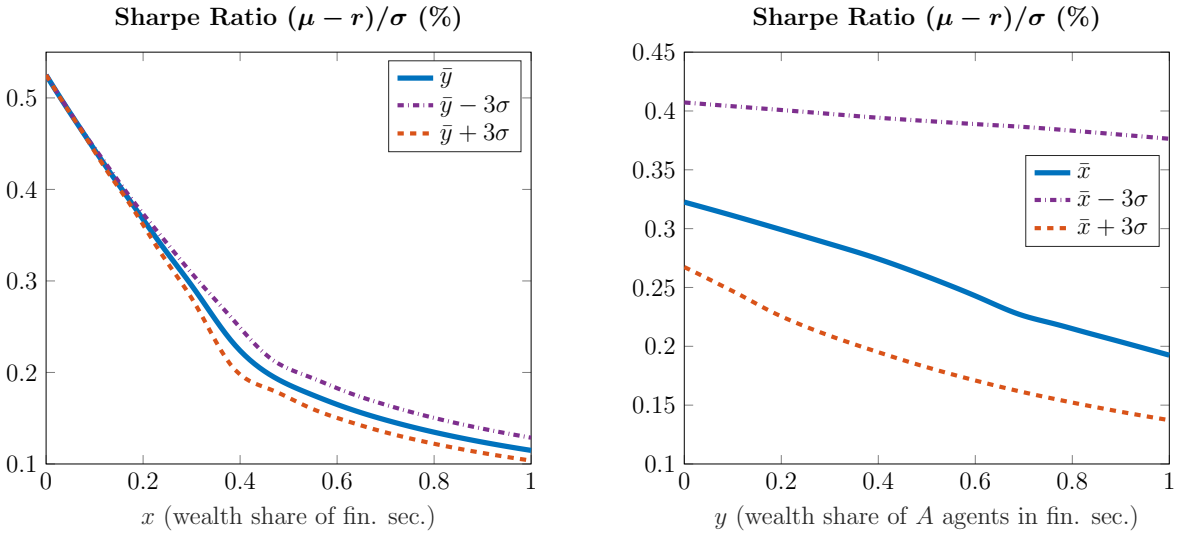


Figure 1.8. Price of risk

This figure presents the Sharpe ratio of the risk asset in the economy with time-varying margin constraints as functions of state variable x (wealth share of the financial sector i.e. type A and B agents) and y_t (wealth share of type A agents in the financial sector) under the benchmark parameters in Table 1.1.

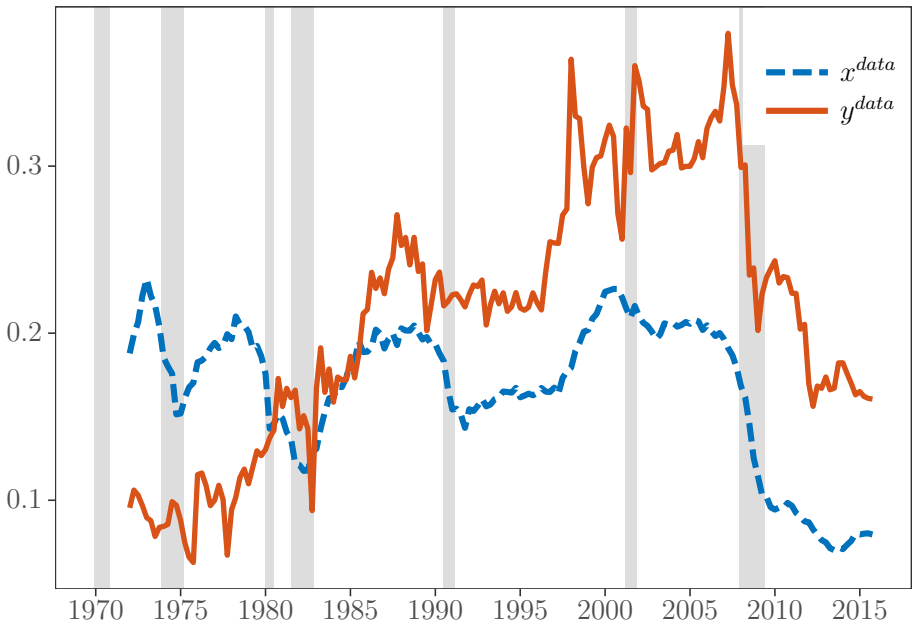


Figure 1.9. State variables x and y in the data

This figure presents the three-month moving average of monthly wealth share of the financial sector, x^{data}, and quarterly book equity share of the broker-dealers in the financial sector, y^{data}, defined in equations (1.29) and (1.30), respectively. Financial sector is identified as firms in the CRSP universe for whom the first two digits of the header SIC code (HSICCD in CRSP) equals 60–67. Book equity for BDs and BHCs are computed from the Flow of Funds Tables L.130 and L.131. Sample period is from 1970 to 2017. The vertical shaded bars indicate NBER recessions.

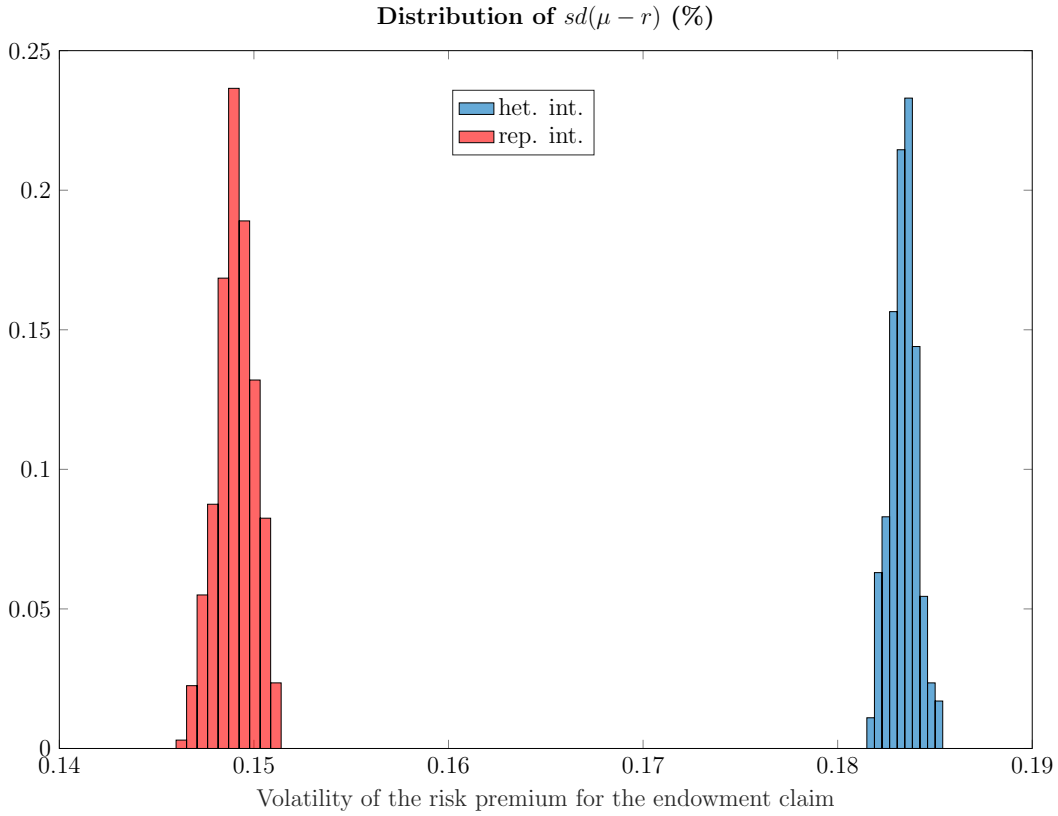
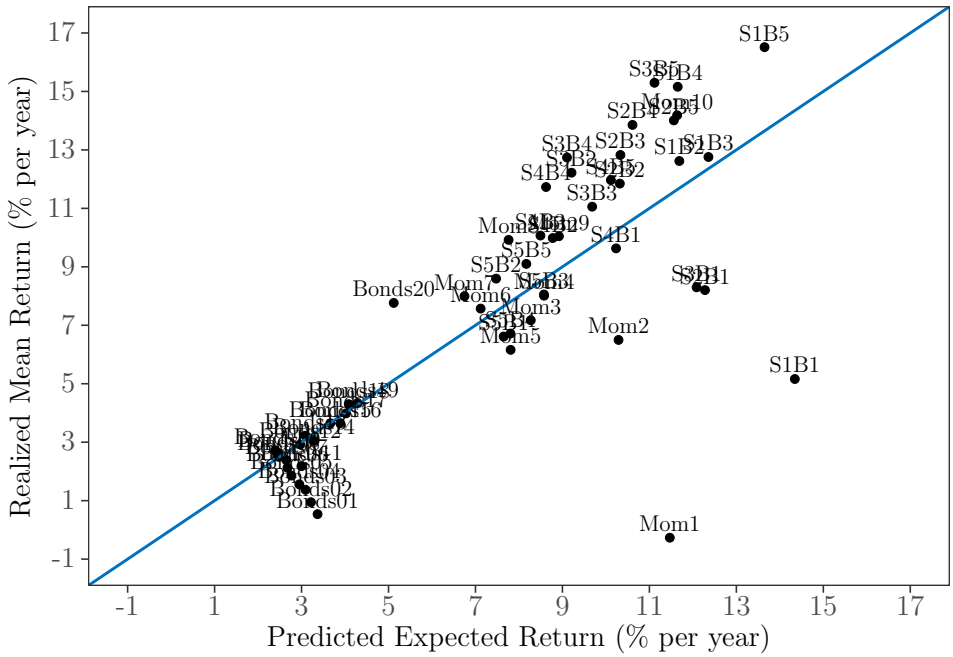
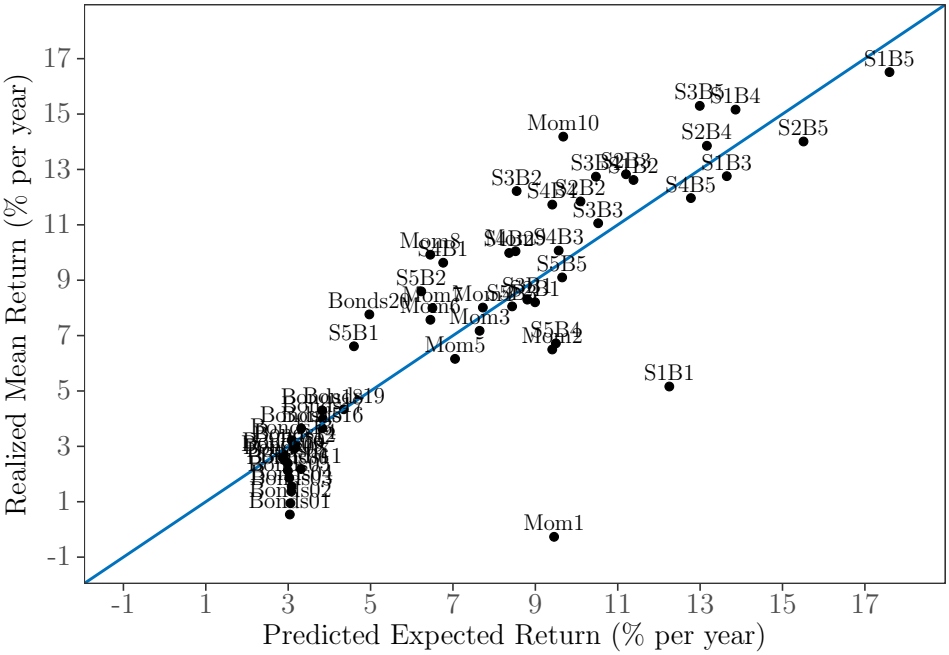


Figure 1.10. Heterogeneous and representative intermediaries

This figure presents distribution of risk premia volatility in models with representative (red) and heterogeneous (blue) intermediaries. I simulate each model 20,000 times for 3,000 quarters. Notice that horizontal axis is the volatility of the risk premium for the endowment claim, which has relatively low volatility ($\sigma_D = 3.5\%$) relative to the market (approximately 16%). Therefore, equity premium volatility implied by the model is about five to six times larger than that of the endowment claim (approximately 1% for the heterogeneous intermediary model, for example).



(a) HIFac only



(b) HIFac and AEM

Figure 1.11. Realized versus predicted mean returns: Heterogeneous intermediary factor

This figure presents the realized mean excess returns of 35 equity portfolios (25 size and book-to-market-sorted portfolios and 10 momentum-sorted portfolios) and 10 Treasury bond portfolios (sorted by maturity), and 10 US corporate bond portfolios (sorted by yield spread) against the mean excess returns predicted by the single heterogeneous intermediary risk factor when only the heterogeneous intermediary factor (HIFac) (panel a) and HIFac and AEM factors (panel b) are used as pricing factor, respectively. The sample is quarterly from 1970Q1 to 2017Q4. Returns are reported in percent per year (quarterly percentages multiplied by four).

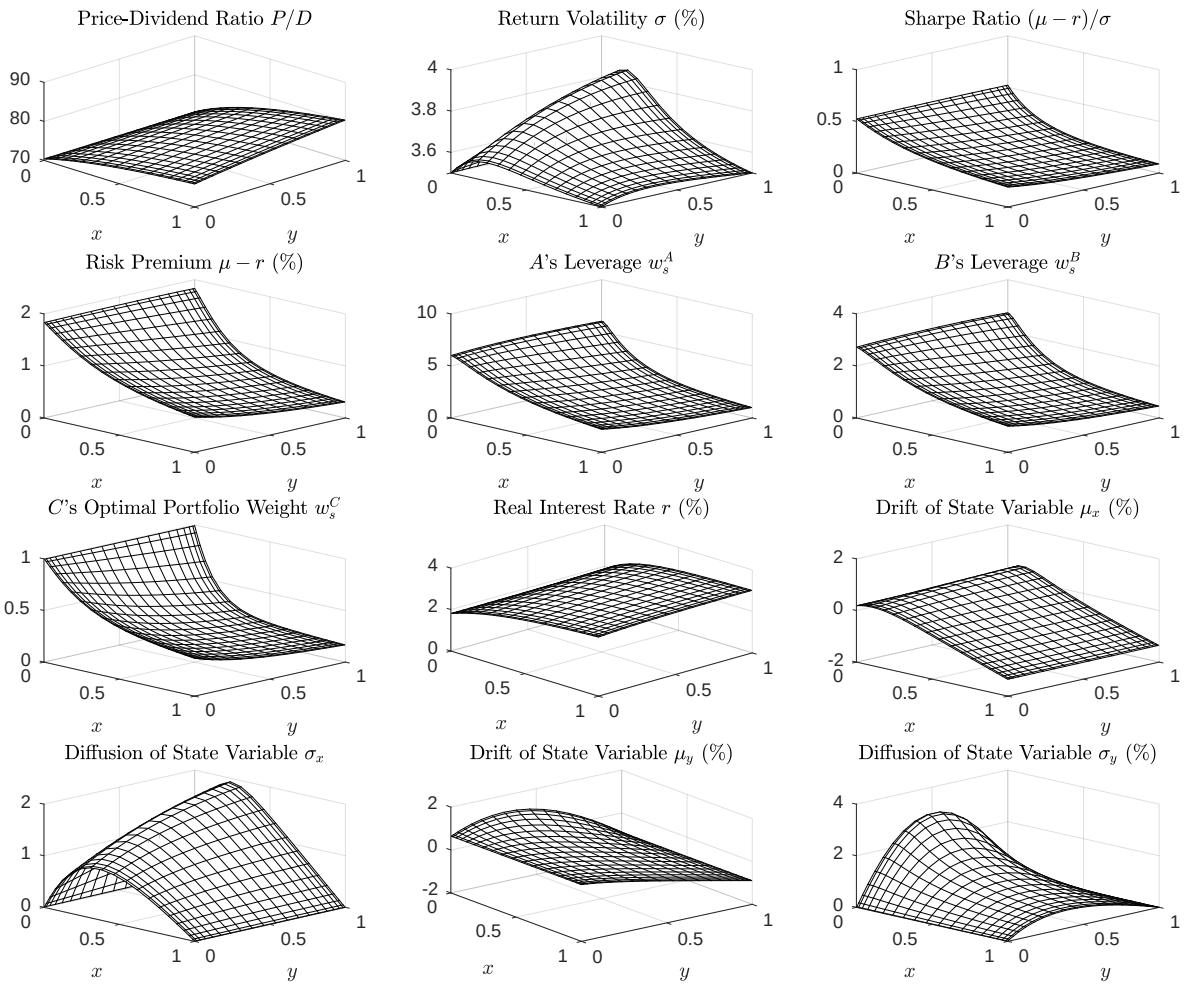


Figure 1.12. Equilibrium in the unconstrained economy

This figure presents price-dividend ratio $1/F$, return volatility σ, Sharpe ratio and risk premium on the endowment claim, optimal portfolio weights of each type of agent (w_s^A, w_s^B, and w_s^C) as well as the real interest rate r_t and the drift and diffusion of state variables x and y (μ_x, σ_x, μ_y, and σ_y, respectively) in the *frictionless* economy as functions of state variable x (wealth share of the financial sector i.e. type A and B agents) and y_t (wealth share of type A agents in the financial sector) under the benchmark parameters in Table 1.1.

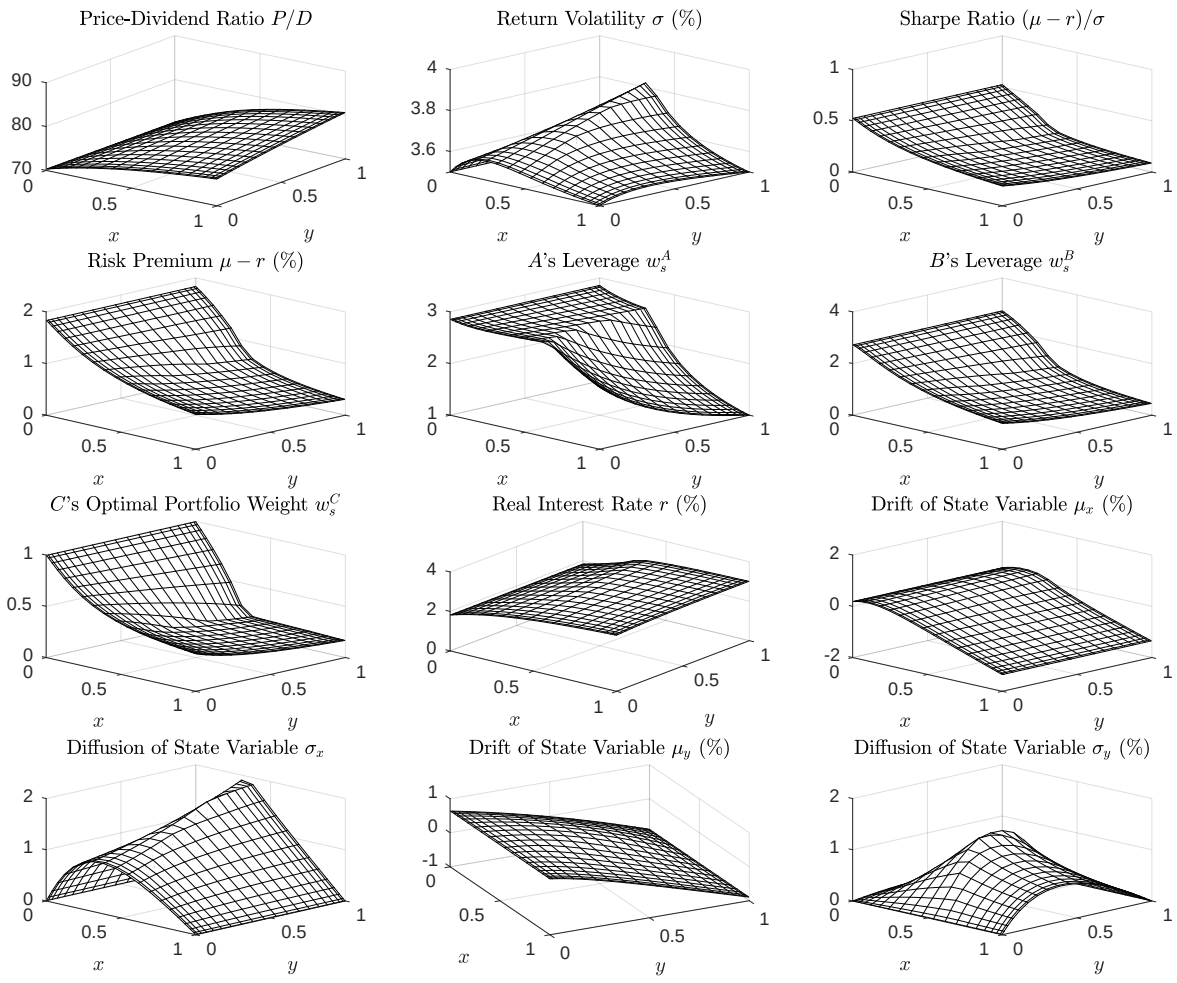


Figure 1.13. Equilibrium in the economy with time-varying margin constraints

This figure presents price-dividend ratio $1/F$, return volatility σ, Sharpe ratio and risk premium on the endowment claim, optimal portfolio weights of each type of agent (w_s^A, w_s^B, and w_s^C) as well as the real interest rate r_t and the drift and diffusion of state variables x and y (μ_x, σ_x, μ_y, and σ_y, respectively) in the economy with *time-varying margin constraints* as functions of state variable x (wealth share of the financial sector i.e. type A and B agents) and y_t (wealth share of type A agents in the financial sector) under the benchmark parameters in Table 1.1.

1.14 Bibliography

Acharya, V. V., L. H. Pedersen, T. Philippon, and M. Richardson (2017). Measuring Systemic Risk. *The Review of Financial Studies* 30.1, pp. 2–47.

Adrian, T. and N. Boyarchenko (2015). Intermediary Leverage Cycles and Financial Stability. Working Paper, Federal Reserve Bank of New York Staff Reports.

Adrian, T., E. Etula, and T. Muir (2014). Financial Intermediaries and the Cross-Section of Asset Returns. *The Journal of Finance* 69.6, pp. 2557–2596.

Adrian, T., E. Moench, and H. S. Shin (2014). Dynamic Leverage Asset Pricing. Federal Reserve Bank of New York Staff Report No. 625.

Adrian, T. and H. S. Shin (2014). Procyclical Leverage and Value-at-Risk. *Review of Financial Studies* 27.2, pp. 373–403.

Ang, A., S. Gorovyy, and G. B. van Inwegen (2011). Hedge Fund Leverage. *Journal of Financial Economics* 102.1, pp. 102–126.

Bansal, R., D. Kiku, and A. Yaron (2012). An Empirical Evaluation of the Long-Run Risks Model for Asset Prices. *Critical Finance Review* 1.1, pp. 183–221.

Bansal, R. and I. Shaliastovich (2013). A Long-Run Risks Explanation of Predictability Puzzles in Bond and Currency Markets. *The Review of Financial Studies* 26.1, pp. 1–33.

Bansal, R. and A. Yaron (2004). Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. *The Journal of Finance* 59.4, pp. 1481–1509.

Basak, S. and D. Cuoco (1998). An Equilibrium Model with Restricted Stock Market Participation. *The Review of Financial Studies* 11.2, pp. 309–341.

Begenau, J., S. Bigio, and J. Majerovitz (2017). Lessons from the Financial Flows of the Great Recession. Working Paper.

Ben-David, I., F. Franzoni, and R. Moussawi (2012). Hedge Fund Stock Trading in the Financial Crisis of 2007–2009. *The Review of Financial Studies* 25.1, pp. 1–54.

Bernanke, B. S. and M. Gertler (1989). Agency Costs, Net Worth, and Business Fluctuations. *The American Economic Review* 79.1, pp. 14–31.

Bernanke, B. S., M. Gertler, and S. Gilchrist (1999). The financial accelerator in a quantitative business cycle framework. *Handbook of Macroeconomics* 1, pp. 1341–1393.

Bhamra, H. S. and R. Uppal (2009). The Effect of Introducing a Non-Redundant Derivative on the Volatility of Stock-Market Returns When Agents Differ in Risk Aversion. *The Review of Financial Studies* 22.6, pp. 2303–2330.

Bhamra, H. S. and R. Uppal (2014). Asset Prices with Heterogeneity in Preferences and Beliefs. *The Review of Financial Studies* 27.2, pp. 519–580.

Bollerslev, T., G. Tauchen, and H. Zhou (2009). Expected Stock Returns and Variance Risk Premia. *The Review of Financial Studies* 22.11, pp. 4463–4492.

Brunnermeier, M. K., T. M. Eisenbach, and Y. Sannikov (2012). Macroeconomics with Financial Frictions: A Survey. Working Paper 18102, National Bureau of Economic Research.

Brunnermeier, M. K. and L. H. Pedersen (2009). Market Liquidity and Funding Liquidity. *Review of Financial Studies* 22.6, pp. 2201–2238.

Brunnermeier, M. K. and Y. Sannikov (2014). A Macroeconomic Model with a Financial Sector. *American Economic Review* 104.2, pp. 379–421.

Campbell, J. Y. and S. B. Thompson (2008). Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *The Review of Financial Studies* 21.4, pp. 1509–1531.

Chabakauri, G. (2013). Dynamic Equilibrium with Two Stocks, Heterogeneous Investors, and Portfolio Constraints. *The Review of Financial Studies* 26.12, pp. 3104–3141.

Cochrane, J. H. (2005). *Asset pricing*. Princeton university press Princeton, NJ.

Coimbra, N. and H. Rey (2017). Financial Cycles with Heterogeneous Intermediaries. Working Paper 23245, National Bureau of Economic Research.

Danielsson, J., H. S. Shin, and J.-P. Zigrand (2012). Procyclical Leverage and Endogenous Risk. Working Paper.

Di Tella, S. (2017). Uncertainty Shocks and Balance Sheet Recessions. *Journal of Political Economy* 125.6, pp. 2038–2081.

Drechsler, I., A. Savov, and P. Schnabl (2018). A Model of Monetary Policy and Risk Premia. *The Journal of Finance* 73.1, pp. 317–373.

Duffie, D. and P.-L. Lions (1992). PDE Solutions of Stochastic Differential Utility. *Journal of Mathematical Economics* 21.6, pp. 577–606.

Duffie, D. and L. G. Epstein (1992). Stochastic Differential Utility. *Econometrica* 60.2, pp. 353–394.

Dumas, B. (1989). Two-Person Dynamic Equilibrium in the Capital Market. *Review of Financial Studies* 2.2, pp. 157–188.

Epstein, L. G. and S. E. Zin (1989). Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework. *Econometrica* 57.4, pp. 937–969.

Fama, E. F. and J. D. MacBeth (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy* 81.3, pp. 607–636.

Gârleanu, N. and S. Panageas (2015). Young, Old, Conservative, and Bold: The Implications of Heterogeneity and Finite Lives for Asset Pricing. *Journal of Political Economy* 123.3, pp. 670–685.

Gârleanu, N. and L. H. Pedersen (2011). Margin-based Asset Pricing and Deviations from the Law of One Price. *The Review of Financial Studies* 24.6, pp. 1980–2022.

Gertler, M. and N. Kiyotaki (2010). "Chapter 11 - Financial Intermediation and Credit Policy in Business Cycle Analysis". Ed. by B. M. Friedman and M. Woodford. Vol. 3. Handbook of Monetary Economics. Elsevier, pp. 547–599.

Gertler, M., N. Kiyotaki, and A. Prestipino (2016). "Chapter 16 - Wholesale Banking and Bank Runs in Macroeconomic Modeling of Financial Crises". Ed. by J. B. Taylor and H. Uhlig. Vol. 2. Handbook of Macroeconomics. Elsevier, pp. 1345–1425.

Giglio, S., B. Kelly, and S. Pruitt (2016). Systemic Risk and the Macroeconomy: An Empirical Evaluation. *Journal of Financial Economics* 119.3, pp. 457–471.

Gorton, G. and A. Metrick (2012). Securitized Banking and the Run on Repo. *Journal of Financial Economics* 104.3, pp. 425–451.

Goyal, A. and I. Welch (2008). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies* 21.4, pp. 1455–1508.

Haddad, V. and T. Muir (2018). Do Intermediaries Matter for Aggregate Asset Prices? Working Paper.

Hansen, L. P., J. C. Heaton, and N. Li (2008). Consumption Strikes Back? Measuring Long-Run Risk. *Journal of Political Economy* 116.2, pp. 260–302.

He, Z., B. Kelly, and A. Manela (2017). Intermediary Asset Pricing: New Evidence from Many Asset Classes. *Journal of Financial Economics* 126.1, pp. 1–35.

He, Z., I. G. Khang, and A. Krishnamurthy (2010). Balance Sheet Adjustments During the 2008 Crisis. *IMF Economic Review* 58.1, pp. 118–156.

He, Z. and A. Krishnamurthy (2012). A Model of Capital and Crises. *The Review of Economic Studies* 79.2, pp. 735–777.

He, Z. and A. Krishnamurthy (2013). Intermediary Asset Pricing. *American Economic Review* 103.2, pp. 732–70.

Hodrick, R. J. (1992). Dividend Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement. *The Review of Financial Studies* 5.3, pp. 357–386.

Judd, K. L. (1992). Projection Methods for Solving Aggregate Growth Models. *Journal of Economic Theory* 58.2, pp. 410–452.

Judd, K. L. (1998). *Numerical Methods in Economics*. MIT Press.

Kiyotaki, N. and J. Moore (1997). Credit Cycles. *Journal of Political Economy* 105.2, pp. 211–248.

Kogan, L., I. Makarov, and R. Uppal (2007). The Equity Risk Premium and the Riskfree Rate in an Economy with Borrowing Constraints. *Mathematics and Financial Economics* 1.1, pp. 1–19.

Lettau, M. and S. Ludvigson (2001). Consumption, Aggregate Wealth, and Expected Stock Returns. *The Journal of Finance* 56.3, pp. 815–849.

Lewellen, J., S. Nagel, and J. Shanken (2010). A Skeptical Appraisal of Asset Pricing Tests. *Journal of Financial Economics* 96.2, pp. 175–194.

Longstaff, F. A. and J. Wang (2012). Asset Pricing and the Credit Market. *Review of Financial Studies* 25.11, pp. 3169–3215.

Ma, S. (2017). Heterogeneous Intermediaries and Asset Prices. Working Paper.

Merton, R. C. (1973). An Intertemporal Capital Asset Pricing Model. *Econometrica* 41.5, pp. 867–887.

Moreira, A. and A. Savov (2017). The Macroeconomics of Shadow Banking. *The Journal of Finance* 72.6, pp. 2381–2432.

Nozawa, Y. (2017). What Drives the Cross-Section of Credit Spreads?: A Variance Decomposition Approach. *The Journal of Finance* 72.5, pp. 2045–2072.

Protter, P. E. (2004). *Stochastic Integration and Differential Equations*. 2nd ed. Springer-Verlag.

Rytchkov, O. (2014). Asset Pricing with Dynamic Margin Constraints. *The Journal of Finance* 69.1, pp. 405–452.

Santos, T. and P. Veronesi (2016). Habits and Leverage. Working Paper 22905, National Bureau of Economic Research.

Schwert, G. W. (1989). Why Does Stock Market Volatility Change Over Time? *The Journal of Finance* 44.5, pp. 1115–1153.

Shanken, J. (1992). On the Estimation of Beta-Pricing Models. *The Review of Financial Studies* 5.1, pp. 1–33.

Stambaugh, R. F. (1999). Predictive Regressions. *Journal of Financial Economics* 54.3, pp. 375–421.

Wang, J. (1996). The Term Structure of Interest Rates in a Pure Exchange Economy with Heterogeneous Investors. *Journal of Financial Economics* 41.1, pp. 75–110.

CHAPTER 2

Student Loans, Marginal Costs, and Markups: Estimates from the PLUS Program

(with William Mann)

The growth of student loans to finance college attendance, and the effects of that debt burden on later life decisions, have led to great interest in how financial aid shapes the market for higher education.[1] A growing body of research supports the "Bennett Hypothesis" that financial aid leads to higher college prices (e.g. Lucca, Nadauld, and Shen (2018)). The usual interpretation is that colleges, not students, are the main beneficiaries of financial aid. However, this interpretation implicitly assumes significant market power in higher education, which to date has not been shown. In this work, we provide the first empirical evidence that market power is the underlying mechanism behind the Bennett Hypothesis.

To motivate our analysis, we present novel facts about private colleges' expenditures, which are not as frequently studied as their enrollment and tuition. While tuition revenues have grown in recent years, colleges have spent those revenues almost as quickly, such that average instruction-related spending per student is close to average tuition revenue throughout the years in our sample. This paints a very different picture than the usual interpretation of the Bennett Hypothesis: If the cost of enrolling a student is close to the price that student pays, then colleges do not unduly benefit from financial aid. In fact, the Bennett Hypothesis in this setting could imply that financial aid inefficiently benefits *students*, not colleges.[2]

[1]Student debt outstanding totaled $1.34 trillion as of the first quarter of 2017 (Federal Reserve Bank of New York, Research and Statistics Group (2017)). For examples of the real effects of student loans on household investment, see Fos, Liberman, and Yannelis (2017) on postgraduate education, and Krishnan and Wang (2018) on entrepreneurship.

[2]This situation is known as advantageous selection. Because market power and advantageous selection

However, these figures only capture average costs. To assess market power, one requires estimates of marginal costs. The main focus of this chapter is to estimate marginal costs at private colleges using an empirical strategy that addresses the simultaneity of enrollment and expenditures. This approach yields a very different conclusion: Marginal costs are much smaller than tuition charges at the typical private college. Our finding supports the common (often implicit) view that the Bennett Hypothesis arises from market power, not advantageous selection. On the other hand, it contrasts the only prior evidence on markups in higher education, from Epple, Romano, and Sieg (2006).

Our strategy is to study an inward demand shock that occurred due to the abrupt removal of an important loan program from many students in 2011. The program in question is the Parent Loans to Undergraduate Students (PLUS) program, which has been largely overlooked in academic research on student aid, even though it accounts for over 15% of federal student borrowing.[3] Unlike the more widely-studied Stafford loan programs, the PLUS program imposes effectively no borrowing limit.[4] This unique feature makes the PLUS program a deep source of tuition revenue, and a potentially powerful setting to observe effects of aid availability on college pricing. We show that only the PLUS program has exhibited growth in average loan amounts in recent years, matching the 4% average annual tuition increase in our sample of private colleges.

In October 2011, the Department of Education completed the consolidation of the Federal Family Education Loan (FFEL) program with the newer Direct Loan (DL) program. It discovered that PLUS loans under the two programs had used different definitions of "adverse credit history," a disqualifying condition. Needing to harmonize the language, the Department chose the stricter FFEL definition. This redefinition was expected to have min-

---

can both generate pass-through effects like the Bennett Hypothesis, one requires further analysis (as in our study) to disentangle them. In a related example, Cabral, Geruso, and Mahoney (2018) document subsidy pass-through effects in the Medicare Advantage program, then disentangle market power from advantageous selection as the explanation.

[3]To our knowledge, the only other paper on PLUS loans is Bhole (2017), who shows that graduate PLUS loans crowd out the private market. Our analysis focuses on undergraduate PLUS loans.

[4]The only limit is the cost of attendance, which is at the college's discretion.

imal real effect, but in fact it disqualified many parents who previously would have been able to borrow. Analysts have suggested that PLUS denial rates doubled, and many colleges blamed the change for enrollment declines, revenue shortfalls, and budget cuts.

To validate the 2011 credit standards tightening as an inward demand shock, we obtain Department of Education data on enrollments, financial aid, and expenditures for private US colleges offering federal loans. We calculate the fraction of financial aid recipients that came from low-income families as of 2010, and use this fraction as our cross-sectional measure of treatment exposure. After the standards tightening, PLUS loan volume fell by about 31% in relative terms for private colleges with an above-median fraction of low-income students. Enrollment and tuition revenue also fell, confirming that the standards tightening represented an inward demand shift for enrollment at these schools.

We find evidence of the Bennett Hypothesis in response to this shock: The growth rate in tuition charges slowed, in relative terms, at colleges more affected by the standards tightening. This evidence implies that the PLUS program increased the price paid to attend affected colleges. Our finding complements prior studies, in that we study a decrease rather than an increase in financial aid. It also holds both for advertised tuition and for average tuition revenue per student (which accounts for tuition discounts and scholarships).

The main goal of our analysis is to investigate the underlying causes of the Bennett Hypothesis. We do this by estimating the marginal costs of enrolling the affected students relative to the prices they paid. To that end, we exploit data on colleges' total expenditures. We regress total expenditures on enrollment, instrumenting enrollment with the tightening of credit standards for PLUS loans. This yields an instrumental-variables estimate of marginal cost. Under perfect competition, even a non-profit college would not be able to charge above marginal cost, so a positive markup is evidence of market power.

As expected, we find that the standards tightening caused a decrease in college spending, with the largest effects in the instructional expense category. However, the implied mean (median) marginal cost per student was only \$8,700 (\$5,300). This is far below advertised tuition and fees of \$21,800 (\$20,600), and also below the effective price of \$15,500 (\$11,800),

defined as the price net of scholarships and discounts. We thus estimate that the average Lerner index for disqualified students was 0.7, far above the typical values even for high-end luxury goods.[5] We conclude that markups were large at private colleges in our sample.

We next explore the cross-sectional distribution of estimated markups to refine this argument. First, we predict that for-profit schools take greater advantage of their pricing power than do non-profits, and therefore have larger markups, due to their explicitly profit-maximizing objective function. Second, we predict that the presence of public colleges should be associated with smaller markups at private colleges, as this competition should erode private colleges' market power. We find evidence consistent with both implications: Markups are higher at for-profit schools, and lower in states with a high number of public schools.

In sum, we find that the pass-through effect captured by the Bennett Hypothesis is explained by market power at incumbent colleges. This conclusion is important for several reasons. First, one cannot analyze welfare effects of student loans based only on the Bennett Hypothesis, without supporting evidence such as ours on its underlying causes. As we describe in the next section, a subsidy pass-through effect, on its own, is equally consistent both with college enrollment being inefficiently low, and with its being inefficiently high. Second, our conclusion has not been documented before. Indeed, as we explain in the next section, it contrasts the only prior evidence on markups in higher education. Third, our cross-sectional analysis provides guidance as to where the intended effects of financial aid are most likely to be blunted by market power.

Our work contributes to a growing literature on the real effects of student loans. Several prior studies document the Bennett Hypothesis at work, mostly using the Stafford or Pell Grant programs. For example, Lucca, Nadauld, and Shen (2018) find that expansions of these programs led to higher tuition charges, and L. J. Turner (2017) and Singell and Stone (2007) find that a large fraction of Pell dollars is captured through higher tuition or smaller discounts. Beyond these specific programs, Long (2004) finds that colleges captured about 30% of the Georgia HOPE scholarship, and N. Turner (2012) find larger effects for education

[5]The Lerner index is a measure of market power, defined as the markup over marginal cost as a fraction of price. The highest value estimated in a sample of cars in Berry, Levinsohn, and Pakes (1995) is 0.3.

tax credits.

An important strand of this literature studies for-profit colleges specifically: Cellini and Goldin (2014) report that they charge more for the same degrees when eligible for federal student aid. Eaton, Howell, and Yannelis (2018) find that they achieve greater profits by scaling up after being acquired by private-equity funds, consistent with our finding that they take advantage of their market power by charging very large markups.

Other papers take a structural approach. Epple, Romano, and Sieg (2006) estimate a model of college admissions, and find markups of less than 3%, a much smaller magnitude than ours. Gordon and Hedlund (2016) use the same framework to argue that colleges face large fixed costs. Other structural papers include Fu (2014), Lau (2015), and Fillmore (2016).

Because the tightening of PLUS credit standards involved a redefinition of the term "adverse credit history," our study also contributes to a growing literature in household finance documenting the real impact of credit histories. Examples include González-Uribe and Osorio (2014), Bos and Nakamura (2014), Bos, Breza, and Liberman (2018), Liberman (2016), Garmaise and Natividad (2017), and Musto (2004).

However, we cannot assess the optimality of a student's decision to borrow (see Card (2001) for a review of the returns-to-schooling literature), nor the impact of student loans on later life outcomes (see Cox (2017) for a study of refinancing behavior, and Fos, Liberman, and Yannelis (2017) and Krishnan and Wang (2018) for evidence that student loans constrain later investment).

2.1 Expenditures and costs in higher education

Despite widespread attention to the revenues colleges obtain from tuition and financial aid, there is little analysis of how colleges spend those revenues. This absence is surprising, as one cannot fully understand the implications of findings like the Bennett Hypothesis without characterizing quantities like markups, which requires data on expenditures and costs. In this section, we use IPEDS data to summarize recent patterns in these quantities.

In Figure 2.1, we plot aggregate revenues and expenditures at private colleges in the United States. The black line shows that aggregate revenues in higher education have trended upwards over time, although the series exhibits a large degree of volatility due to the importance of endowment income for a few colleges that have very large endowments. The dashed line shows that colleges do not simply save these funds, but rather spend them at the same pace: Aggregate expenditures have risen at roughly the same rate as aggregate revenues. Moreover, they appear to be rather inflexible, rising at a stable pace even when they far exceed aggregate revenues during the downturn of 2007-2008.

Total revenues includes endowment income, which is not closely tied to tuition charges or financial aid. Similarly, total expenditures includes research-related expenditures, which are not closely tied to student education. These distinctions are especially important at the colleges with the largest operating budgets in the sample. To focus more closely on the issues raised by the Bennett Hypothesis, the lower two lines in Figure 2.1 plot aggregate tuition and fee revenue, and aggregate student-related expenses (defined as "instructional" plus "student service" expenses). These categories are more likely to respond to financial aid increases.

The upward trends in tuition revenue and student expenses are both quite stable compared to the total revenue series at the top of the figure. Moreover, they do not immediately suggest that colleges are pocketing large amounts of surplus from the growth in financial aid. While aggregate tuition revenue grows steadily throughout the sample period, consistent with prior research, spending on instruction and student services rises almost as quickly.

Of course, the time period covered in the figure has also seen a large expansion in enrollment. To provide a more precise view of these patterns, Figure 2.2 plots the same information on a per-student basis: For each college and award year, we scale the values of the series in Figure 2.1 by that college's full-time-equivalent (FTE) student body.[6] We winsorize these ratios at the 1% and 99% level in the pooled panel, then average them for each award year in the sample, and plot the resulting averages in Figure 2.2.

[6]A full-time equivalent is a standard enrollment measure, defined as 30 credit-hours of instruction.

Average tuition revenue per FTE rises from about $12,000 to about $17,000 from the beginning to the end of the sample, an increase of roughly 3% per year. This increase is consistent with the Bennett Hypothesis, and evidence like this has been widely examined in the literature on higher education. However, average expenditures on instruction and student services nearly keep pace with this increase, a pattern that has not been as widely appreciated. Towards the beginning of the sample they are nearly equal, and towards the end the gap between them is only about $1000, or 6% of average tuition revenue per student.

Based on these patterns, we conclude that the usual interpretation of the Bennett Hypothesis – that it reflects an inefficient distortion away from students and towards colleges – is not immediately obvious. Formally, this conclusion can only hold if the market for higher education is fairly uncompetitive, such that colleges can set large markups (can charge far above the marginal cost of enrolling students). Based on Figure 2.2, one might instead be led to conclude that college has gotten more expensive simply because educating students has gotten more expensive.

This would entirely change the interpretation of the Bennett Hypothesis: If anything, it would suggest that colleges enroll too *many* students, not too few, following a pattern of advantageous selection. In fact, this would be consistent with the only existing empirical analysis on markups in higher education of which we are aware: Epple, Romano, and Sieg (2006) estimate a model of higher education and find that markups are only 2% to 3% of tuition charges, suggesting that higher education is actually quite competitive. In their counterfactual analysis, a small change in price leads to a dramatic reshuffling of students across schools, consistent with the effects of advantageous selection.

However, the structural estimation in that prior paper was not primarily focused on identification of the cost function. Its cost parameters came from an OLS regression of expenditures on a quadratic function of enrollment. Both that approach, and our simpler plots in Figure 2.2, suffer from the inability to distinguish *average* from *marginal* costs, due to the simultaneity of quantities and costs.

In short, data on college expenditures is helpful but insufficient to disentangle the com-

peting explanations behind the Bennett Hypothesis. The second necessary ingredient is an empirical strategy that can convincingly identify the marginal costs of enrolling students in college. Starting in the next section, we describe such a strategy, then use its results to shed light on the issues raised in this section.

2.2 Demand shock: The PLUS standards tightening

2.2.1 The PLUS program

The Parent Loans for Undergraduate Students (PLUS) program provides loans to parents of dependent undergraduate students who are unable to cover the cost of college attendance even after exhausting other forms of federal aid. Approximately 13% of parents of full-time dependent undergraduate students have taken out PLUS loans, with an average amount of $13,000 a year (Johnson, Bruch, and Gill (2015)). Figure 2.3 shows that aggregate borrowing has been roughly constant in dollar terms, and increasing as a fraction of total federal loan borrowing, representing over 15% by the end of the 2014–2015 academic year.

Uniquely among federal student loan programs, the PLUS program imposes no aggregate borrowing limit, and the only annual borrowing limit is the cost of attendance reported by the institution, less any other financial aid received.[7] Thus, there is effectively no limit on the *intensive* margin of credit for parents who qualify for PLUS loans. For this reason, the PLUS program is a particularly likely mechanism for student aid subsidies to affect college pricing. Figure 2.4 shows that the average amount borrowed under the PLUS program (conditional on positive) increased steadily during 2010–2014, while average borrowing under subsidized and unsubsidized Stafford programs stayed roughly constant at these programs' borrowing limits. The growth rate in aggregate borrowing in the figure matches the 4% average growth rate of tuition charges in our sample of private colleges during the same time period.

[7]See Fishman (2014). In contrast, Stafford subsidized and unsubsidized loans to dependent undergraduate students have (combined) annual borrowing limits ranging between $5,500 to $7,500, and aggregate limits of $31,000. Students whose parents fail to qualify for PLUS loans are eligible for an additional $4,000 to $5,000 of annual unsubsidized borrowing, and an additional $26,500 in aggregate borrowing, but still cannot automatically borrow up to the full cost of attendance as with PLUS loans.

However, and again unlike other forms of aid, PLUS loans are limited along the *extensive* margin. Despite being non-dischargeable in bankruptcy (like other student loans), PLUS loans are in the parents' names and cannot be transferred to the student beneficiary. For this reason, PLUS loans are not made to parents with an "adverse credit history." Our natural experiment centers on a redefinition of this term that happened in 2011.

2.2.2 The 2011 tightening of PLUS credit standards

The redefinition stemmed from the 2010 consolidation of the older Federal Family Education Loan (FFEL) program with the newer Direct Loan (DL) program. In October 2011 the Department of Education began applying for all PLUS borrowers a strict credit standard that had previously applied only under FFEL. The main difference was an expanded definition of the term "adverse credit history" that included having unpaid debt in collection, or having written-off student loans in the previous five years. The change took effect without any public announcement from the Department of Education (Nelson (2012)).

A *Washington Post* article described the effects:

> At the time, federal officials considered the matter routine. They tightened the screening process for loan applications to ensure that certain kinds of unpaid debts were considered in a review of a parent's credit record. That made it more likely that some applicants would be deemed to have an "adverse credit history" and therefore ineligible.
>
> Acting Deputy Education Secretary Jim Shelton said the action was taken by "middle management" officials in an effort to fix what they saw as "a glitch in the system." He said that top officials did not review the decision before it was implemented, but that the department stood by it as consistent with laws and regulations.
>
> Anderson (2013)

Data support the view that this was a significant event: Using data from the 2007–2008 National Postsecondary Student Aid Study (NPSAS), Kantrowitz (2009) suggests that prior to consolidation, PLUS loan denial rates were 42% in the FFEL program but only 21% in the DL program. Goldrick-Rab, Kelchen, and Houle (2014) suggest that PLUS denial rates jumped from 22% to 42% between 2010 and 2012. Media coverage focused especially

on historically black colleges and universities (HBCUs), which feature high rates of PLUS borrowing and have little endowment to cushion the blow of lost tuition.[8]

However, these effects varied greatly across educational institutions. Some were unaffected by the change while others faced severe problems. Our empirical strategy is to combine variation in pre-treatment exposure to the standards tightening with panel data on college charges and enrollments.

2.3 Implementation

2.3.1 Data

The data sources that enable our analysis are both publicly available, and are the standard sources for empirical work on student loan programs.

Our first data source is the annual Title IV Program Volume Reports, published by the Department of Education. The name refers to Title IV of the Higher Education Act of 1965, which authorizes federal student aid programs. The Volume Reports provide the number of recipients and loan origination volume for each school participating in the Title IV aid, separated by loan program.

Our second data source is the Integrated Postsecondary Education Data System (IPEDS), also published by the Department of Education. Every college whose students are eligible for Title IV aid must annually complete a series of surveys that are published on IPEDS. The survey components include institutional characteristics, enrollment, student financial aid, and school finances, among others.

[8]The UNCF reported that the number of PLUS borrowers at HBCUs dropped by 45% from 2011–2012 to 2012–2013, removing $155 million from college budgets. Howard University blamed the standards change for a 6 percent drop in enrollment, a $17 million decline in revenues, and a credit downgrade from Moody's, while Morehouse College cut its budget by $2.5 million and laid off 50 employees. See "The Parent PLUS Loan Crisis" on the UNCF website (link); "Change to PLUS loan program hits HBCUs hard," aired on APM Marketplace October 28, 2014 and available at http://www.marketplace.org/; and Johnson, Bruch, and Gill (2015). For-profit colleges also feature high rates of PLUS borrowing that dropped sharply following the standards tightening (Fishman (2014)).

Appendix 2.7 discusses the details of merging the IPEDS and Title IV datasets. We note one detail here: IPEDS data are reported separately for each physical *campus*, but the Title IV reports are only available at the coarser level of the college *system*. To merge the datasets, we aggregate the IPEDS data across all campuses within a college system. This is largely a technicality, as 90% of college systems in the sample have only one campus. However, strictly speaking, all our references to "colleges" or "schools" (which we use interchangeably) are actually to college systems.

Both datasets are timed by award year. The Department of Education defines an award year as the school year for which financial aid is used to fund a student's education. Generally, this is the 12-month period that begins on July 1 of a given year and ends on June 30 of the following year. Hence, in all of the following analysis, year t is defined as the award year beginning on July 1 of calendar year t and ending on June 30 of calendar year $t + 1$.

We restrict the sample to private schools, in order to analyze the effects of demand shocks on pricing and quantities. The objective function and market structure for public schools are likely be very different from private schools, so we set them aside for the main analysis. We furthermore require that each school is present in the Title IV data (to observe their usage of federal aid programs), as well as having non-missing data on enrollment (to quantify the demand shock), expenses (to estimate marginal costs), and advertised tuition and realized tuition revenue (to calculate markups). Finally, we also retain only schools that, after applying the above filters, appear in the data for at least the 2010, 2011, and 2012 award years (one year before and after the standards tightening).

2.3.2 Cross-sectional treatment intensity

To examine the real impact of the 2011 PLUS standards tightening, we sort colleges according to a proxy for the expected intensity of the effects of this event. Ideally, we would like to know the ex-ante fraction of students at each college who relied on PLUS loans to finance their attendance, and whose parents would have qualified under the old credit history regime but not under the new one. This would require a student-level match between college choice,

financial aid, and parents' credit histories, which to our knowledge does not exist anywhere—even in confidential or administrative data at colleges or government agencies.

Instead, we construct a college-level proxy for student credit constraints based on the IPEDS data. Our proxy is a ratio, $frac\_low\_income$, that measures the fraction of students at the school who come from low-income backgrounds. We interpret this as a proxy for the importance of financial constraints to an individual student, as low-income students are less likely to have resources or credit to finance their eduction outside of the PLUS program.

To construct $frac\_low\_income$, we consult the campus-specific IPEDS Student Financial Aid form, which is available between 2008 and 2013. This form reports the number of newly-enrolled undergraduates that received any federal student aid under Title IV, referred to as "Group 4," and further breaks down this number into brackets of annual household income, of which the lowest two brackets are \$0-\$30,0000 and \$30,000-\$48,000. For each campus, we record both the total number of Group 4 students, and the number that come from these two brackets. We sum both these numbers across each campus in a college system, then calculate their ratio as our value of $frac\_low\_income$.

Our empirical approach is to see if a high value of this ratio as of 2010 – before the credit standards tightening – predicts a relative decline in PLUS loan usage and enrollment thereafter, controlling for school fixed effects. In the motivating figures, we will divide the sample into two bins by values of $frac\_low\_income_{i,2010}$, so this is a difference-in-difference, where the above-median schools are referred to as "treated" and the below-median as "untreated" (really, more-treated and less-treated). In the instrumental-variables estimation of marginal costs and markups, we will exploit continuous variation in $frac\_low\_income_{i,2010}$.

2.3.3 Characteristics of sample schools

Table 2.1 displays summary statistics for private schools in our sample as of the 2010–2011 award year. For the average school in our sample in 2010, following the Title IV data, total PLUS loan volume was \$3.5 million spread among 242 borrowers, implying an average borrowing of \$14,000, close to the \$13,000 reported in Johnson, Bruch, and Gill (2015).

Enrollment is measured in multiple ways in the IPEDS data. The measure on which we will focus is twelve-month undergraduate headcount, which is defined as the number of distinct undergraduates who attended during a twelve-month period. Our conclusions are also robust to an alternative measure, undergraduate full-time equivalents (UFTEs), which are defined as total credit hours of instruction to undergraduates, divided by 30 (to reflect the standard 30 credit hours in a full-time load). The first measure has a physical interpretation as a number of students, while the second approximates the quantity of education "sold." The sample averages of these enrollment measures as of 2010 were 3,158 and 2,624 respectively.

Average advertised tuition and fees in our sample in 2010 were $21,000. However, realized tuition and fee revenue per student was significantly lower, at $15,000, which reflects institutional scholarships and tuition discounts for individual students compared to the advertised price. We will focus on the smaller value, realized revenue per student, as our concept of price, which is a conservative approach for our purposes as it leads to smaller estimated markups. Total tuition and fee revenue was $19 million for the median school in the sample.

Finally, we note that the typical private college is not wealthy and selective, as is sometimes assumed in discussions of higher education. At the median in our sample, the ratio of endowment dollars to students is only $5,800, and revenue from investment income constitutes only 5% of total revenues. On the other hand, 48% of tuition and fee revenue comes from federal student loan originations. These figures show that most private colleges rely heavily on federal loan programs to supply revenue from tuition and fees, so that a decrease in loan availability should greatly affect their operations, as we will demonstrate. The extreme example is for-profit colleges, which represent 23% of our sample. At the median for-profit, two-thirds of tuition and fee revenue came from federal loan originations in 2010.

2.4 Impact of the standards tightening

2.4.1 Graphical analysis

First, we demonstrate visually how the standards tightening constituted a demand shock for exposed schools. To do this, we sort schools into two bins of 2010 treatment intensity, according to whether a school's value of $frac\_low\_income_{i,2010}$ was above or below the median value for schools in the same state. We create the cutoffs within-state as the purchasing power of the income cutoff in $frac\_low\_income_{i,2010}$ can be very different across regions.

For the graphical analysis in this section, we refer to the above-median subsample as "treated" and the below-median as "untreated," and we track these subsamples based on several outcome variables during 2009–2013. This is analogous to the difference-in-difference approach taken, for example, in Havnes and Mogstad (2011) in studying the effect of subsidized childcare on maternal employment in Norway.

We first confirm that the standards tightening indeed reduced PLUS loan usage. Figure 2.5 tracks the average log number of PLUS loan recipients per college across the treated and untreated subsamples, relative to each subsample's 2010 levels. Prior to 2011, the two subsamples followed the same upward trend. But in 2011, the average log number of recipients in the treated subsample fell sharply, and a large disparity opened up between the two subsamples thereafter.

Since the goal of financial aid is to promote access to college, one would expect a removal of that aid to decrease attendance. We show that this was the case: In Figure 2.6, the outcome variable is the log of the school's undergraduate headcount, again shifted by the 2010 mean value for each subsample. Figure 2.6 shows that enrollment, like PLUS loan activity in the previous figures, exhibited a steep differential drop across the more- and less-exposed colleges beginning in 2011 and persisting thereafter. The figure employs both of the alternative measures of enrollment described in the previous section. Figure 2.7 further shows that this translated into a substantial decrease in tuition and fee revenue at the treated schools.

The Bennett Hypothesis claims that student loan programs increase not only attendance, but also tuition. If so, the removal of PLUS loans should have led to a decrease in tuition. Figure 2.8 demonstrates exactly this result. The figure plots realized tuition and fee revenue divided by undergraduate headcount. This measures the average revenue the school actually extracts per student, accounting for tuition discounts on advertised tuition and fee charges. The figure demonstrates that average tuition revenue grew at a similar rate in the treated and untreated subsamples pre-2010. However, between 2010 and 2012, average tuition revenue grew much more slowly at the treated schools. Towards the end of the sample, the growth rates converge again, but the disparity in average tuition revenue persists.

Taken at face value, these results suggest that financial aid programs do indeed allow colleges to charge higher prices to students. However, there is a subtle distinction between these findings and the Bennett Hypothesis: Our results do not establish that any *individual* student paid a higher price due to the existence of the PLUS program. Due to tuition discounts, in principle every student pays a specific price, not necessarily equal to the advertised price nor to the average realized price. It could be that the students who failed to enroll due to the standards tightening would have paid the highest prices before, so the average price fell simply through a change in the composition of the student body.[9]

From a policy perspective, there may be nothing inefficient about subsidized loan programs in this case. Restricting financial aid would decrease the average realized price of college, but only by preventing attendance among students who pay the most; it would not benefit any inframarginal students who would enroll regardless. The same issue would apply to any empirical analysis of the Bennett Hypothesis that focuses only on price effects, in the absence of detailed data on actual individual charges. These observations call for a different approach to investigating colleges' pricing power in the presence of subsidized student loans.

Our focus is instead on estimating the marginal costs of enrolling the students who were affected by the policy, and the markups that they paid over marginal cost. Estimating cost

[9]For example, this could be the outcome of the models in Epple, Romano, and Sieg (2006) and related papers, in which schools offer discounts to some high-achieving students, in order to increase reputation and attract other students who pay higher rates.

functions is fundamental to an empirical analysis of any market, yet it has not received much attention in research on the market for higher education. The only prior evidence of which we are aware comes from Epple, Romano, and Sieg (2006) and Epple, Romano, Sarpça, and Sieg (2017), who estimate colleges' cost functions via OLS regressions of expenditures on a quadratic function of enrollment.

To estimate marginal costs, we introduce data from IPEDS on colleges' expenditures, and examine how it responded to the credit standards tightening. Following the motivation of this section, Figures 2.9 and 2.10 present trends in expenditures at the above- and below-median treatment-exposure schools around the date of the credit standards tightening. Figure 2.9 focuses on total expenses, and Figure 2.10 focuses only on the three largest categories.[10] The figures show that spending followed similar growth rates prior to 2011, but declined at treated schools in relative terms thereafter. However, the magnitude of this effect is noticeably smaller than the tuition revenue effect in Figure 2.7. This is the intuitive pattern that we will exploit in estimating markups.

We next turn to a quantitative analysis of all these findings.

2.4.2 Regression analysis

In this section, we quantify the above results in a regression framework. We start by presenting difference-in-difference results that correspond to the figures already presented, then move to specifications that exploit continuous variation in the treatment-exposure proxy $frac\_low\_income_{i,2010}$. We retain only the award years 2010 and 2012 in order to tighten the identification as much as possible around the credit standards tightening, while excluding 2011 itself, as the standards tightening happened midway through that award year.

We start by estimating the following difference-in-difference specification:

$$y_{it} = \alpha_i + \phi \times \text{After}_t + \gamma \times \text{Bin } 2/2_i \times \text{After}_t + \epsilon_{it} \tag{2.1}$$

[10] As summarized in Table 2.1, these are instructional, support, and research and service.

where i indexes schools, $t \in \{2010, 2012\}$, After_t is an indicator for $t = 2012$, and $\text{Bin } 2/2_i$ indicates that the school's value of $frac\_low\_income_{i,2010}$ was above the median for its state. The outcome variables y_{it} will be at the school level, and will include PLUS loan volumes, enrollment, revenues, and average revenue per student.

The causal effect of the credit standards tightening on y_{it} is captured by γ, granted the assumption that the subsamples would have followed parallel trends in the absence of the standards tightening, as suggested by the previous figures. School fixed effects are captured by α_i. We cluster standard errors in all tables by school system.

Table 2.2 reports regression results corresponding to Figures 2.5 and 2.6. In Columns 1 and 2, we establish the steep decline of PLUS usage in response to the standards tightening. Schools with above-median treatment intensity saw approximately a 31% decrease in both the number and volume of PLUS loans relative to those below-median. Columns 3 and 4 demonstrate, as in Figure 2.6, that undergraduate headcount and UFTEs fell by about 7% and 9% respectively. We will focus on measuring enrollment via headcount in our later analysis, but both measures yield qualitatively similar results.

Table 2.3 reports further regressions that capture reduced-form effects of the demand shock illustrated by Figures 2.7, 2.8, and 2.9. Column 1 demonstrates a decrease of roughly 5% in total tuition and fee revenue in response to the standards tightening. Column 2 shows that advertised tuition and fees per student fell by about $1370 as well, which is suggestive evidence of the Bennett Hypothesis effect. However, the magnitude may not be meaningful, as students do not necessarily pay the advertised tuition. Column 3 focuses instead on realized tuition revenue per student, and shows that this was lower by about $478 at schools more affected by the credit standards tightening. This represents about 4% of the median realized tuition and fee revenue per student documented in Table 2.1.

As discussed in the previous section, our main goal is to estimate colleges' marginal costs of enrolling the students affected by the standards tightening. Column 4 of Table 2.3 provides evidence in this direction: It shows that total spending fell by roughly 2.5% in response to the standards tightening. While precisely estimated (and significantly different from zero),

this magnitude is noticeably smaller than the effect in Column 1. This is the intuitive basis for our finding that colleges charged well above marginal cost.

Dividing the sample into two bins is useful for illustrative figures and examining parallel trends, but the variation in the standards-tightening instrument is potentially much richer. In Tables 2.4 and 2.5 we exploit this variation more fully, repeating the analysis of Tables 2.2 and 2.3, respectively. The top panel of each table splits the treatment intensity into four bins instead of two. The lower panel replaces the bin instruments with the continuous fraction of low-income students that was used to construct them. Across all specifications in the tables, the effects of the standards tightening on PLUS loan usage, enrollment, charges, revenues, and expenses are monotonically increasing in the fraction of low-income students at the college as of 2010. This is encouraging evidence that our instrument convincingly captures exposure to the demand shock represented by the standards tightening.

For the remainder of the chapter, we will instrument for enrollment using the continuous treatment proxy $frac\_low\_income_{i,2010}$, as in the lower panels of these two tables.

2.5 Estimating markups charged to PLUS recipients

2.5.1 Instrumental-variables estimates of costs and markups

Table 2.6 reports instrumental-variables estimation of the marginal effect of enrollment on colleges' spending and revenues. All the specifications are in logs to remove scale effects, and (endogenous) enrollment is measured via the twelve-month student headcount. The instruments for enrollment in these columns are the interaction between the 2012 indicator and the continuous treatment exposure measure $frac\_low\_income_{i,2010}$.

In Columns 1 and 2 of Table 2.6, we focus on quantifying marginal costs. Column 1 estimates the elasticity of total college spending with respect to enrollment: A 10% increase in enrollment would lead to a 2.8% increase in total spending. Table 2.1 reported a breakdown of total expenses across the five largest categories in the data. The largest category is instructional expense, at about one-third of the total. Column 2 shows that this expense

category mostly drives the proportional adjustment in spending captured by Column 1. Untabulated results show that all other categories exhibit adjustments that are economically small and statistically insigificant. As is intuitive, instructional spending seems to be the primary category of variable costs for colleges in the sample.

In Columns 3 and 4 of Table 2.6, we focus on quantifying marginal revenues, in order to compare this magnitude with the marginal costs estimated in the first two columns. Column 3 of Table 2.6 shows elasticity estimates of tuition and fee revenue with respect to enrollment. However, this does not perfectly capture the effective price paid by students, as not all student-level sources of revenue are reported with tuition and fees in IPEDS. For example, some government aid (such as Pell grants) and many types of fees (such as residence halls, meals, activities, and health services) are reported as revenue in other categories. Nevertheless, these would be included in the total expense figures in Column 1. Therefore, to be consistent, in Column 4 we report the elasticity of total revenue with respect to enrollment, and this effect will be our focus in comparing revenue and cost effects.

We can convert the coefficients from these log-log specifications to dollar terms as follows: For example, the instrumented marginal effect of headcount on total expenses, reported in Column 1, is 0.283. The mean (median) ratio of total expense to headcount in the sample as of 2010 was 29.8 (17.9). Multiplied together, these magnitudes imply a mean (median) marginal cost of \$8,400 (\$5,100), much lower than either the advertised tuition of \$21,000 (\$20,000), or realized tuition per student of \$15,000 (\$11,000), that were reported in Table 2.1. On the other hand, the mean (median) ratio of total revenues to headcount was 38.9 (20.3), so the coefficient in Column 4 implies a marginal effective price (including tuition, fees, and any other sources of college revenue) of \$31,400 (\$16,400). The discrepancy between these numbers is the essential reason that we estimate large markups in higher education.

We can think about these results in terms of the standard Lerner index, which is defined as the ratio of markup to price. This can be expressed $1 - \frac{MC}{P}$, where MC is marginal cost and P is price. We can write the ratio in this expression as $\frac{MC}{P} = \frac{dC/dQ}{dR/dQ} = \frac{C}{R} \times \frac{d\ln C/d\ln Q}{d\ln R/d\ln Q}$. The derivatives in these expressions refer to marginal costs and revenues, meaning the changes in costs and revenues induced by an exogenous change in enrollment— for example, by the

PLUS loans credit standards tightening that we study. We can therefore substitute in for these values the estimated coefficients from Columns 1 and 4 of Table 2.6. Finally, in the data, the median ratio $\frac{C}{R}$ as of 2010 was 0.86. Altogether, then, our estimates imply a median Lerner index of $1 - 0.86 \times \frac{.283}{.808} = 0.699$. For comparison, the Lerner index values that Berry, Levinsohn, and Pakes (1995) estimate for cars range from 0.16 for a Mazda 323 to 0.3 for a BMW 735i (see their Table VIII). Thus, the markups charged by private colleges to PLUS loan recipients seem to far exceed those attainable even for high-end luxury cars.

How should we interpret these findings? Intuitively, it may not seem surprising that the marginal cost of enrolling a student in college is far below the price he or she pays, but our estimated markups are much larger than the only prior evidence of which we are aware, from Epple, Romano, and Sieg (2006). One possible interpretation is that colleges can (and must) charge high markups to recover large fixed costs of operation. This would be a source of large barriers to entry in higher education. Next, we attempt to gain further insight into this interpretation.

2.5.2 Heterogeneity in estimated markups

In the final section of our analysis, we show that the identification strategy we pursue is rich enough to identify cross-sectional variation in private colleges' markups, yielding valuable insights about the determinants of colleges' pricing power.

To do this, we start by returning to the instrumental-variables regressions in Table 2.6. Suppose we modified the specification in Column 1 of that table by replacing log headcount with log total revenues as the endogenous variable. The resulting IV coefficient would be equal to the ratio of the coefficients in columns 1 and 4 of Table 2.6, which is $\frac{0.283}{0.808} = 0.35$. Recall that this cost-price ratio was the basis for our calculation of markups above. Therefore, we will use this modified specification as the basis for exploring heterogeneity in markups, adding interaction terms as appropriate to model that heterogeneity.

We also modify the previous specification by dropping the school fixed effects, in order to

mitigate the attenuation bias common to fixed-effects specifications.[11] When we estimated *average* markups in the previous table, we had sufficient statistical power to identify significant relationships even with these fixed effects. However, when we explore heterogeneity, power quickly becomes an issue. Therefore, in presenting the main results of this section in Table 2.7, we will replace the fixed effects with just the main effects of any cross-sectional explanatory variables in the model, and we will rely for identification on the difference-in-difference strategy across treated and untreated groups. In Table 2.8, we will further show that all the results are robust to including *state* fixed effects.

Our specification for estimating heterogeneity in markups is therefore:

$$\begin{aligned} \ln C_{it} \; &= \lambda_0 \times \ln R_{it} \; + \; \lambda_1 \times \ln R_{it} \times x_{i,2010} \\ &+ \; frac\_low\_income_{i,2010} \; + \; After_t + \; x_{i,2010} \; + \; \epsilon_{it} \end{aligned} \tag{2.2}$$

The cross-sectional $x_{i,2010}$ represents any dimension along which we examine heterogeneity in markups. We treat log total revenue, and its interaction with $x_{i,2010}$, as endogenous variables. Therefore, we also estimate a first-stage regression in which the terms $\ln R_{it}$ and $\ln R_{it} \times x_{i,2010}$ are replaced by the instruments $frac\_low\_income_{i,2010} \times \text{After}_t$ and $frac\_low\_income_{i,2010} \times \text{After}_t \times x_{i,2010}$.

Note that the first of these two instruments is the same as was used in Table 2.6, and the analysis here is equivalent to the approach in the prior sections if we drop $x_{i,2010}$. We report this benchmark case in the Column 1 of Table 2.7. In this specification, even without school fixed effects, we obtain the same cost-price ratio of 0.35 as we calculated earlier from the ratio of the coefficients in Columns 1 and 4 of Table 2.6.

As with Table 2.6, we note that this log-log coefficient is not equal to one minus the markup charged. The adjustment to this interpretation would involve multiplying by the ratio $\frac{C_{i,2010}}{R_{i,2010}}$, as in the previous section. Since this ratio is less than one for 90% of observations, markups should generally be larger than one minus the coefficients reported in the table,

[11] See Angrist and Pischke (2009), page 226.

which is again consistent with the earlier section. For simplicity, in discussing this table, we will focus only on the regression cofficients, without working through this translation.

Thus, the coefficient of interest in this section is λ_1, which captures how the cost-price ratio varies in the cross-section with $x_{i,2010}$, relative the average value of 0.35 that we calculated above based on the coefficients in Table 2.6. This cost-price ratio maps to (one minus) the Lerner index, following the intuition that was described before. In principle, any cross-sectional variable of interest could be used for $x_{i,2010}$. In Table 2.7, we focus on three choices that are particularly relevant to the market for higher education:

In the first specification, $x_{i,2010}$ is an indicator for being a for-profit school. The for-profit sector is known to rely on federal loan programs for revenue (e.g. Cellini and Goldin (2014)), and compared to non-profit schools, their objective function should make them particularly aggressive about taking advantage of any pricing power afforded by the availability of PLUS loans. We therefore expect $\lambda_1 < 0$, which would mean that for-profit schools have a lower cost-price ratio (and thus a higher Lerner index) than non-profit schools.

Column 2 in Table 2.7 shows that this is indeed the case. The first row shows that non-profit schools have a cost-price ratio of 0.844, whereas at for-profit schools this ratio is $0.844 - 0.535 = 0.309$. This result confirms that, where for-profit schools serve a greater fraction of the educational market, one should be more concerned about the possibility that the surplus created by a subsidized loan program accrues less to students and more to the incumbent colleges. This result is important in light of research showing that for-profit schools are a growing segment of the market for higher education,[12] but to our knowledge this is the first evidence quantifying the size of the markup in the for-profit education sector.

In our second specification, $x_{i,2010}$ is the number of bachelor's-granting public schools in the same state as a given school, calculated from the IPEDS data. Our hypothesis here is that the markup charged by a private school reflects, at least partially, a barrier to entry. Competition from public schools should then lower the markup that can be charged in equilibrium by a private school, so we expect that $\lambda_1 > 0$. If this effect is quantitatively large,

[12]See Cellini and N. Turner (2016) and Cellini and Goldin (2014), for example.

this suggests a natural policy—opening public schools and providing education directly—that can serve as a substitute for simply expanding loan programs, and that would increase students' share of the surplus from any subsidized loans that are provided.

This insight is explored in Column 3 of Table 2.7. We fix the number of public schools in the state as of 2010, and we also demean and standardize it across the sample. Thus, the coefficient in the first row, which corresponds to λ_0 in specification (2.2) above, can be interpreted as the cost-price ratio of the average school in the sample. This average is 0.364, close to the value of 0.35 obtained based on the prior tables.

The coefficient of 0.092 in the third row of Column 3 corresponds to λ_1 in specification (2.2). This magnitude shows that, relative to the sample average, private colleges in states with more public schools exhibit a higher cost-price ratio, which translates to a lower markup and Lerner index. Again, to our knowledge, this is the first quasi-experimental evidence that the presence of public schools can reduce the prices charged by public schools. This spillover effect constitutes a potential benefit of subsidizing higher education by opening public schools, as opposed to subsidized loans to attend incumbent private schools.

In the final specification in Table 2.7, we consider an alternative approach to capturing this spillover: Rather than the number of public schools in IPEDS, we use the 2010 higher-education appropriations (HEA) in the state government's budget.[13] This provides a continuous measure of the state's support for public schools, which should lower private-college markups and lead to $\lambda_1 > 0$ by the same mechanism as described above. As above, we fix HEA as of 2010, and demean and standardize it in the regression. Consistent with the prior tables, the average cost-price ratio is estimated to be 0.355 in Column 4. Meanwhile, the interaction term in the fourth row demonstrates, as in Column 3, that greater government support for higher education is associated with a higher ratio and therefore a lower markup.

As a robustness check for these findings, in Table 2.8, we repeat all the same specifications,

[13]Data on higher-education appropriations come from the Center on Budget and Policy Priorities (CBPP), as recorded in the annual Grapevine report from Illinois State University. HEA is not available for schools in the District of Columbia or Puerto Rico, so we drop these schools throughout Tables 2.7 and 2.8.

but add fixed effects for the state in which the school is located, and cluster all standard errors by state. The magnitudes of all the effects are virtually unchanged from Table 2.7, and all of them remain statistically and economically significant.

We draw several conclusions from the analysis in this section:

First, markups appear to be both large and heterogeneous in higher education. Heterogeneity in markups means that the incidence of a subsidy to consumers, and thus the optimal design of subsidies for college attendance, may also be different across different colleges and regions. This basic insight is important, in light of the fact that federal financial aid programs are currently implemented with the same rules for all colleges and students. A one-size-fits-all approach may have very different efficiency implications depending on the schools that are in the choice set of any individual student.

Second, for-profit schools appear able to extract a relatively larger margin from their attendees, suggesting that there are large barriers to entry even for this seemingly-less-selective segment of the higher education market. Possible sources of these barriers to entry include a track record of job placements, which would take time for an entrant to build; or geographic segmentation, if for-profit schools locate in areas without many other educational options. In any case, the findings lend support to concerns that the efficiency of student loans may be falling over time as for-profit schools constitute an increasing share of the education market.

Third, greater support for higher education from state government is associated with lower markups charged by private colleges in that state. This last finding suggests a novel policy implication: Where markups are large, so that students receive relatively little of the surplus from a subsidized loan program, opening a public school instead of expanding student aid may shift the market towards a more competitive structure. This approach would reflect the classic logic of Hotelling (1938), who pointed out that subsidizing fixed costs in industries with declining average costs would increase efficiency by driving prices down towards marginal cost. Opening a public school is effectively paying the fixed cost of a new entrant. In contrast, student loan subsidies scale with the size of college enrollment, thus constituting

a subsidy to variable instead of fixed costs. In the presence of large markups, such a subsidy mainly benefits producers (colleges), arguably leading to many of the inefficiencies identified in the modern empirical literature on the Bennett Hypothesis.

2.6 Conclusion

We present novel evidence for concerns about the incidence of student aid, based not on pass-through effects of aid programs on college tuition (which are the focus of prior papers), but rather on quasi-experimental estimates of low marginal costs and high markups. We exploit variation in aid programs as an identification strategy: Based on a 2011 tightening of credit standards for PLUS loans, we estimate that affected colleges faced relatively small marginal costs, and therefore their tuition charges reflected large markups. One interpretation of this finding is that the fixed costs of operating a college are large, which both enables and requires colleges to charge large markups and "recover" those fixed costs.

Moreover, we find important cross-sectional heterogeneity: Markups are higher at for-profit schools, and at schools in states with fewer public schools and less support for higher education. This heterogeneity demonstrates that financial aid, which is usually implemented as a flat subsidy to all consumers, can have heterogeneous impacts and incidence across different regions. Finally, the large average markup that we estimate suggests that a simple subsidy to consumers may not necessarily be the ideal financial aid design, and that policymakers should instead target barriers to competition in the form of large fixed costs.

Our results provide novel evidence for concerns about the incidence of student loan programs, such as the Bennett Hypothesis. Colleges seem to have, and to exploit, large pricing power, even outside of elite schools that are relatively unaffected by our natural experiment, and especially within the for-profit sector. Our results also suggest some settings where these concerns are likely to be greater, and potential alternative policies to subsidizing loan programs in the presence of these concerns.

2.7 Appendix 1: Aggregating and merging IPEDS and Title IV

The IPEDS institutional identifier is the UnitID variable. The Office of Postsecondary Education ID (OPEID) is the 8-digit unique identifier for the Title IV data. OPEID is also included in IPEDS. The first 6 digits of the OPEID variable identify a school system for financial aid purposes. Schools with OPEID ending in 00 are either independent entities or are the primary reporting institution for the Title IV program. For example, the primary reporting institution for the University of Washington system is the Seattle campus which has the OPEID of 00379800. The other two campuses, in Tacoma and Bothell, have OPEIDs that do not end in 00 but they share the same first 6 digit OPEIDs (003798) with the main Seattle campus. There are several instances where each individual campus of the same university system reports Title IV data separately. For example, each of the ten campuses in the University of California system has an OPEID ending in 00 and appears as a separate entity in the Title IV data.

Since the Title IV data is only available at the system level, we use the first 6 digits of the OPEID variable to merge IPEDS and Title IV datasets. In order to merge with Title IV, we also need to aggregate the institutional-level IPEDS data to create variables at the system level. For the indicator variables, we take the maximum across colleges in the same school system to create the aggregate variables. These variables indicate, for example, if the school is a HBCU, if it offers various degrees (bachelors, masters, doctorate, etc.), if it is public/private, for-profit/non-profit, etc. For variables such as tuitions and fees, grants, enrollment, school revenues and expenses, etc., we sum across the campuses within the same school system to generate the aggregate system-level variables. See the discussion in Kelchen (2014) for dealing with issues in merging IPEDS with Title IV.

The office of Federal Student Aid delivers aid to students through loan, grant, and work-study programs. The Title IV Program Volume Reports provide the number of recipients and volume by loan program for each school participating in the Title IV programs. The data is available annually from award year 1999–2000 to 2005–2006. From 2006–2007 award year onward, the data are reported quarterly and we use cumulative measures through Q4.

Prior to the award year 2009–2010, the Title IV data include information on both FFEL and DL programs. As mentioned above, after the elimination of the FFEL program in 2010, only the DL program is available to borrowers. Thus from award year 2010–2011 onward, the Title IV Reports only include the volume and number of recipients for the DL program. The annual data are available from award year 1999–2000 to 2005–2006.

2.8 Appendix 2: Tables and Figures

Table 2.1. Summary statistics

Summary statistics for sample school systems as of the 2010-2011 award year. Source: IPEDS/Title IV.

	Median	Mean	SD
PLUS Loan Volume ($1,000)	1365.21	3458.86	7301.12
PLUS Loan Recipients	134.00	241.71	416.55
Frac of aid recipients less than 48k$ income ($frac\_low\_income$)	0.49	0.55	0.26
Fraction of students receiving PLUS loans	0.10	0.11	0.09
Undergraduate 12-month headcount	1594.00	3158.55	7381.92
Undergraduate FTEs	1384.07	2624.83	5443.84
Advertised tuition and fees	20026.00	21017.88	9364.06
Total revenues ($1,000)	33369.30	136631.88	559414.14
Tuition and fee revenue ($1,000)	18990.93	46547.65	97831.82
Tuition and fee revenue per student	11426.29	14989.87	28607.18
Total expenses ($1,000)	27517.02	100324.65	378479.41
Instructional expense ($1,000)	9269.79	32886.63	115883.60
Support expense ($1,000)	13309.81	33490.52	86760.05
Research and service expense ($1,000)	0.00	11395.20	80908.30
Auxiliary enterprises expense ($1,000)	2955.71	9594.67	30354.85
Net grant aid expense ($1,000)	0.00	505.96	4666.83
Endowment dollars / headcount	5777.75	43473.30	166676.39
Investment income / total revenue	0.05	0.10	0.14
Tuition and fees / total revenue	0.60	0.60	0.25
Federal student loans / tuition and fees	0.48	0.47	0.23
For-profit	0.00	0.23	0.42
Federal loans / tuition and fees, for-profit only	0.66	0.65	0.19

Table 2.2. Effect of the credit standards tightening on PLUS loan usage and enrollment

This table reports OLS regression estimates of specification (2.1) for the effect of the credit standards tightening on PLUS loan usage and enrollment. Bin 2/2 is the above-median indicator for $frac\_low\_income_{i,2010}$. In Columns 1 and 2, the outcome variables are respectively the number of PLUS loan recipients, and the volume of PLUS loans disbursed, both in log terms and measured at the level of the school system. In Columns 3 and 4, the outcomes are respectively undergraduate 12-month headcount and full-time fall undergraduate enrollment. The sample includes the award years 2010 and 2012, for the 1,530 private colleges with nonmissing data for both of these years. All regressions include year and school system fixed effects, as described in specification (2.1) in the text. Standard errors are clustered by school system.

	(1) Ln(PLUS num)	(2) Ln(PLUS vol)	(3) Ln(Headcount)	(4) Ln(UFTEs)
Bin 2/2 × After	-0.311^{***}	-0.312^{***}	-0.0716^{***}	-0.0926^{***}
	(0.0263)	(0.0287)	(0.0133)	(0.0144)
School FE?	Yes	Yes	Yes	Yes
Year FE?	Yes	Yes	Yes	Yes
R^2	0.356	0.187	0.0240	0.0446
Obs.	2924	2924	3060	3060

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.3. Effect of the credit standards tightening on college revenue, pricing, and spending

This table reports OLS regression estimates of specification (2.1) for the effect of the credit standards tightening on college revenue, pricing, and spending. Bin 2/2 is the above-median indicator for $frac\_low\_income_{i,2010}$. In Column 1, the outcome variable is the log of total tuition and fee revenue. In Column 2, it is advertised undergraduate tuition and fees. In Column 3, it is realized tuition and fee revenue per unit of undergraduate headcount. Finally, in Column 4, it is the log of total spending. The sample includes the award years 2010 and 2012, for the 1,530 private colleges with nonmissing data for both of these years. All regressions include year and school system fixed effects, as described in specification (2.1) in the text. Standard errors are clustered by school system.

	(1) Ln(Tuition rev.)	(2) Tuition	(3) Rev/student	(4) Ln(Expenses)
Bin 2/2 × After	-0.0520^{***}	-1367.6^{***}	-478.4^{***}	-0.0253^{**}
	(0.0171)	(110.5)	(158.1)	(0.0110)
School FE?	Yes	Yes	Yes	Yes
Year FE?	Yes	Yes	Yes	Yes
R^2	0.0283	0.393	0.0761	0.0855
Obs.	3060	3060	3060	3060

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.4. Effect of the credit standards tightening on PLUS loan usage and enrollment: Alternative approaches to measuring exposure to the standards-tightening treatment

This table repeats the specifications of Table 2.2, but with alternative approaches to measuring exposure to the standards-tightening treatment. In Panel (a), exposure is measured with four separate bins of $frac\_low\_income_{i,2010}$, instead of only two as in the prior table. In Panel (b), it is measured with the continuous value of $frac\_low\_income_{i,2010}$. The sample includes the award years 2010 and 2012, for the 1,530 private colleges with nonmissing data for both of these years. All regressions include year and school system fixed effects, as described in specification (2.1) in the text. Standard errors are clustered by school system.

	(1) Ln(PLUS num)	(2) Ln(PLUS vol)	(3) Ln(Headcount)	(4) Ln(UFTEs)
Bin 2/4 × After	−0.0790***	−0.0840***	−0.00324	−0.00806
	(0.0220)	(0.0247)	(0.0131)	(0.0128)
Bin 3/4 × After	−0.253***	−0.267***	−0.0338**	−0.0497***
	(0.0300)	(0.0324)	(0.0158)	(0.0160)
Bin 4/4 × After	−0.468***	−0.457***	−0.116***	−0.147***
	(0.0445)	(0.0485)	(0.0191)	(0.0224)
School FE?	Yes	Yes	Yes	Yes
Year FE?	Yes	Yes	Yes	Yes
R^2	0.373	0.202	0.0362	0.0591
Obs.	2924	2924	3060	3060

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Panel (a)

	(1) Ln(PLUS num)	(2) Ln(PLUS vol)	(3) Ln(Headcount)	(4) Ln(UFTEs)
Frac_low_income$_{2010}$ × After	−0.840***	−0.833***	−0.141***	−0.196***
	(0.0669)	(0.0711)	(0.0289)	(0.0333)
School FE?	Yes	Yes	Yes	Yes
Year FE?	Yes	Yes	Yes	Yes
R^2	0.409	0.240	0.0243	0.0491
Obs.	2924	2924	3060	3060

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Panel (b)

Table 2.5. Effect of the credit standards tightening on college revenue, pricing, and spending: Alternative approaches to measuring exposure to the standards-tightening treatment

This table repeats the specifications of Table 2.3, but with alternative approaches to measuring exposure to the standards-tightening treatment. In Panel (a), exposure is measured with four separate bins of $frac\_low\_income_{i,2010}$, instead of only two as in the prior table. In Panel (b), it is measured with the continuous value of $frac\_low\_income_{i,2010}$. The sample includes the award years 2010 and 2012, for the 1,530 private colleges with nonmissing data for both of these years. All regressions include year and school system fixed effects, as described in specification (2.1) in the text. Standard errors are clustered by school system.

	(1) Ln(Tuition rev.)	(2) Tuition	(3) Rev/student	(4) Ln(Expenses)
Bin 2/4 × After	−0.0130	−471.4***	−492.6***	−0.00372
	(0.0135)	(98.97)	(151.3)	(0.0116)
Bin 3/4 × After	−0.0392**	−1284.3***	−732.3***	−0.00853
	(0.0185)	(157.8)	(161.2)	(0.0128)
Bin 4/4 × After	−0.0790***	−1933.9***	−700.7**	−0.0472***
	(0.0289)	(148.3)	(277.0)	(0.0161)
School FE?	Yes	Yes	Yes	Yes
Year FE?	Yes	Yes	Yes	Yes
R^2	0.0302	0.404	0.0792	0.0893
Obs.	3060	3060	3060	3060

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Panel (a)

	(1) Ln(Tuition rev.)	(2) Tuition	(3) Rev/student	(4) Ln(Expenses)
Frac_low_income$_{2010}$ × After	−0.0879**	−3407.0***	−1126.5***	−0.0400*
	(0.0444)	(228.7)	(406.5)	(0.0240)
School FE?	Yes	Yes	Yes	Yes
Year FE?	Yes	Yes	Yes	Yes
R^2	0.0267	0.432	0.0786	0.0844
Obs.	3060	3060	3060	3060

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Panel (b)

Table 2.6. Effects of enrollment on college expenses and revenue: IV regression

This table reports IV regressions of the effects of enrollment on college expenses and revenue. The sample includes the award years 2010 and 2012, for the 1,530 private colleges with nonmissing data for both of these years. Enrollment is measured with the unduplicated twelve-month institutional headcount, and is instrumented with the interaction between the year-2012 indicator and the continuous treatment exposure $frac\_low\_income_{i,2010}$. All regressions include year and school system fixed effects. Standard errors are clustered by school system.

	(1) Ln(Total exp.)	(2) Ln(Instrl exp.)	(3) Ln(Tuition rev.)	(4) Ln(Total rev.)
Ln(Headcount)	0.283**	0.326*	0.621**	0.808***
	(0.141)	(0.191)	(0.268)	(0.159)
School FE?	Yes	Yes	Yes	Yes
Year FE?	Yes	Yes	Yes	Yes
Obs.	3060	3060	3060	3060

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.7. Cross-sectional heterogeneity in the estimated markups

This table uses an instrumental-variables strategy to examine cross-sectional heterogeneity in the estimated markups at private colleges. The sample includes the award years 2010 and 2012, for the 1,502 private colleges with nonmissing data for both of these years. We exclude schools in the District of Columbia and Puerto Rico, for which data on higher-education appropriations are not available. The outcome variable in each column is log total expenses, and the key explanatory variable is log total revenues. A higher (instrumented) effect of log revenues on log expenses reflects a lower markup and Lerner index. In columns 2, 3, and 4, we examine interaction effects between log revenues and three cross-sectional variables: an indicator for being a for-profit school; the number of public schools in a state; and the dollar amount of higher-education appropriations (HEA) in the state budget, respectively. Let $x_{i,2010}$ denote any of these three additional cross-sectional variables. Each is measured as of 2010, demeaned, and standardized. In each column, $\text{Ln(Total revenues)}_{it}$ and $\text{Ln(Total revenues)}_{it} \times x_{i,2010}$ are treated as endogenous. The first stage regression includes the interaction terms $frac\_low\_income_{i,2010} \times After_t$ and $frac\_low\_income_{i,2010} \times After_t \times x_{i,2010}$, as well as the main effects $frac\_low\_income_{i,2010}$, $x_{i,2010}$, and $After_t$. Here, $frac\_low\_income_{i,2010}$ is the fraction of students with family income less than \$48,000 in 2010, and $After_t$ is an indicator for the year 2012. In the second stage, we exclude $frac\_low\_income_{i,2010} \times After_t$ and $frac\_low\_income_{i,2010} \times After_t \times x_{i,2010}$, so that these two terms are the instruments identifying the coefficients on the endogenous terms reported in the table. (Note that the first of these two instruments is the same instrument used in Table 2.6.) Standard errors are clustered by school.

	(1) Ln(Expenses)	(2) Ln(Expenses)	(3) Ln(Expenses)	(4) Ln(Expenses)
Ln(Total revenues)	0.349**	0.844***	0.364**	0.355**
	(0.155)	(0.123)	(0.171)	(0.175)
Ln(Tot. rev.) × For-profit		−0.535***		
		(0.178)		
Ln(Tot. rev.) × # Public			0.0920**	
			(0.0411)	
Ln(Tot. rev.) × HEA				0.0651**
				(0.0268)
Other Controls	Yes	Yes	Yes	Yes
Obs.	3004	3004	3004	3004

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.8. Heterogeneity with *state* fixed effects

This table repeats the analysis of Table 2.7, but includes fixed effects for the *state* in which the school is located. All standard errors are clustered by state.

	(1) Ln(Expenses)	(2) Ln(Expenses)	(3) Ln(Expenses)	(4) Ln(Expenses)
Ln(Total revenues)	0.349**	0.854***	0.366**	0.355**
	(0.171)	(0.157)	(0.171)	(0.175)
Ln(Tot. rev.) × For-profit		−0.546***		
		(0.160)		
Ln(Tot. rev.) × # Public			0.106**	
			(0.0466)	
Ln(Tot. rev.) × HEA				0.0599**
				(0.0254)
Fixed effect	State	State	State	State
Other Controls	Yes	Yes	Yes	Yes
Obs.	3004	3004	3004	3004

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

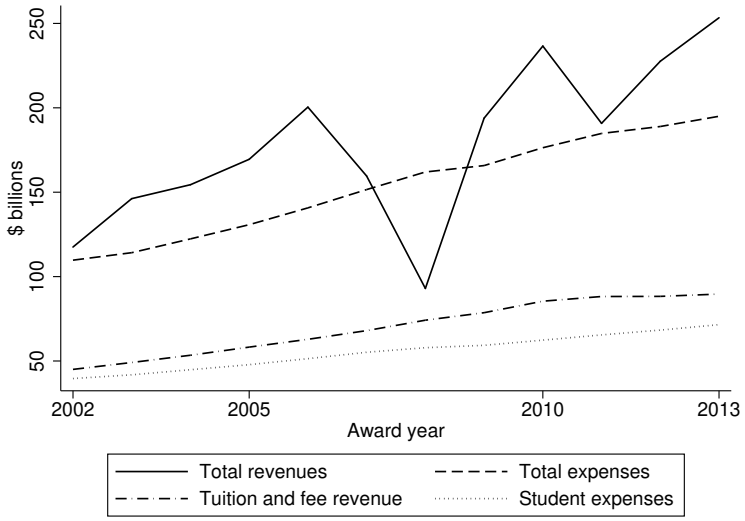


Figure 2.1. Aggregate revenues and expenses for private colleges

Source: Title IV.

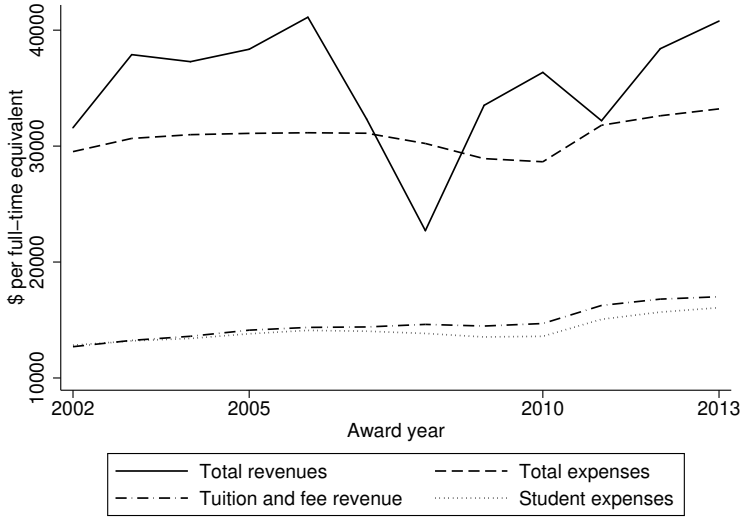


Figure 2.2. Average revenues and expenses per student for private colleges

Source: Title IV.

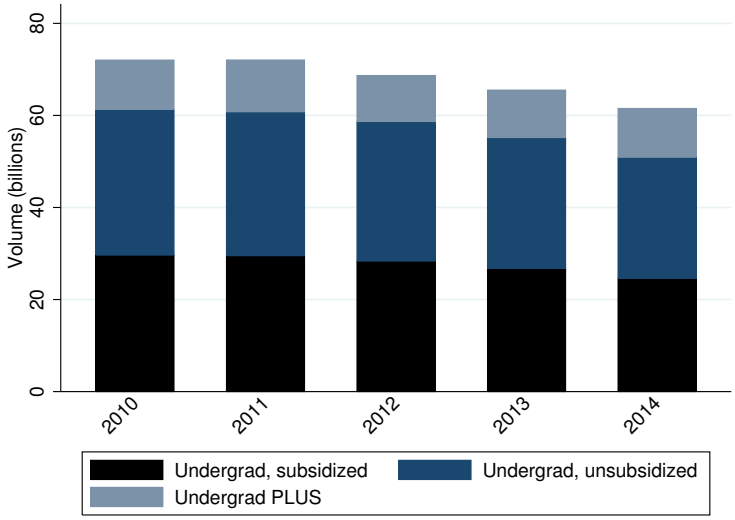


Figure 2.3. Total borrowing through subsidized Stafford, unsubsidized Stafford, and PLUS programs since 2010

Source: Title IV.

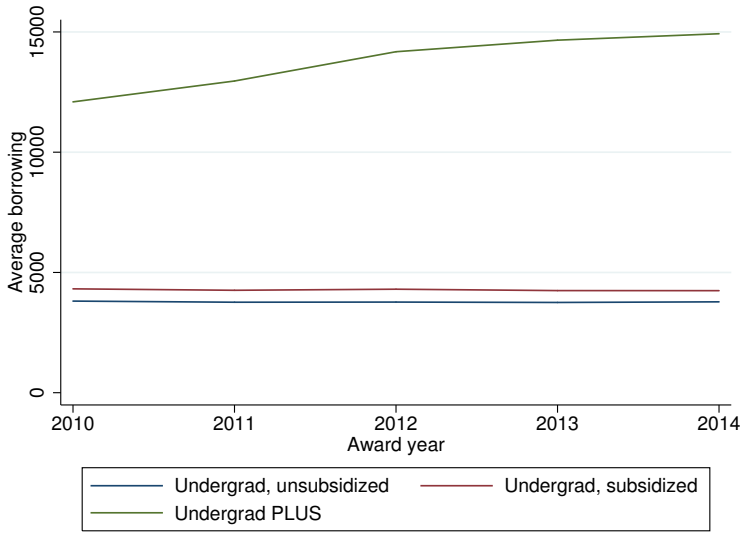


Figure 2.4. Average borrowing through Stafford and PLUS programs since 2010

Source: Title IV. Averages are calculated by dividing aggregate loan volumes over total number of borrowers reported across all schools. This figure is similar to one in Cooper (2016).

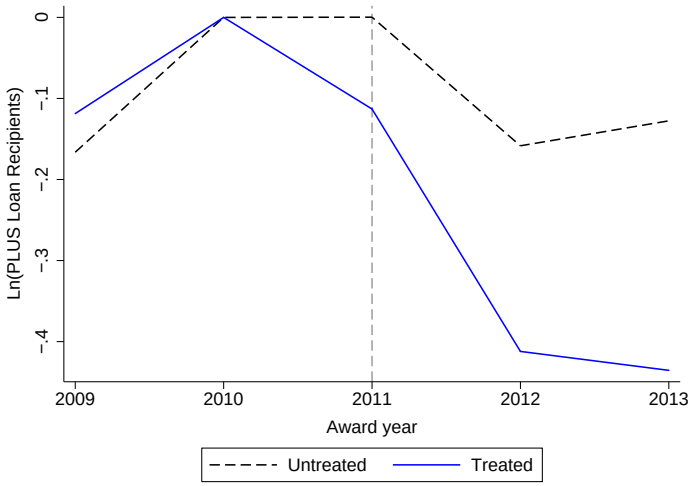


Figure 2.5. Number of PLUS loan recipients for treated and untreated school systems

Both series use annual data and are in log scale and shifted by the subsample's 2010 mean. Schools are sorted into two bins according to whether a school's value of $frac\_low\_income_{i,2010}$ was above the median value (treated) or below the median value (untreated) for schools in the same state. Source: IPEDS/Title IV.

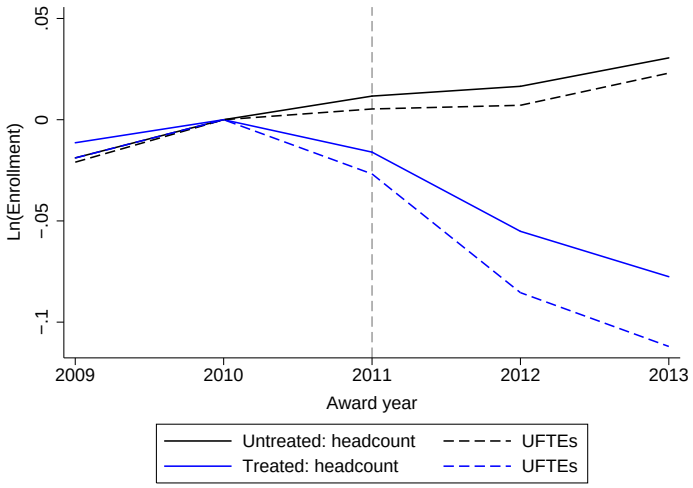


Figure 2.6. Enrollment at treated and untreated schools

Both series are in log scale and shifted by the subsample's 2010 mean. Schools are sorted into two bins according to whether a school's value of $frac\_low\_income_{i,2010}$ was above the median value (treated) or below the median value (untreated) for schools in the same state. Two measures of enrollment are used: Headcount and undergraduate full-time equivalents. Source: IPEDS/Title IV.

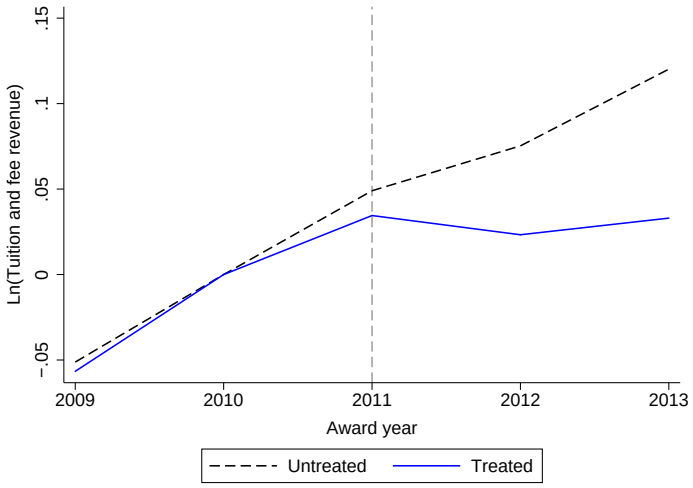


Figure 2.7. Total tuition and fee revenue at treated and untreated private schools

Both series are in log scale and shifted by the subsample's 2010 mean. Schools are sorted into two bins according to whether a school's value of $frac\_low\_income_{i,2010}$ was above the median value (treated) or below the median value (untreated) for schools in the same state. Source: IPEDS/Title IV.

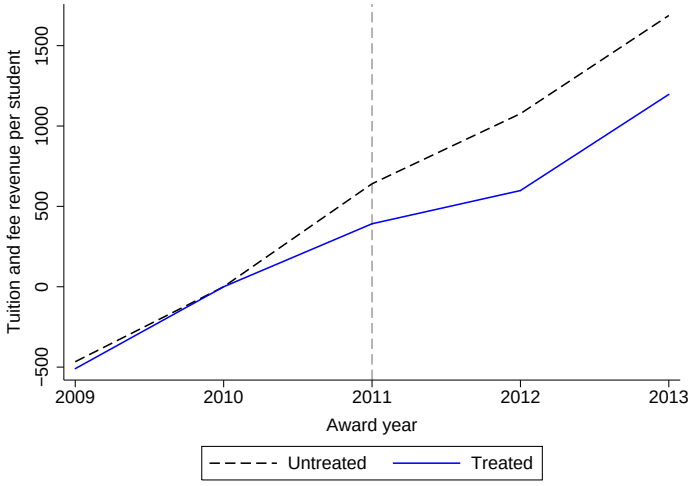


Figure 2.8. Tuition and fee revenue per undergraduate student at treated and untreated private schools

Both series are shifted by the subsample's 2010 mean. Schools are sorted into two bins according to whether a school's value of $frac\_low\_income_{i,2010}$ was above the median value (treated) or below the median value (untreated) for schools in the same state. Source: IPEDS/Title IV.

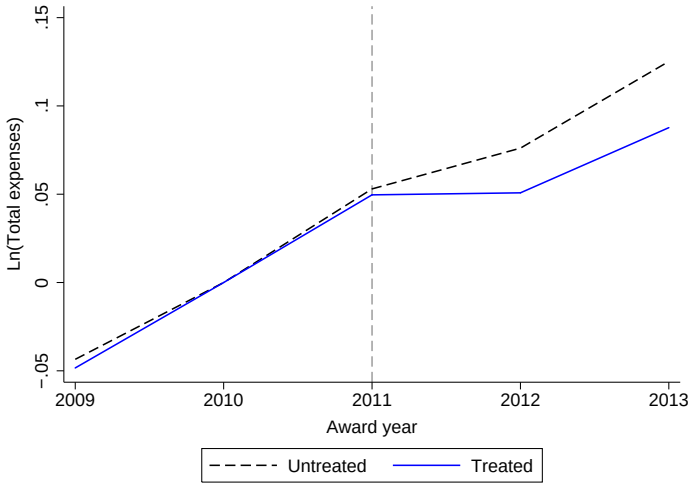


Figure 2.9. Total expenditures at treated and untreated private schools

Both series are in log scale and shifted by the subsample's 2010 mean. Schools are sorted into two bins according to whether a school's value of $frac\_low\_income_{i,2010}$ was above the median value (treated) or below the median value (untreated) for schools in the same state. Source: IPEDS/Title IV.

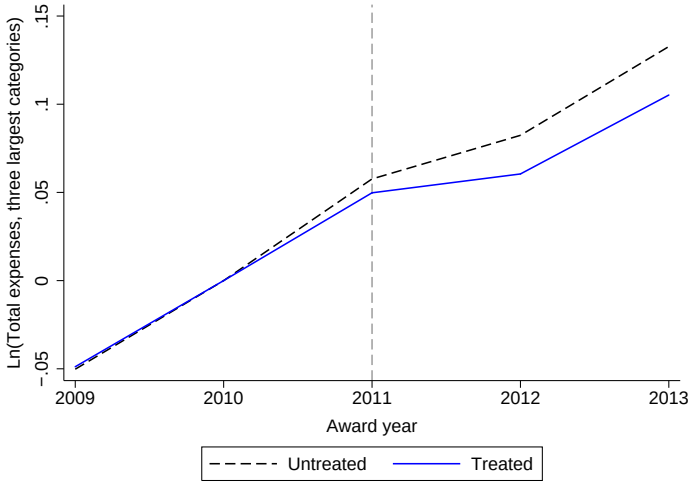


Figure 2.10. Total expenditures at treated and untreated private schools on only the three largest categories

The three largest expense categories are instructional, support, and research/service, respectively. Both series are shifted by the subsample's 2010 mean. Colleges are sorted into two bins according to whether a school's value of $frac\_low\_income_{i,2010}$ was above the median value (treated) or below the median value (untreated) for schools in the same state. Source: IPEDS/Title IV.

2.9 Bibliography

Anderson, N. (2013). Tighter Federal Lending Standards Yield Turmoil for Historically Black Colleges. *The Washington Post*.

Angrist, J. and J. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile Prices in Market Equilibrium. *Econometrica* 63.4, pp. 841–890.

Bhole, M. (2017). Why do Federal Loans Crowd Out the Private Market? Evidence from Graduate PLUS Loans. Working Paper.

Bos, M., E. Breza, and A. Liberman (2018). The Labor Market Effects of Credit Market Information. *The Review of Financial Studies* 31.6, pp. 2005–2037.

Bos, M. and L. I. Nakamura (2014). Should Defaults be Forgotten? Evidence from Variation in Removal of Negative Consumer Credit Information. Federal Reserve Bank of Philadelphia Working Paper No. 12-29/R.

Cabral, M., M. Geruso, and N. Mahoney (2018). Do Larger Health Insurance Subsidies Benefit Patients or Producers? Evidence from Medicare Advantage. *American Economic Review* 108.8, pp. 2048–87.

Card, D. (2001). Estimating the Return to Schooling: Progress on some Persistent Econometric Problems. *Econometrica* 69.5, pp. 1127–1160.

Cellini, S. R. and C. Goldin (2014). Does Federal Student Aid Raise Tuition? New Evidence on For-Profit Colleges. *American Economic Journal: Economic Policy* 6.4, pp. 174–206.

Cellini, S. R. and N. Turner (2016). Gainfully Employed? Assessing the Employment and Earnings of For-Profit College Students Using Administrative Data. Working Paper 22287, National Bureau of Economic Research.

Cooper, P. (2016). The Negatives of PLUS Loans. The Manhattan Institute, Retrieved from http://economics21.org/html/negatives-plus-loans-1630.html.

Cox, N. M. (2017). Pricing, Selection, and Welfare in the Student Loan Market: Evidence from Borrower Repayment Decisions. Working Paper.

Eaton, C., S. T. Howell, and C. Yannelis (2018). When Investor Incentives and Consumer Interests Diverge: Private Equity in Higher Education. Working Paper.

Epple, D., R. Romano, S. Sarpça, and H. Sieg (2017). A General Equilibrium Analysis of State and Private Colleges and Access to Higher Education in the U.S. *Journal of Public Economics* 155, pp. 164–178.

Epple, D., R. Romano, and H. Sieg (2006). Admission, Tuition, and Financial Aid Policies in the Market for Higher Education. *Econometrica* 74.4, pp. 885–928.

Federal Reserve Bank of New York, Research and Statistics Group (2017). Quarterly Report on Household Debt and Credit. https://www.newyorkfed.org/medialibrary/interactives/householdcredit/data/pdf/HHDC_2017Q1.pdf.

Fillmore, I. (2016). Price Discrimination and Public Policy in the US College Market. Working Paper.

Fishman, R. (2014). The Parent Trap: Parent PLUS Loans and Intergenerational Borrowing. *New America Education Policy Program* 3.

Fos, V., A. Liberman, and C. Yannelis (2017). Debt and Human Capital: Evidence from Student Loans. Working Paper.

Fu, C. (2014). Equilibrium Tuition, Applications, Admission, and Enrollment in the College Market. *Journal of Political Economy* 122.2, pp. 225–281.

Garmaise, M. J. and G. Natividad (2017). Consumer Default, Credit Reporting and Borrowing Constraints. *The Journal of Finance* 72.5, pp. 2331–2368.

Goldrick-Rab, S., R. Kelchen, and J. Houle (2014). The Color of Student Debt: Implications for Federal Loan Program Reforms for Black Students and Historically Black Colleges and Universities. Working Paper.

González-Uribe, J. and D. Osorio (2014). Information Sharing and Credit Outcomes: Evidence From a Natural Experiment. Working Paper.

Gordon, G. and A. Hedlund (2016). Accounting for the Rise in College Tuition. Working Paper 21967, National Bureau of Economic Research.

Havnes, T. and M. Mogstad (2011). Money For Nothing? Universal Child Care and Maternal Unemployment. *Journal of Public Economics* 95.11, pp. 1455–1465.

Hotelling, H. (1938). The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates. *Econometrica* 6.3, pp. 242–269.

Johnson, M., J. Bruch, and B. Gill (2015). Changes in Financial Aid and Student Enrollment at Historically Black Colleges and Universities After the Tightening of PLUS Credit Standards. *"U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic,* Retrieved from http://ies.ed.gov/ncee/edlabs.

Kantrowitz, M. (2009). Parent PLUS Loan Denial Rates in the FFEL and Direct Loan Programs. Retrieved from http://www.finaid.org/educators/20090831parentplusdenial.pdf.

Kelchen, R. (2014). Exploring Trends and Alternative Allocation Strategies for Campus-Based Financial Aid Programs. Working Paper.

Krishnan, K. and P. Wang (2018). The Cost of Financing Education: Can Student Debt Hinder Entrepreneurship? *Management Science*. (forthcoming).

Lau, C. V. (2015). The Incidence of Federal Subsidies in For-profit Higher Education. Working Paper.

Liberman, A. (2016). The value of a good credit reputation: Evidence from credit card renegotiations". *Journal of Financial Economics* 120.3, pp. 644–660.

Long, B. T. (2004). How Do Financial Aid Policies Affect Colleges? The Institutional Impact of the Georgia HOPE Scholarship. *The Journal of Human Resources* 39.4, pp. 1045–1066.

Lucca, D. O., T. Nadauld, and K. Shen (2018). Credit Supply and the Rise in College Tuition: Evidence from the Expansion in Federal Student Aid Programs. *The Review of Financial Studies* 32.2, pp. 423–466.

Musto, D. K. (2004). What Happens When Information Leaves a Market? Evidence from Postbankruptcy Consumers. *The Journal of Business* 77.4, pp. 725–748.

Nelson, L. A. (2012). Cracking Down on PLUS Loans. *Inside Higher Ed.* Retrieved from https://www.insidehighered.com/news/2012/10/12/standards-tightening-federal-plus-loans.

Singell, L. D. and J. A. Stone (2007). For Whom the Pell Tolls: The Response of University Tuition to Federal Grants-in-aid. *Economics of Education Review* 26.3, pp. 285–295.

Turner, L. J. (2017). The Economic Incidence of Federal Student Grant Aid. Working Paper.

Turner, N. (2012). Who Benefits from Student Aid? The Economic Incidence of Tax-based Federal Student Aid. *Economics of Education Review* 31.4, pp. 463–481.