Causal Inference for Asset Pricing[*]

Valentin Haddad Zhiguo He Paul Huebner
Péter Kondor Erik Loualiche

June 2025

Download the latest version here.

Abstract

This paper provides a guide for using causal inference with asset prices and quantities. Our framework revolves around an elementary assumption about portfolio demand: homogeneous substitution conditional on observables. Under this assumption, standard cross-sectional instrumental variables or difference-in-difference regressions identify the relative demand elasticity between assets with the same observables, the difference between own-price and cross-price elasticity. In contrast, identifying aggregate elasticities and substitution along specific characteristics requires joint estimation using multiple sources of exogenous time-series variation. The same principles apply to the estimation of multipliers measuring the price impact of supply or demand shocks. Our assumption maps to familiar restrictions on covariance matrices in classical asset pricing models, encompass demand models such as logit, and accommodate rich substitution patterns even outside of these models. We discuss how to design experiments satisfying this condition and offer diagnostics to validate it.

[*] Haddad: Anderson School of Management, UCLA and NBER, valentin.haddad@anderson.ucla.edu; He: Graduate School of Business, Stanford University and NBER, hezhg@stanford.edu; Huebner: Stockholm School of Economics, paul.huebner@hhs.se; Kondor: London School of Economics and CEPR, p.kondor@lse.ac.uk; and Loualiche: Carlson School of Management, University of Minnesota, elouali c@umn.edu. We thank Mikhail Chernov, Carter Davis, Zhiyu Fu, William Fuchs, Satoshi Fukuda, Xavier Gabaix, Sergei Glebkin, Paul Goldsmith-Pinkham, Elise Gourier, Kristy Jansen, Ralph Koijen, Jane Li, Daniel Neuhann, Hanbin Yang, Motohiro Yogo, and seminar and conference participants at Minnesota Carlson, Chicago Booth Alumni Insights, UCLA Anderson, Stockholm School of Economics, Demand in Asset Markets Working Group, USC Macro-Finance Workshop, Adam Smith Workshop, Spring Finance Workshop, SFS Cavalcade, FIRS, CFF Asset Pricing and Machne Learning, LBS Summer Symposium, and SoFiE Pre-Conference for helpful comments and suggestions.

Introduction

Causal inference methods that leverage plausibly exogenous sources of variation have become essential tools in empirical economics (Angrist and Pischke, 2009). Recently, these methods have gained traction in asset pricing to better understand the demand for financial assets, through both specific experiments like index inclusions or central banks' asset purchases (Shleifer, 1986; Chang et al., 2014; Krishnamurthy and Vissing-Jorgensen, 2011), and as building blocks for demand systems (Koijen and Yogo, 2019; Haddad et al., 2024). However, these approaches differ sharply from traditional empirical methods in asset pricing (see, e.g., Cochrane, 2005; Campbell, 2017), which instead prioritize tests of equilibrium relationships such as Euler equations or the CAPM.

We provide a framework for using causal inference in the asset pricing context. We put forward elementary conditions that allow the use of the standard toolbox of causal inference while entertaining a rich set of finance models.[1] Under these conditions, we fully characterize what sources of variation and estimation procedures identify portfolio demand and its equilibrium impact.

Consider the example of a researcher who has detailed data on a pension fund's corporate bond holdings, and wants to understand how the fund's portfolio choices respond to prices. Regressing portfolio positions on prices suffers from the classic endogeneity issue: the holding of a bond might respond to its price or the price might respond to heightened demand by the fund and other investors. However, the researcher has found a source of exogenous variation in prices: the Fed decides to engage in a one-off experiment and randomly purchases some bonds but not others. The researcher handles this natural experiment as usual. They run the following instrumental variable (IV) regression specification to estimate an elasticity of demand $\widehat{\mathcal{E}}$:

$$\Delta D_i = \widehat{\mathcal{E}} \Delta P_i + \theta' X_i + e_i, \tag{1}$$

$$\Delta P_i = \lambda Z_i + \eta' X_i + u_i, \tag{2}$$

where ΔD_i is the change in demand for asset i, ΔP_i is the change in price, the instrument Z_i is the quantity purchased by the Fed, and X_i are observables like bond characteristics. However, the researcher recalls a central tenet of finance: it is portfolio choice, not individual security demand. The pension fund does not choose each holding in isolation but instead

[1] This framework offers a complement to approaches making stronger structural assumptions about portfolio choice so that estimation can be done with fewer sources of exogenous variation (e.g., Koijen and Yogo (2019) with asset demand of logit form) or without exogenous variation (e.g., Hansen and Singleton (1982) with CRRA utility and rational expectations).

forms a portfolio; the price of all assets affects the demand for all assets. For example, if the fund sells the bond of a green firm following a price increase, it might replace this position by investing disproportionately more in other green bonds than in brown bonds in order to manage its portfolio's environmental tilt. Portfolio choice implies that demand is characterized by a matrix $\boldsymbol{\mathcal{E}}$:

$$\mathit{\Delta} D = \boldsymbol{\mathcal{E}} \mathit{\Delta} P + \epsilon, \tag{3}$$

where the off-diagonal elements capture substitution patterns. In general, this cross-dependence implies that the regression of equations (1)-(2) is misspecified. This is the well-known challenge of demand estimation with multiple goods: the SUTVA assumption of canonical causal inference is violated, and the many prices of other assets are omitted variables.[2]

We propose a solution to the researcher in the spirit of the causal inference literature: make an elementary assumption about demand (technically, a restriction on the matrix $\boldsymbol{\mathcal{E}}$) which is flexible enough to accommodate a wide range of plausible investor behaviors while being restrictive enough to facilitate estimation. Informally, the assumption of *homogeneous substitution conditional on observables* holds if, when choosing with which bonds to substitute, the pension fund differentiates bonds with different observables — say greenness and duration — but treats symmetrically bonds with the same value of these observables.

Under this assumption, the good news is that the coefficient $\widehat{\mathcal{E}}$ consistently measures relative elasticity: the response in demand for one asset relative to another one with the same observables to a change in the relative price of these assets. The bad news is that the researcher cannot estimate substitution across assets with different greenness or duration within a cross-sectional setting like the Fed's experiment. Hence the researcher cannot completely characterize how portfolio holdings would change with different prices — the entire matrix $\boldsymbol{\mathcal{E}}$ — with this experiment alone. Not all is lost though: we demonstrate that the researcher can get at those substitution patterns by focusing on exogenous time series variation in the aggregate price of bonds and the price of portfolios based on greenness and duration. For example, one might imagine another series of Fed experiments which vary over time the amount of bonds the Fed purchases, and the tilt of these purchases between green and brown firms, and long-term and short-term bonds.

The paper derives these results formally and generally, and discusses how to apply them. We spell out: a) which assumptions one needs to defend to use causal inference techniques, b) diagnostics to assess the plausibility of these assumptions, c) which technique and source of variation is appropriate for different economic questions, d) how to interpret causal estimates.

[2] An early statement of this challenge is in Deaton and Muellbauer (1980). Berry and Haile (2021) reviews it in the context of the industrial organization literature, while Fuchs et al. (2025) does so for asset pricing.

With this guide, we hope the reader will be empowered to use and interpret evidence from natural experiments in asset markets as well as to understand their limits. The remainder of the introduction highlights the main takeaways.

Multipliers. The framework also applies to the measurement of multipliers or price impacts, i.e., the effect of an exogenous supply or demand shock for an asset on its price. For example, how do Fed asset purchases affect bond prices? This type of question flips prices and quantities compared to demand estimation, with the multiplier matrix $\boldsymbol{\mathcal{M}}$ measuring how demand for all assets affects the price of all assets. Under our assumption, cross-sectional causal inference only estimates the relative multiplier $\widehat{\mathcal{M}}$: the relative response in price for two assets with the same observables to a change in the relative supply of these assets.[3] Estimating cross-multipliers is necessary to understand the price impact of broader shifts such as a change in the demand for green bonds; doing so requires exogenous shifts in the time-series of demand across the observables.

Formal assumptions. Our main assumption is homogeneous substitution conditional on observables: the demand for all assets in the estimation with the same values of observables must react similarly to the price of all other assets in the investor's investment set. For example, this condition restricts the funds' demand for 10-year bonds of Ford and General Motors to respond in the same way to the price of 5-year bonds of First Solar. Meanwhile, bonds with different duration or greenness are allowed to respond differently to First Solar's bond price. Importantly, this condition must apply both to substitution with respect to other assets inside the estimation and outside of the sample.[4] This assumption is central to our identification results. For the cross-sectional regression, it ensures that the response of demand to other prices is proportional to observables and hence is absorbed (but not estimated) in the coefficient vector θ.

To simplify the analysis, our baseline adds an assumption of constant relative elasticity. In this case, when the instrument Z_i satisfies the usual exclusion and relevance conditions, the cross-sectional estimate $\widehat{\mathcal{E}}$ reveals this value: the difference between own and cross-price elasticity for any two assets with the same values of observables. We show how to relax this condition and obtain either conditional estimates or local average effects.[5]

What if the assumption is not satisfied? If variation in substitution is not captured by

[3] Under our assumptions, we analytically show that relative elasticity and multiplier are inverse of each other with $\widehat{\mathcal{M}} = \widehat{\mathcal{E}}^{-1}$, implying that both estimation approaches convey the same information.

[4] Excluded assets outside of the sample can arise because of the common issue that the econometrician does not observe all of the investor's holdings, or by design so as to make the two conditions more plausible in the sample.

[5] We favor the simple setting with homogeneous treatment as a baseline because standard regression methods do not always lead to average treatment effect estimation in presence of controls. Goldsmith-Pinkham et al. (2024) explain this challenge and propose some alternative estimation approaches.

observables, the relative elasticity will generally be biased. Consider a situation where a bond outside of the sample is a closer substitute to treated bonds than control bonds and its price increases. The differential change in demand across the two groups when the pension fund buys substitutes for this position will be wrongfully attributed to variation in the instrument.

Equilibrium spillovers and exclusion restriction. One might worry that because prices are an equilibrium outcome, the natural experiment will generate spillovers across assets. In our example, even if the Fed does not buy a bond, the price of this bond might respond to purchases of its substitutes. This type of spillovers does not affect our identification result. Instrument exogeneity only means that treatment status (whether the Fed bought the bond or not) is unrelated to shifts in the investor's demand curve such as changes in their preferences or their views about the assets: Z_i is orthogonal to the shift in demand ϵ_i, conditional on the observables X_i. In other words, our assumptions about the structure of demand are sufficient to address the challenge of cross-asset spillovers in estimation of relative elasticity highlighted in Fuchs et al. (2025).

Estimating substitution. Our other identification result is that, under our assumptions, substitution can be completely estimated with exogenous variation in the price of portfolios based on each observable and an aggregate portfolio. Combined with relative elasticity from the cross-section, these estimates achieve identification of the entire matrix $\boldsymbol{\mathcal{E}}$.

We first show a general decomposition result that holds if demand satisfies homogeneous substitution conditional on observables. In this case, after adjusting for relative elasticity, there is no need to track substitution for each pair of assets. Instead it is enough to only consider substitution across the observables.[6] Concretely, how will the pension fund adjust its tilts towards green bonds or long-duration bonds and its total holdings of bonds if the price of green bonds increases relative to brown bonds (more precisely, if the price of a greenness-weighted long-short portfolio increases)? When researchers answer this question, as well as measure adjustments in response to changes in the price of a duration-weighted portfolio and of an aggregate bond portfolio fully, they have characterized substitution entirely. The flip side of this result is that answering questions at the meso (along observables) or macro (the aggregate) levels cannot be done with the cross-sectional estimates and requires fully estimating substitution.

Once the researcher is down to these few portfolios, they can simply consider them as distinct synthetic assets — even if they have overlapping holdings — and analyze demand for them. With only a few portfolios (three in our example) and no additional structure on their substitution, only exogenous time series variation in all of their prices allows estimation. This must be done jointly across the portfolios. Say the researcher is interested in the fund's

[6] Gabaix and Koijen (2021) derive a case without observables.

macro elasticity, that is, how the pension fund's overall demand for bonds responds to the aggregate price of bonds. A simple approach would regress total demand on the aggregate price instrumented by a shock orthogonal to shifts in the fund's demand curve. In a framework with observables this is not enough: the researcher must incorporate additional instruments for the prices of the greenness- and duration-sorted portfolios as well.[7]

Using the framework. Homogeneous substitution conditional on observables accommodates a large variety of portfolio demands. This versatility allows the framework to fit different types of natural experiments.

A simple version of this approach without including observables is particularly suitable to settings where the natural experiment only touches a few assets. Then the researcher must argue that demand for each of these assets would respond in the same way to the price of every other asset, including outside of the sample. This concern can guide the design of the estimation sample: the researcher can restrict themself to assets that are as similar as possible (in the same industry, with the same size, etc.). To substantiate their choice, and because risk often drives substitution, the researcher could present evidence of similar covariances of treated and control with various portfolios of assets, that is evidence of balance in betas.

Using observables to capture heterogeneous substitution opens up the possibility of larger samples with more heterogeneity. One could have multiple groups of assets, where elasticities between assets in each group are symmetric, but elasticities across groups differ. Then, the observables are group fixed effects (Chaudhary et al., 2022). Often though, substitution is driven by continuous variables that cannot be delimited by a group. This occurs when the investor is managing aggregate statistics of the portfolio, and each asset contributes to these statistics with its own loadings. For example, in standard risk-based portfolio optimization, the investor manages the portfolio's betas on various factors.[8] This implies that observables should include the betas (directly or through characteristics proxying for them) of each asset with respect to these factors (Koijen and Yogo, 2019). Heterogeneous substitution can also be driven by other motives than risk. An investor might balance their portfolio's carbon emissions (e.g. our pension fund facing pressure from its stakeholders) or target some regulatory constraint, and hence substitute across assets based on the corresponding characteristic for each asset; incorporating such characteristics is crucial as well.

Relation to structural models. This framework also helps revisit results of the literature using structural models for demand estimation. Koijen and Yogo (2019) and Koijen et al. (2023) introduce a model of portfolio demand in the logit form. While they prove *existence* of factor

[7] Simply controlling for the price of these portfolios could lead to a bad control (Angrist and Pischke, 2009) situation which introduces endogeneity.

[8] This occurs, for example, in a mean-variance model with constant volatility and expected returns that depend linearly on the price. Then, elasticity is proportional to the covariance matrix.

models that yield asset demand in a logit form as the outcome of log-utility maximization, logit demand cannot capture the rich substitution patterns of *generic* factor models.[9] To see why, simply go back to our pension fund reacting to an increase in the price of a single green bond, assuming that the fund's portfolio is otherwise equally composed of brown and green bonds. With logit demand, the fund would replace the shocked green bond equally by green and brown bonds. Meanwhile, with log utility and a simple factor model, they would tilt how much they replace it between the two categories depending on risk factors such as the risk of the green-minus-brown portfolio.[10] This implies that the two models generically lead to different responses of the pension fund's demand to changes in prices, and that they would imply different portfolios and equilibrium prices in counterfactual exercises. Our framework encompasses logit and all factor models (and many others). Because both models satisfy homogeneous substitution conditional on observables, the process for estimating relative elasticity within each of them is the same. In logit, the strong structure of demand implies that substitution can be inferred from relative elasticity; generally, estimating substitution requires additional sources of variations.

Related Literature. A long tradition in finance uses plausibly exogenous sources of variation to understand portfolio decisions and the price impact of shifts in demand. Prominent examples include the effect of index inclusion (Shleifer, 1986; Harris and Gurel, 1986; Chang et al., 2014; Pavlova and Sikorskaya, 2022; Greenwood and Sammon, 2024), institutional ownership and fund flows (Gompers and Metrick, 2001; Coval and Stafford, 2007; Lou, 2012; Ben-David et al., 2022; Hartzmark and Solomon, 2022), central bank asset purchases (Krishnamurthy and Vissing-Jorgensen, 2011; Selgrad, 2023; Haddad et al., 2021, 2025), or financial constraints (Du et al., 2018; Greenwood and Vissing-Jorgensen, 2018; Haddad and Muir, 2021; Chen et al., 2023). This work often incorporates thorough analysis of exogeneity, in particular in the wake of the "credibility revolution" (e.g., Angrist and Pischke, 2009). However, this literature is often more scant in considering a central feature of asset pricing theory, substitution across assets, and whether it affects the validity of inference and the interpretation of estimates. Our framework provides a simple bridge between classical discussions of causal inference and the role of substitution.

Another approach fully specifies and estimates models of portfolio demand and their equilibrium implications. In a seminal article, Koijen and Yogo (2019) derive and estimate a logit model of portfolio choice in the stock market. Subsequent work uses this model either structurally, or as a semi-structural simplification to introduce other mechanisms (e.g.

[9] Appendix D reproduces their result and establishes formally this distinction.

[10] Furthermore, the substitution portfolio would be different if the initially shocked bond was a brown bond.

Haddad et al. (2024)). Applications include quantifying the impact of the rise of passive investing, preferences for sustainable assets (Koijen et al., 2023; Van der Beck, 2021), or the transmission of monetary policy (Lu and Wu, 2023), and have found echo in other settings: the stock market overall Gabaix and Koijen (2021), corporate bonds (Bretscher et al., 2022), treasuries (Jansen et al., 2024; Fang, 2023; Fang and Xiao, 2024), or exchange rates (Koijen and Yogo, 2024; Jiang et al., 2024).

As we discuss in the text, we build on some of the insights from estimation inside of these models. Some important ideas are controlling for common exposures (Koijen and Yogo, 2019), the distinction between micro and macro elasticity (Gabaix and Koijen, 2021; Li and Lin, 2022), heterogeneous substitution (Chaudhary et al., 2022; Aghaee, 2024), substitution along factors (An et al., 2024; An and Huber, 2025; Peng and Wang, 2023), and accounting flexibly for spillovers (Fuchs et al., 2025). Naturally, our simple conditions cannot cover every model; we leave aside considerations of strategic responses (Haddad et al., 2024), dynamics (Greenwood et al., 2018; Gabaix and Koijen, 2021; Huebner, 2024; He et al., 2025), state-contingent demand shocks (Haddad et al., 2025), intermediary distress (He et al., 2022), or bidding in auctions (Allen et al., 2018). In this context, the contribution of our framework is twofold: it not only provides a unifying formalism to discuss identification across models but also allows discussion of what can be learned from the data before espousing a specific model.

Finally, the role of spillovers is not limited to asset pricing and has been recognized in many other contexts. The industrial organization literature often relaxes the assumption of independence of irrelevant alternative (IIA) and includes heterogeneous substitution in discrete choice models, such as Berry et al. (1995). While without observables, our assumptions would be closely related to IIA, including the observables entertains heterogeneous substitution. Our setting of portfolio choice in line with finance theory does so without introducing nonlinearities, lending itself to using linear regressions. Berg et al. (2021) discuss spillovers in corporate finance. In macroeconomics, a key concern is the missing intercept problem due to general equilibrium effects, with some recent contributions such as Chodorow-Reich et al. (2021), Guren et al. (2021), Huber (2023), and Wolf (2023).

1 The Challenge of Causal Inference in Asset Pricing

We set up the basic regression framework for estimating the demand for assets using canonical causal inference. We contrast this setting with how standard asset pricing theory works. The key distinction is the emphasis on strong patterns of substitution across assets.

1.1 The causal inference framework

We focus on a generic setting for identifying the demand for financial assets. Section 3 considers the related problem of price impact from demand shocks. Intuitively, we want to understand how an investor's demand for an asset responds to the price of this asset. We consider the following experiment: a shock exogenous to demand happens and affects the price P_i of various assets indexed by i, with intensity Z_i.

Inspired by standard causal inference, running an instrumental variable estimation on a sample $\mathcal{S}$ of assets is natural in this setting. In this model, one regresses the change in demand for each asset $\varDelta D_i$ on the change in the price of this asset $\varDelta P_i$, using Z_i as an instrument for the price change. This corresponds to the two-stage least square specification:

$$\varDelta D_i = \widehat{\mathcal{E}} \varDelta P_i + \theta' X_i + \epsilon_i, \tag{4}$$

$$\varDelta P_i = \lambda Z_i + \eta' X_i + u_i, \tag{5}$$

where X is a set of observables for each asset to be specified. For example, X_i could include the maturity of a bond or the industry of a firm. These observables allow narrowing the identification to comparable assets. For simplicity of notation, we always assume that X contains a constant and is of small enough dimension that there is enough variation to run the regression.

The two standard conditions for this regression model to be identified are the relevance and exclusion restrictions. Exclusion is the idea that the instrument does not affect demand through other channels than the price: $Z_i \perp \epsilon_i | X_i$. In other words, the instrument is not correlated with unobservable shifts in the demand curve in the cross-section of assets. For example, even if the experiment leads to general equilibrium effects such as changing the risk-free rate, the exclusion restriction can still be satisfied if the impact of these effects across assets does not correlate with which asset is treated. Relevance is the idea that the instrument Z_i creates variation in prices: $\lambda \neq 0$. In practice, it is not enough for the first stage to be significant at standard confidence levels; it must be strong to avoid issues related to the weak-instrument problem (Stock and Yogo, 2005; Olea and Pflueger, 2013).

One can imagine running this specification in levels or logs depending on the model of demand. For example, models like CARA preferences are better behaved in levels, while logit demand aligns with logs. In practice, the choice of units should be driven by regularity in the data and the type of model the researcher believes best match this regularity. We abuse the language of demand estimation slightly and call coefficients in such regressions demand elasticities irrespective of log or levels. Section 2.4.1 reviews the appropriate units for standard models. Also, while we focus on writing specifications in changes to match the standard difference-in-difference framework, similar arguments apply without changes.

A simpler benchmark To better understand the behavior of this regression, it is useful to study a simplified version. There is no shift in the demand curve, but simply a shock that triggered movements in prices, with the movement in the price of asset 1 larger than in the price of asset 2. There are still many other assets $(3, \ldots, N)$ that might also experience price changes. For example, the shock could be a surprise increase in the supply of asset 1 but not asset 2. In this case, the counterpart to the IV estimator is the relative change in demand for assets 1 and 2 divided by the relative change in price:

$$\widehat{\mathcal{E}} = \frac{\Delta D_1 - \Delta D_2}{\Delta P_1 - \Delta P_2}. \tag{6}$$

To see this result, note that the sample is just the two assets $\mathcal{S} = \{1, 2\}$, the instrument representing the experiment is $Z_1 = 1$ and $Z_2 = 0$, and there are no controls or error terms.[11]

1.2 Standard asset pricing structure

The setting of equation (4) differs sharply from how standard asset pricing theory specifies the demand for assets. A key insight going back to Markowitz (1952) is that assets are not distinct goods but instead alternative means of saving with different risk and reward. Investors choose portfolios optimally combining these assets. This substitutability implies that the demand for one asset depends not only on its own price but also on the price of other assets. How many shares of Apple you purchase depends on the price of Apple and also on the price of Nvidia.

The most standard example of this approach is mean-variance optimization: an investor chooses their portfolio to maximize $\mathbf{E}(W) - \frac{\gamma}{2}\,\text{var}(W)$ where W is their future wealth, and γ measures their absolute risk aversion. If assets have constant mean payoffs M and covariance matrix Σ, the vector of demand is:

$$D = \frac{1}{\gamma}\Sigma^{-1}(M - P). \tag{7}$$

Absent demand shocks, this implies that changes in demand can be written as

$$\Delta D = \mathcal{E}\Delta P \iff \Delta D_i = \sum_j \mathcal{E}_{ij}\Delta P_j, \tag{8}$$

[11] The first stage regresses ΔP_i on the dummy, so $\lambda = \Delta P_1 - \Delta P_2$. The second stage regresses the change in demand on the predicted value from the first stage, $(\Delta P_1 - \Delta P_2)1_{\{i=1\}}$, leading to equation (6).

with the matrix of elasticity $\mathcal{E}$ determined by risk aversion and the covariance between assets: $\mathcal{E} = -\gamma^{-1}\Sigma^{-1}$.[12] When assets are correlated with each other, they become close substitutes, and their demands respond to each other's prices.

In addition to these elasticities, there can also be shifts in demand, for example, due to changing beliefs about expected payoffs. Hereafter, we represent these movements by a component ϵ_i.

Modern finance research acknowledges many deviations from this simple setting: investors have different beliefs and various cognitive limitations, institutions face many regulations and constraints that influence their portfolio decisions. Still the basic idea of portfolio choice as opposed to asset choice remains. Generally, any model of asset demand will imply its own matrix of elasticities $\mathcal{E}$. The diagonal elements of $\mathcal{E}$ measure the own-price elasticities, while the off-diagonal elements capture cross-price elasticities. If the model is not linear (or log-linear) in prices, we focus on a local approximation of demand; Appendix D discusses the nonlinear case. Such an unrestricted elasticity matrix is reminiscent of the almost-ideal demand system of Deaton and Muellbauer (1980).

1.3 The challenge

The distinction between the two approaches is clearly visible: causal inference focuses on a univariate relation between price and demand — the coefficient $\widehat{\mathcal{E}}$ — while standard asset pricing emphasizes a multivariate relation — the matrix $\mathcal{E}$.[13] This univariate focus is a key element of standard causal inference; under the stable unit treatment value assumption (SUTVA), treatment on one unit (for us, an asset) does not affect other units.

This feature implies that, in general, the estimation equation (4) is misspecified. Concretely, the presence of cross-elasticities implies that the prices of all other assets are omitted variables in equation (4). When we have non-zero elasticity of substitutions between assets, the change in the price of other assets affects the demand for the original asset. In changes, the demand system of equation (8) gives:

$$\Delta D_i = \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{ij}\Delta P_j + \epsilon_i. \tag{9}$$

A standard natural experiment focuses on a situation where the instrument is orthogonal to shifts in demand, so $Z_i \perp \epsilon_i$. However, other prices naturally respond to the treatment so

[12] If M and Σ respond to changes in prices, the elasticity formula would be different. Koijen and Yogo (2019) characterize demand curves in such a setting. Koijen et al. (2023) add hedging demands.

[13] While we focus on a static setting, the elasticity matrix also arises in dynamic settings; see, e.g., Gabaix and Koijen (2021).

the other terms in the sum create an omitted variable bias. Fuchs et al. (2025) discuss at length the theoretical foundations of this challenge.

As an example, go back to the deterministic case comparing two assets 1 and 2, and consider the effect of the change in the price of a third asset, say asset 3. This change results in a contribution $(\mathcal{E}_{13} - \mathcal{E}_{23})\Delta P_3$ to the numerator of (6). If the two cross-elasticities differ from each other, this leads to a bias away from the own-price elasticity. This is the standard problem of demand estimation with multiple goods.

In the face of this challenge, one can deem causal inference hopeless for asset pricing and throw their hands in the air. However, there is a more constructive approach: acknowledge that additional assumptions about the nature of spillovers are necessary and that the coefficient $\widehat{\mathcal{E}}$ will only reveal a specific dimension of the matrix $\mathcal{E}$. After all, this is the second part of Markowitz' argument: basic economics can inform us about the structure of substitution across assets. In the rest of the paper, we follow this path and put forward simple, flexible conditions guided by these economic principles.

An alternative, followed for example in Koijen and Yogo (2019), is to fully specify a structural model. The modern empirical industrial organization literature does so as well, albeit with different foundations. We show later how our results intersect with these approaches. Another alternative would be to include all prices in the demand estimation regression. This is often not possible in practice because it would require exogenous sources of variation for each one of the individual prices.

2 Making Causal Inference Work with Asset Pricing

We provide a framework for using cross-sectional causal inference regressions in asset pricing. We give two natural conditions on the structure of substitution that are sufficient for these regressions to identify a meaningful quantity. In the context of risk-based models, the two conditions have a simple interpretation in terms of the statistical structure of asset returns. However, applying these conditions does not require espousing the view that risk is the only driver of investment decisions. We show how they lend themselves to settings with other considerations, such as regulatory constraints or even non-pecuniary objectives.

2.1 Conditions for valid estimation

We state the two conditions leading to valid estimation. First, we put some structure on substitution between assets.

Assumption A1 (Homogeneous substitution conditional on observables) *Any pair of assets in the estimation sample $\mathcal{S}$ with the same observables shares the same cross-price elasticity with respect to each third asset, within or outside of the estimation sample:*

$$\mathcal{E}_{il} = \mathcal{E}_{jl}, \quad \text{for all } i, j \in \mathcal{S} \text{ such that } X_i = X_j, \text{ and } l \neq i, j, \tag{10}$$

where X_i is the $K \times 1$ vector of observables for asset i. These cross-elasticities are parametrized by a bilinear form $\mathcal{E}_{cross}$: $\mathcal{E}_{il} = \mathcal{E}_{cross}(X_i, X_l) = X_i' \mathcal{E}_X X_l$, where $\mathcal{E}_X$ is a $K \times K$ matrix.

Assumption A1 states that for two assets comparable along observables, if the price of any third asset, either within or outside the estimation sample, moves, then substitution between the third asset and the two comparable assets is the same. That is, for the pair of comparable companies Ford and General Motors, if the price of Netflix moves, the response of the demand for Ford will be the same as the response of the demand for General Motors. This assumption is crucial to deal with the omitted variable problem coming from the substitution effect when prices of other assets change. When the condition holds without observables, substitution effects are constant in the sample, so they are absorbed in the constant of the cross-sectional regression. With observables, substitution responses are equal across assets conditional on X_i and thus absorbed into regression coefficients on the observables.

Importantly, assuming that the investor substitutes in the same way with Ford and General Motors is not the same as assuming that the two bonds are identical. Features specific to each of the bonds are still allowed to affect how much the investor demands of them, as materialized by the residual ϵ_i in equation (9). For example, in risk-based models, two assets with the same observables can have the same comovement with other assets, which drives substitution, but each of them has a distinct idiosyncratic component to their returns; see Section 2.3.3 below.

Analytically, homogeneous substitution implies that cross-price elasticities are a function of the observables, which we write as $\mathcal{E}_{cross}(X_i, X_l)$. To make the model tractable, we further parametrize this function as a bilinear form in observables, $\mathcal{E}_{cross}(X_i, X_l) = X_i' \mathcal{E}_X X_l$. The substitution matrix $\mathcal{E}_X$ may not necessarily be symmetric. This simple specification encompasses a large space of potential substitution patterns because observables could already include nonlinear transformations of more primitive variables or dummies for their levels. We demonstrate this versatility and practicality in Section 2.3.

The second assumption ensures that there is a single number to estimate. In the language of causal inference methods, this corresponds to assuming a homogeneous treatment effect. Section 2.3.5 extends the framework to consider situations where the relative elasticity is not constant and either depends on observable or unobservable sources of variations.

Assumption A2 (Constant relative elasticity) *Assets in the estimation sample have the same value of relative elasticity $\mathcal{E}_{relative}$ with respect to other assets with the same characteristics:*

$$\mathcal{E}_{ii} - \mathcal{E}_{ji} = \mathcal{E}_{relative}, \quad \text{for all } i, j \in \mathcal{S} \text{ such that } X_i = X_j. \tag{11}$$

Assumption A2 ensures a form of symmetry in how the investor responds to the price of assets with the same observables in the sample. It focuses on a specific dimension: the difference between the own-price and cross-price elasticity. We call this difference the relative elasticity. It represents how the demand for one asset relative to another shifts when the price of the asset changes relative to the other. The next section explains why this quantity is the natural target of cross-sectional regressions.[14]

As we will show shortly in Section 2.3, these assumptions are valid in a wide variety of contexts. For example, mean-variance optimizing investors who face assets with a covariance matrix satisfying a factor structure and constant idiosyncratic risk have such a demand function. This situation arises in the classic model of Vayanos and Vila (2021) with constant idiosyncratic risk for each bond, where arbitrageurs solve an optimal portfolio problem when absorbing the supply shocks from habitat investors.

The next proposition gives the mathematical structure of the elasticity matrix $\mathcal{E}$ once we combine assumptions A1 and A2, with proof provided in Appendix A.2. Clearly, if the elasticity satisfies the assumptions for a set of observables X, then it also does so for a linear transformation—such as demeaning or standardizing—of these observables.

Proposition 1 *Assumptions A1 and A2 are equivalent to an elasticity matrix $\mathcal{E}$ with the following representation*

$$\mathcal{E} = \mathcal{E}_{relative}\mathbf{I} + X\mathcal{E}_X X'. \tag{12}$$

Assumptions A1 and A2 can be viewed as guidance for the econometrician to choose their sample and their observables appropriately. For example, one might choose to focus on a narrow set of highly comparable assets, making the assumptions plausible. If they want to consider a much larger asset space, the econometrician has to confront more substantial heterogeneity, for example, in risk and how the assets comove with one another. They will have to judiciously choose observables such that the assumptions are credible conditional on those observables. Similarly, the choice of units to define elasticities — demand vs. portfolio shares, change in price vs. return — also affects whether the assumptions hold; Section 2.4.1

[14] If an asset i does not have a "twin" j with identical observables. This often occurs when the observables are continuous variables, such as the sales of a firm. In this situation, we replace Assumption A2 with its natural extension: $\mathcal{E}_{ii} - \mathcal{E}_{\text{cross}}(X_i, X_i) = \mathcal{E}_{relative}$.

shows how in the context of standard models. In practice, the econometrician should focus on units that make assets more comparable. We discuss empirical design in light of the two assumptions in Section 2.3.

2.2 What does it estimate?

We are now ready to state our main proposition.

Proposition 2 *Under assumptions A1 and A2, as well as the standard relevance and exclusion restrictions, the two-stage least square estimation of equations* (4) *and* (5) *identifies the relative elasticity:*

$$\widehat{\mathcal{E}} = \mathcal{E}_{relative}. \tag{13}$$

When the IV estimation is well specified, it identifies the relative elasticity: the difference between the own-price elasticity and the cross-price elasticity for two assets in the sample with the same observables. While this result stands in contrast to the intuition of measuring "how demand for each asset responds to its own price," it is natural. A cross-sectional regression is a comparison across assets in the sample. Even if only the price of the treated asset is shocked, the regression coefficient will still be driven by the response of demand for this asset relative to that for the comparable control asset—hence the relative intensity of the own- and cross-elasticity conditional on observables. In other words, $\widehat{\mathcal{E}}$ answers the question: how does the demand for one asset relative to another comparable asset respond to the relative price of these assets?

Proof for the simple case. Appendix A.1 proves Proposition 2. To understand the mechanics of this result, let us go back to the deterministic case comparing 2 assets (say Ford and General Motors) with the same observables. The changes in demands are:

$$\varDelta D_1 = \mathcal{E}_{11}\varDelta P_1 + \mathcal{E}_{12}\varDelta P_2 + \sum_{k>2} \mathcal{E}_{1k}\varDelta P_k; \tag{14}$$

$$\varDelta D_2 = \mathcal{E}_{22}\varDelta P_2 + \mathcal{E}_{21}\varDelta P_1 + \sum_{k>2} \mathcal{E}_{2k}\varDelta P_k. \tag{15}$$

Assumption A1 implies that the cross-elasticities with respect to other assets ($k > 2$) are identical:

$$\sum_{k>2} \mathcal{E}_{1k}\varDelta P_k = \sum_{k>2} \mathcal{E}_{2k}\varDelta P_k. \tag{16}$$

When computing the difference $\Delta D_1 - \Delta D_2$, this response to other prices disappears, effectively removing the omitted variable problem due to other assets:

$$\Delta D_1 - \Delta D_2 = (\mathcal{E}_{11} - \mathcal{E}_{21})\Delta P_1 - (\mathcal{E}_{22} - \mathcal{E}_{12})\Delta P_2. \tag{17}$$

Assumption A2 implies that the coefficients on each of the prices are the relative elasticity:

$$\mathcal{E}_{11} - \mathcal{E}_{21} = \mathcal{E}_{22} - \mathcal{E}_{12} = \mathcal{E}_{relative} \tag{18}$$

Both the response of demand to the own price (measured by $\mathcal{E}_{11}$) and the response to the price of the other asset asset (measured by $\mathcal{E}_{21}$) shape this comparison. Hence, the regression coefficient is the relative elasticity:

$$\widehat{\mathcal{E}} = \frac{\Delta D_1 - \Delta D_2}{\Delta P_1 - \Delta P_2} = \mathcal{E}_{relative}. \tag{19}$$

The role of observables. In the richer case with observables, it is important to control for the asset's own observables X_i if they vary in the sample. When the assets in the sample differ, they potentially respond differently to the price of other assets. However, assumption A1 ensures that these responses only depend on each asset's observables X_i:

$$\begin{aligned}
\Delta D_i &= \mathcal{E}_{ii}\Delta P_i + \sum_{j \neq i} X_i' \mathcal{E}_X X_j \Delta P_j + \epsilon_i && (20)\\
&= \left(\mathcal{E}_{ii} - X_i'\mathcal{E}_X X_i\right)\Delta P_i + \sum_j X_i'\mathcal{E}_X X_j \Delta P_j + \epsilon_i && (21)\\
&= \underbrace{\left(\mathcal{E}_{ii} - X_i'\mathcal{E}_X X_i\right)}_{\widehat{\mathcal{E}},\text{relative elasticity}} \Delta P_i + X_i' \underbrace{\sum_j \mathcal{E}_X X_j \Delta P_j}_{\text{constant across assets}} + \epsilon_i. && (22)
\end{aligned}$$

The second term in this expression highlights that substitution, while depending on all other prices, is proportional to X_i. This implies that controlling for X_i in a cross-sectional regression absorbs the effects of substitution.[15] Furthermore, the first term in (22) shows that the regression (after controlling for X_i) is equivalent to making pairwise comparisons of assets that have the same observables. Hence, following the same reasoning as in the simple case, the estimate $\widehat{\mathcal{E}}$ recovers the relative elasticity.

At this stage, it might be tempting to conclude that the demand curve of equation (9) in which all prices matter for all demands is equivalent to a demand curve that only depends on

[15] This reasoning shows that a weaker form of assumption A1 is necessary for Proposition 2 to hold: $\mathcal{E}_{il} = \mathcal{E}_{cross,l}(X_i) = X_i'Y_l$ for arbitrary vectors Y_l. In other words, the dependence to other assets for a given X_i can be arbitrary and does not need to be parametrized by observable characteristics X_l.

the own price and characteristics, as in the regression equation (4). This would be incorrect: the equivalent representation only holds when fixing a specific vector of prices. In other words, while equilibrium quantities demanded satisfy equation (4), the demand curve does not. This distinction is transparent when examining what determines the coefficients θ on the observables X_i in the cross-sectional regression. Equation (22) highlights that these coefficients depend on realized changes in prices ΔP_k, and would therefore differ for another realization of prices, such as in a counterfactual exercise.[16]

Robustness to deviations in the assumptions. In practice, assumptions A1 and A2 are approximations of reality. In Appendix B.1, we assess whether the result of Proposition 2 is robust to small deviations. We show that, as long as the first stage is strong, the two-stage least square estimator recovers the relative elasticity up to a bias that is proportional to the distance to the assumptions; small deviations, small bias. An economically meaningful situation that leads to weak instruments is when assets are perfect substitutes, for example if they satisfy a no-arbitrage relation.[17] In this case, there is no change in relative prices that can lead to identification. Fuchs et al. (2025) highlight how this case creates challenges for demand estimation.

The potential for this weak-instrument issue in the first stage highlights a key consideration when selecting an appropriate control group for a given treated asset. While a control asset should be similar enough to the treated to satisfy Assumptions A1 and A2, it should not be identical, ensuring that the law of one price does not hold between these two groups. As discussed later in Sections 2.3.3 and 2.3.4, this difference may arise due to idiosyncratic risk or other non-risk considerations (e.g., regulatory constraints), preventing the market as a whole from pricing them identically in equilibrium.

2.3 Using the identification result

Assumptions A1 and A2 provide general conditions for the causal cross-sectional regression to identify the relative elasticity. To use this result, the econometrician must take a stand on what is the appropriate estimation sample and which are the relevant observables. We discuss a few different approaches to do so, with choices that are intuitive, close to common empirical practice, and line up with standard finance theory.

[16] The coefficients θ are also be driven by the covariance of the demand residual ϵ_i with X_i.

[17] Here, what is the important is that these assets are perfect substitutes at the aggregate level (so that their equilibrium prices are tied together), not so much that the investor whose demand is estimated treats them as such.

2.3.1 Homogeneous estimation sample

A common scenario: you want to assess the effects of a local experiment on a few highly comparable assets. In this case, it is not necessary to introduce observables to differentiate the assets. The only control needed in the regression is a cross-sectional constant. For example, firms in a narrowly defined industry might have similar risk and similar relation with stocks in other industries. Another example could be multiple corporate bonds from the same issuer with similar maturity (see Coppola (2025)). The simplest manifestation of this example is the case of two assets: a treated and a control. There, the regression is equivalent to examining the spread in return between treated and control, a common practice of empirical asset pricing.

In this case, assumption A1 implies that the cross-price elasticity is the same for all assets in the estimation sample, while assumption A2 additionally implies that the own-price elasticity is the same for all assets in the estimation sample:

$$\mathcal{E}_{ii} = \mathcal{E}_{own}, \quad \text{for all } i \in \mathcal{S}, \quad \text{and} \quad \mathcal{E}_{ij} = \mathcal{E}_{cross}, \quad \text{for all } i, j \in \mathcal{S}. \tag{23}$$

Moving on to outside assets, which could be a vast set, the substitution between them and the assets in the estimation sample is generally not constant. Still, assumption A1 implies that, for each outside asset, all assets in the estimation sample have the same cross-elasticity. In other words, the demand for any asset in the sample responds in the same way to a change in the prices of each outside assets. Figure 1 illustrates such an elasticity matrix. This setting corresponds to a situation in which observables are constant within the estimation sample, while they can vary arbitrarily across assets outside the estimation sample.

In simple risk-based models like in Section 1.2, in which elasticities are proportional to the inverse of the covariance matrix, this means that all assets in the sample have the same variance and covariance with each other. It also corresponds to assuming that for any outside asset k, the covariance of its return with that of any asset in the sample is constant: $\text{cov}(R_i, R_k) = \text{cov}(R_j, R_k)$. In practice, outside assets are plentiful and this condition cannot be fully assessed. Still, one should present some corroborating evidence. For example, one can compute the covariances, or betas, with a set of broad portfolios for assets in the sample. This can take the form of a table of "balance on covariances," reporting these average covariances for treated and control assets, or high and low values of the instrument Z.

2.3.2 Groups of assets

Sometimes homogeneous substitution is not plausible across the whole sample, for example when the treatment affects an heterogeneous set of assets. Yet one might be able to delineate

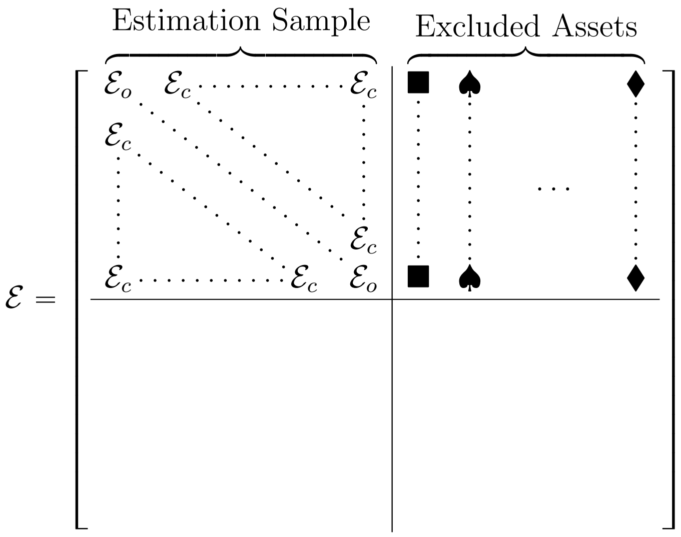


Figure 1: Elasticity matrix satisfying assumptions A1 and A2 for a local experiment.

groups of assets such that homogeneous substitution is plausible within each group. For example, homogeneity might hold for a set of firms in a narrow industry but not across these industries. In this case, it is possible to get estimates by pooling all groups while including group fixed effects to focus on within-group variation. Within the general setting of Assumption A1, this is the special case where the observables X_i are group dummies. The two-stage least squares regressions takes the form:

$$\mathit{\Delta} D_i = \widehat{\mathcal{E}} \mathit{\Delta} P_i + \theta_{g(i)} + \epsilon_i, \tag{24}$$

$$\mathit{\Delta} P_i = \lambda Z_i + \eta_{g(i)} + u_i. \tag{25}$$

Here, $g(i)$ denotes the group of assets (industries in our example) which contains asset i; θ_g and η_g are group fixed effects. Since this fits our framework, this regression with group fixed effects correctly identifies the relative elasticity. A special case of this situation arises in the nested logit model, but with stronger restrictions: in this model, substitution is completely symmetric across groups, while we allow arbitrary asymmetry through the matrix $\mathcal{E}_X$.

Chaudhary et al. (2022) explain how omitting group fixed effects in such a situation leads to biased inference. They document the relevance of this bias when measuring the effect of fund flows on corporate bond prices.

2.3.3 Heterogeneous risk exposures

When the investor cares about risk, the exposures of assets to different risks affects how they substitute. For example, in a downturn, when many asset prices fall, the demand for a more cyclical asset might change differently from that of a less cyclical asset. Empirical asset pricing often highlights many such sources of heterogeneity with risk factors. Even if we focus on a specific industry, different assets might have a different response to inflation or duration. These factors naturally affect patterns of substitutability, because they affect the covariance matrix of returns. In this case, an important continuous observable is the exposure or beta of the asset return to the factor.

To understand the mapping between factor models and our assumptions, assume there is a set of common factors F_t with loadings β, and constant idiosyncratic risk:

$$R_{i,t} = \beta_i' F_t + \nu_{i,t}, \quad \nu_i \perp \nu_j, \ \mathrm{var}(\nu_i) = \sigma^2_{idio}. \tag{26}$$

The corresponding covariance matrix is $\Sigma = \sigma^2_{idio}\mathbf{I} + \beta \Sigma_F \beta'$. In the mean-variance framework, the elasticity matrix will have the same structure:

$$\mathcal{E} = \gamma^{-1}\Sigma^{-1} = \widehat{\mathcal{E}}\mathbf{I} + \beta \Psi \beta', \tag{27}$$

where $\widehat{\mathcal{E}} = 1/(\gamma\sigma^2_{idio})$, $\beta = \left[1, \beta^{(f_1)}, \beta^{(f_2)}, \ldots, \beta^{(f_{K-1})}\right]$ is the set of factor loadings, and Ψ a $K \times K$ symmetric matrix.

The type of elasticity matrix in (27) satisfies assumptions A1 and A2. Intuitively, the relative elasticity for two assets with the same factor exposure depends only the amount of idiosyncratic risk and the investor's risk aversion. This is because the idiosyncratic component is the only risk taken when "arbitraging" between two assets that have the same risk profile. The factor structure matters for how investors respond to prices, but substitution is homogeneous for comparable assets. In practice, one might be reluctant to assume constant idiosyncratic volatility to ensure that assumption A2 is satisfied; Section 2.3.5 shows how to relax this condition.

Proposition 2 applies here, in that a regression controlling for β's, which are the factor loadings, recovers the relative elasticity. Alternatively, under the assumption that the betas are a function of characteristics, it is enough to control for the characteristics (Koijen and Yogo, 2019).

Synthetic controls. A variation of this approach particularly well-suited for event-study settings is to construct synthetic controls in the style of hedging portfolios. If one has a set of treated assets, one can construct portfolios of other assets as the control group for a difference-in-difference study. There are two requirements for this approach to be valid. First, the factor exposures of the control portfolio must be the same as that of the treated asset. Second, each asset in the control portfolio (as opposed to the combined portfolio returns) must have the same residual volatility as the treated asset.

2.3.4 Non-risk drivers of substitution

In practice, portfolio decisions respond to many other drivers than risk and return. Some investors care about non-pecuniary aspects of the stocks they hold, such as their carbon emissions or ESG characteristics. Mutual funds, pension funds, and endowments often operate under mandates that require a specific mix of assets, while others are guided by broader objectives outlined in their prospectus. When hedge funds take on leveraged positions, haircuts apply and they have to post margins. Banks and insurance companies must ensure that their portfolios satisfy various regulatory targets such as capital adequacy ratios, leverage requirements, or liquidity requirements.

All these dimensions affect which assets these investors choose in the first place, but also how they rebalance their portfolio when prices move. For example, if one of your more environmentally friendly stocks appears overpriced, you might shed this position and replace it with another similarly green position. Hence, these motives can play an important role in the elasticity matrix and must be taken into account when evaluating our assumptions.

To understand how, consider a generic representation of such a motive, by adding a quadratic cost and a linear constraint to the mean-variance optimization problem:

$$\max_{D} \quad D'(M-P) - \frac{\gamma}{2} D'\Sigma D - \frac{\kappa}{2}\left(D'X^{(1)}\right)^2 \tag{28}$$

$$\text{such that} \quad D'X^{(2)} \leqslant \Theta. \tag{29}$$

The quadratic cost $\kappa/2\left(D'X^{(1)}\right)^2$ captures smooth investment priorities: the more carbon-emitting stocks an investor holds, the less willing she is to hold additional carbon-emitting stocks. The variable $X^{(1)}$ measures the relative contribution of each asset to this total cost — e.g. its carbon emissions — while κ measures the overall willingness to hold carbon emitting stocks. The linear constraint represents hard targets, such as the liquidity ratio that a bank must hold. There, $X^{(2)}$ measures the contribution of each position to the constraint—e.g. its liquidity weight—and Θ is the maximum value capturing the regulatory requirement.

When prices move, such an investor will balance risk-return with reaching these other non-risk objectives. Hence, all these dimensions will shape substitution patterns. Clearly, to be able to use our results, the covariance matrix Σ has to satisfy the assumptions with respect to a set of observables $X^{(3)}$.[18] We show in Appendix C that the elasticity matrix for this investor satisfies assumptions A1 and A2 with respect to the stacked set of observables $X = [X^{(1)}, X^{(2)}, X^{(3)}]$.

Concretely, this result implies that the econometrician should first take a stand on the motives behind the investor's demand. Then, they should find the relevant observables that capture how each asset contributes to these motives; for example, the carbon emissions of a firm for a fund that has an ESG mandate. Finally, they should include these variables as controls in the instrumental variable regressions in order to recover the relative elasticity. The substitution driven by these motives is also interesting for its own sake, and Section 4 shows how to estimate it.

2.3.5 Entertaining more heterogeneity in elasticities

For some situations, the assumptions of constant relative elasticity or of homogeneous substitution might appear too restrictive. We provide two variations of the basic framework to accommodate these situations.

Observed heterogeneity in relative elasticity. First, we tackle the case of heterogeneity in relative elasticity. Just like Assumption A1 allows cross-elasticities to depend on observables, one can relax Assumption A2 to let the relative elasticity depend on observables. This corresponds to replacing the condition (11) by:

$$\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i) = \mathcal{E}_{relative}(X_i) = \mathcal{E}_r' X_i, \tag{30}$$

with $\mathcal{E}_r$ a vector of dimension K. For example, if the observable captures the size of a company, this relation allows big stocks to have a different relative elasticity than small stocks, an approach taken, for example, in Haddad et al. (2024). Another useful application is in the context of the factor models of Section 2.3.3. There, we have seen that idiosyncratic volatility controls the relative elasticity. Therefore, one could include the idiosyncratic volatility of each asset as an observable, or variables that proxy for this idiosyncratic volatility.

Intuitively, the setting with heterogeneous relative elasticity implies that relative demand responds not only to price changes but also to price changes interacted with the observables, all encoded in $\mathcal{E}_r$. As such, one must include in the regression and identify coefficients on all these components and provide instruments for each of them. Starting from an instrument Z_i for the change in price ΔP_i, one can construct instruments $Z_i X_i$ for its interaction

[18] Specifically, Σ has to be such that the elasticity matrix of the mean-variance problem without constraint and cost function satisfies assumptions A1 and A2.

with the observables $\mathit{\Delta} P_i X_i$. Then, under the exclusion restriction $Z_i X_i \perp \epsilon_i | X_i$ and the relevance conditions, the corresponding two-stage least square regression estimates $\mathcal{E}_r$. Appendix Section A.3 proves this result and provides all details on implementation.

Unobserved heterogeneity. A more complicated situation arises when there is unobserved variation in own- and cross-price elasticities across assets. Handling this case requires taking a stronger stance on variation in the instruments to maintain meaningful identification. Nothing comes for free: accommodating a more flexible elasticity matrix $\mathcal{E}$ is at the cost of stronger assumptions on the sources of variations needed for estimation. Specifically, one needs to assume independence of the instrument with respect to all unobserved sources of heterogeneity — a stronger condition than orthogonality with respect to the demand residual.

Unobserved heterogeneity in elasticities becomes relevant if one believes there is an amount of noise around Assumptions A1 and A2. It is possible that, even in a narrowly defined group like in Section 2.3.1, all assets are slightly different, with small variations in elasticities that have no apparent connection to the experiment at hand. Another case where unobserved heterogeneity is relevant is when the experiment uses pairs of assets that are different on many dimensions, but in a plausibly random way. An example of this case is index inclusions: the included and excluded assets from the index are closely related in size but might be in different industries or have different characteristics.

By formalizing the conditions necessary to handle unobserved heterogeneity, the next proposition pinpoints what the causal inference regression identifies in this case. For a formal proof, see Appendix A.6.

Proposition 3 *Assume that the data-generating process of the first stage follows:*

$$\mathit{\Delta} P_i = \lambda_i Z_i + u_i, \quad \textit{with } Z_i \textit{ independent of } (u_i, \lambda_i), \tag{31}$$

and that the instrument is independent of own- and cross-price elasticities as well as the demand residual

$$(\mathcal{E}_{ii}, \mathcal{E}_{ij}, \epsilon_i) | Z_i \sim (\mathcal{E}_{ii}, \mathcal{E}_{ij}, \epsilon_i). \tag{32}$$

Then, the two-stage least square estimation of equations (4) *and* (5) *without observables identifies the local average of the relative elasticity:*

$$\widehat{\mathcal{E}} = \frac{\mathbf{E}_i \left\{ \lambda_i (\mathcal{E}_{ii} - \mathbf{E}_j(\mathcal{E}_{ji})) \right\}}{\mathbf{E}_i(\lambda_i)}. \tag{33}$$

The two conditions state that the realization of the instrument is not only independent of variation in how the instrument transmits to prices (λ_i) but also how elasticities vary across assets. The proposition highlights that these assumptions lead to estimating an average value of relative elasticity, a form of local average treatment effect. With an added monotonicity condition that the instrument always affects prices in the same direction — all λ_i sharing the same sign — the estimate $\widehat{\mathcal{E}}$ will fall within the range of estimates in the sample. If one expects little variation in elasticity, this result indicates that heterogeneity will not create a large deviation from a situation with exactly constant elasticity. With a wider range of variation in relative elasticity, it becomes interesting to inspect the weights in the average formula. For assets in which the instrument has a greater impact on prices (large λ_i), their relative elasticities are given greater weights. For example, if more illiquid assets have both a higher impact of the instrument λ_i and investors trade them more inelastically (lower $\mathcal{E}_{ii}$), estimates of relative elasticity will be lower than the unweighted average relative elasticity, and overstate how inelastic the typical asset is.

2.4 Estimating elasticity in theoretical models

We take a brief detour through theoretical models. We first show how commonly used models, once considered in appropriate units, relate to our identification assumptions. Then, we provide an example explaining why simple equilibrium considerations do not affect our identification results.

2.4.1 Standard models of asset demand

We discuss how standard models of asset demand relate with the identification assumptions. For each model, we derive the appropriate units and parameter restrictions under which the demand regression is well specified.[19] Table 1 summarizes the results.

Constant absolute risk aversion. In the mean-variance model (CARA) described above, we have seen the direct mapping between covariance matrix and the elasticity matrix when considering a relation between the level of demand the level of prices: $\mathcal{E} = \partial D/\partial P = \gamma^{-1}\Sigma^{-1}$. Section 2.3.3 show that Assumptions A1 and A2 are satisfied if the covariance matrix has a factor structure with factor loadings which depend on the observables.

[19] Petajisto (2009), Davis et al. (2025), and Davis (2024) quantify elasticities in these models.

	CARA	CRRA	Logit
Regression units "demand" LHS	demand D_i	portfolio shares P_iD_i/W	log portfolio shares $\log(P_iD_i/W)$
Regression units "price" RHS	price P_i	log price $\log P_i$	log price $\log P_i$
Own Price Elasticity $\mathcal{E}_{ii}$	$\frac{R_f}{\gamma}[\Sigma^{-1}]_{ii}$	$\frac{1}{\gamma}[\Sigma^{-1}]_{ii}$	$\alpha(1-\omega_i)$
Cross price Elasticity $\mathcal{E}_{ij}$	$\frac{R_f}{\gamma}[\Sigma^{-1}]_{ij}$	$\frac{1}{\gamma}[\Sigma^{-1}]_{ij}$	$-\alpha\omega_j$
Relative Elasticity $\widehat{\mathcal{E}} = \mathcal{E}_{ii} - \mathcal{E}_{ji}$	$\frac{R_f}{\gamma}\left([\Sigma^{-1}]_{ii} - [\Sigma^{-1}]_{ji}\right)$	$\frac{1}{\gamma}\left([\Sigma^{-1}]_{ii} - [\Sigma^{-1}]_{ji}\right)$	α

Table 1: Three standard models of asset demands.

Constant relative risk aversion. Preferences with constant relative risk aversion (CRRA) are the workhorse model of macro-finance. Utility in this case is given by $u(C) = C^{1-\gamma}/(1-\gamma)$, with now γ being the constant relative risk-aversion. Assume that the risk-free rate is r_f and that there are N assets with payoffs $X = \{X_i\}_i$ at time 1, with prices $\{P_i\}$. Hence, asset returns are $R_i = X_i/P_i$.

To solve for the optimal demands, we assume that the payoffs follow a lognormal distribution: $\log X \sim \mathcal{N}(M, \Sigma)$ and log-linearize portfolio returns following Campbell and Viceira (2002).[20] For an investor with wealth W, the optimal demand is:

$$D_i = \frac{1}{\gamma}\frac{W}{P_i}\left[\Sigma^{-1}\left(M - \log P - r_f + \frac{1}{2}\operatorname{diag}(\Sigma)\right)\right]_i \tag{35}$$

This implies that when considering the relation between portfolio weights, $\omega_i = P_iD_i/W$, and log prices, the elasticity matrix is (this relation is exact in continuous time; see Duffie, 2010; He et al., 2025):

$$\mathcal{E} = \frac{\partial \omega}{\partial \log P} = -\frac{1}{\gamma}\Sigma^{-1}. \tag{36}$$

This is the same elasticity as the CARA case, albeit with different units: portfolio weights on log prices. Therefore, our earlier discussion relating properties of the covariance matrix (here of log returns) and the identification assumptions apply to this case as well.

[20] We log-linearize the return of portfolio ω, $r_p = \log R_p$ as:

$$r_p - r_f = \log\left(\omega' \exp\left(\mathbf{r} - r_f\right)\right) \simeq \omega'(\mathbf{r} - r_f) + \frac{1}{2}\omega' \operatorname{diag}(\Sigma) - \frac{1}{2}\omega'\Sigma\omega. \tag{34}$$

Logit. Koijen and Yogo (2019) introduce a model of portfolio demand of the logit form. They show existence of factor models giving rise to this demand for an investor with log utility. Unlike the model we just studied, these factor models feature a covariance matrix and expected returns that depend nonlinearly on prices, and hence have a different elasticity matrix; Appendix D details this distinction. The logit model is also commonly used in industrial organization, as well as in many fields in economics including trade and spatial equilibrium models. There, it is most often motivated by aggregation of a consumer discrete choice model, but can also apply to an individual's choice of consumption basket.[21]

For an investor with initial wealth W, the expenditure shares or portfolio weights are:[22]

$$\omega_i = \frac{P_i D_i}{W} = \frac{\exp\left(-\alpha p_i + \theta' X_i + \epsilon_i\right)}{1 + \sum_l \exp\left(-\alpha p_l + \theta' X_l + \epsilon_l\right)}, \tag{38}$$

where p_i is the log of the price of asset i, X_i observable demand shifters, and ϵ_i the unobserved component of demand.

When considering the relation between log portfolio weights and log prices, the elasticity matrix is:

$$\mathcal{E} = \frac{\partial \log \omega}{\partial \log P} = -\alpha \left(\mathbf{I} - \mathbf{1}\omega'\right), \tag{39}$$

where ω is the vector of portfolio weights given in (38). Note that $\mathcal{E}$ in general is not symmetric in this case. The coefficient α is the only demand parameter that determines the matrix of demand elasticity, as opposed to the whole covariance matrix in the CARA and CRRA cases. Further, this matrix always satisfies assumptions A1 ($\mathcal{E}_{jk} = \mathcal{E}_{ik} = \alpha\omega_k$) and A2 ($\mathcal{E}_{ii} - \mathcal{E}_{ji} = \alpha$), with α being the relative elasticity of demand.

Same relative elasticity vs. same elasticity matrix. Note that the fact that the simple risk-based models and the logit model can both satisfy the two assumptions only implies that they lead to the same estimation of the relative elasticity. Even when these models have the same value of relative elasticity, they exhibit different elasticity matrices. Figure 2 illustrates this nuance, with three distinct elasticity matrices that share the same relative elasticity.

[21] Anderson et al. (1988) derives the utility that leads to logit shares as optimal demand.

[22] If there is not outside asset, the model of expenditure shares becomes:

$$\omega_i = \frac{P_i D_i}{W} = \frac{\exp\left(-\alpha p_i + \theta' X_i + \epsilon_i\right)}{\sum_l \exp\left(-\alpha p_l + \theta' X_l + \epsilon_l\right)}. \tag{37}$$

$$\begin{bmatrix} \widehat{\mathcal{E}} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \widehat{\mathcal{E}} \end{bmatrix}$$

(a) Diagonal matrix.

$$\begin{bmatrix} \frac{1}{1-\rho}\widehat{\mathcal{E}} & \frac{\rho}{1-\rho}\widehat{\mathcal{E}} & \cdots & \frac{\rho}{1-\rho}\widehat{\mathcal{E}} \\ \frac{\rho}{1-\rho}\widehat{\mathcal{E}} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{\rho}{1-\rho}\widehat{\mathcal{E}} \\ \frac{\rho}{1-\rho}\widehat{\mathcal{E}} & \cdots & \frac{\rho}{1-\rho}\widehat{\mathcal{E}} & \frac{1}{1-\rho}\widehat{\mathcal{E}} \end{bmatrix}$$

(b) Symmetric matrix.

$$\begin{bmatrix} (1-\omega_1)\widehat{\mathcal{E}} & -\omega_2\widehat{\mathcal{E}} & \cdots & -\omega_N\widehat{\mathcal{E}} \\ -\omega_1\widehat{\mathcal{E}} & (1-\omega_2)\widehat{\mathcal{E}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\omega_N\widehat{\mathcal{E}} \\ -\omega_1\widehat{\mathcal{E}} & \cdots & -\omega_{N-1}\widehat{\mathcal{E}} & (1-\omega_N)\widehat{\mathcal{E}} \end{bmatrix}$$

(c) Logit-style matrix.

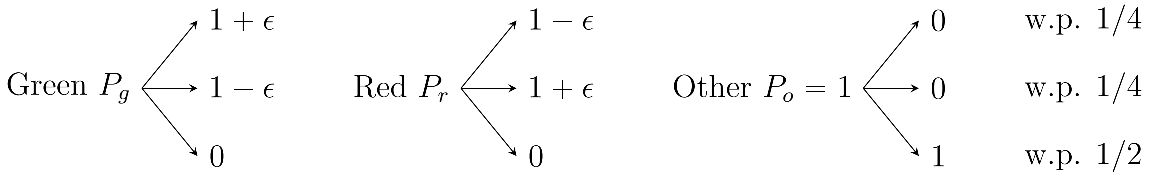
Figure 2: Different elasticity matrices with the same relative elasticity $\widehat{\mathcal{E}}$

2.4.2 What about equilibrium spillovers?

The reader might be surprised that, so far, we have not discussed the concept of equilibrium, which is usually central in asset pricing. This is not because we assume that the world is not in equilibrium: equation (9) is a change in demand in equilibrium. Instead, we can do so because identifying specific sources of variation in prices — the instrument Z_i — and assuming that an investor's demand satisfies Assumptions A1 and A2 is enough to estimate this investor's demand elasticity without understanding the entire structure of the equilibrium.

In this section, we work out a simple equilibrium model to illustrate this insight. The setting is inspired by Fuchs et al. (2025) who point out that endogenous cross-asset spillovers can imply a low measured own-price elasticity even if the true own price elasticity is near infinite. This result considers a different regression from the causal inference framework of this paper. We show that the issue arises because the omitted variable bias that we have pointed out in Section 1.3 is present. We also explain that, because the example satisfies our assumptions A1 and A2, a standard difference-in-difference regression is unbiased, once we recognize that it recovers the relative elasticity.

Setting. The economy is populated by a representative agent with log utility. There are three assets with different payoffs in three possible states of the world, with payoffs as follows:



The "other" asset acts as a numéraire, whose price is normalized to 1. Denote the prices of the green and red assets P_g and P_r. These two assets become closer substitutes as ϵ goes towards 0. Indeed, in the limit, any price difference between P_g and P_r represents an arbitrage opportunity. The representative agent has endowments E_g, E_r, and E_o, which implies that their wealth is $W = P_gE_g + P_rE_r + E_o$.

Demand and equilibrium. We can first derive the demand function, that is, the optimal portfolio share as a function of prices:

$$\omega_g\left(P_g, P_r\right) = \frac{P_g}{2}\frac{\epsilon^2(P_r+P_g)+(P_r-P_g)}{\epsilon^2(P_r+P_g)^2-(P_r-P_g)^2}, \quad \omega_r\left(P_g, P_r\right) = \frac{P_r}{2}\frac{\epsilon^2(P_r+P_g)+(P_g-P_r)}{\epsilon^2(P_r+P_g)^2-(P_r-P_g)^2}. \tag{40}$$

Market-clearing for the two assets, $\omega_gW = P_gE_g$ and $\omega_rW = P_rE_r$ lead to equilibrium prices as functions of the endowments:

$$P_g\left(E_o, E_g, E_r\right) = E_o\frac{\epsilon^2\left(E_g-E_r\right)-\left(E_g+E_r\right)}{\epsilon^2\left(E_g-E_r\right)^2-\left(E_g+E_r\right)^2}, \quad P_r\left(E_o, E_g, E_r\right) = E_o\frac{\epsilon^2\left(E_r-E_g\right)-\left(E_g+E_r\right)}{\epsilon^2\left(E_g-E_r\right)^2-\left(E_g+E_r\right)^2}. \tag{41}$$

As an initial equilibrium, we assume that the endowments are $E_g = E_r = 1/2$ and $E_o = 1$. It is then immediate that $P_r = P_g = 1$.

Demand elasticities. We can compute the demand elasticities: how individual demand would respond to a change in prices. Because utility is CRRA, we measure the sensitivity of portfolio shares to log prices (in line with Section 2.4.1) around the initial equilibrium values of prices:

$$\mathcal{E}_{own} = \frac{\partial \omega_g}{\partial \log P_g} = \frac{1}{8} - \frac{1}{8\epsilon^2}; \tag{42}$$

$$\mathcal{E}_{cross} = \frac{\partial \omega_g}{\partial \log P_r} = -\frac{1}{8} + \frac{1}{8\epsilon^2}. \tag{43}$$

The expressions for ω_r are identical. The relative elasticity is $\mathcal{E}_{own} - \mathcal{E}_{cross} = 1/4 - 1/(4\epsilon^2)$.

These measures show that when the two assets are near-identical, $\epsilon \to 0$, any deviation from parity would lead to a near-infinite increase in demand for the cheaper asset, and near-infinite decrease in demand for the more expensive one. This is the standard arbitrage argument.

Running regressions. We are interested in whether various regressions around a supply shock for one of the assets can identify these elasticities. A shift in supply of the green asset leads to the price changes

$$\frac{\partial \log P_g}{\partial E_g} = -\left(1+\varepsilon^2\right), \qquad \frac{\partial \log P_r}{\partial E_g} = -\left(1-\varepsilon^2\right), \tag{44}$$

around the equilibrium.

Fuchs et al. (2025) correctly point out that regressing the demand for the green asset on the change in its price using such a supply shock does not recover the own price elasticity. This regression corresponds to taking the ratio of the change in portfolio to the change in price across equilibria:

$$\frac{d\omega_g/dE_g}{d\log P_g/dE_g} = -\frac{1}{4}\frac{1-\varepsilon^2}{1+\varepsilon^2} \neq \mathcal{E}_{own}. \tag{45}$$

In particular, when $\varepsilon \to 0$, the regression coefficient on the left-hand-side converges to $-1/4$, while the own-price elasticity goes to infinity. Unpacking the total derivative explains the source of the bias:

$$\frac{d\omega_g/dE_g}{d\log P_g/dE_g} = \frac{\frac{\partial \omega_g}{\partial \log P_g}\frac{\partial \log P_g}{\partial E_g} + \frac{\partial \omega_g}{\partial \log P_r}\frac{\partial \log P_r}{\partial E_g}}{\frac{d\log P_g}{dE_g}} = \mathcal{E}_{own} + \mathcal{E}_{cross}\frac{\partial \log P_r/\partial E_g}{\partial \log P_g/\partial E_g} \tag{46}$$

The change in demand for the green asset is not driven only by the change in its own price but also by the change in price of its substitute the red asset, because $\mathcal{E}_{cross} \neq 0$. In the language of regressions, the price of the red asset is acting as a correlated omitted variable. Intuitively, the induced price drop of the green asset would lead to a large increase of its demand if the price of the red asset remained high. However, in equilibrium the price of the red asset drops too, resulting in only a moderate change in the demand for the green asset.

However, the canonical causal inference framework corresponds to a standard difference-in-difference regression in this setting with two assets. The regression coefficient is the ratio of the difference of change in portfolio weight to the difference of change in price, as in equation (6):

$$\frac{d\omega_g/dE_g - d\omega_r/dE_g}{d\log P_g/dE_g - d\log P_r/dE_g} = \frac{1}{4} - \frac{1}{4\varepsilon^2} = \mathcal{E}_{own} - \mathcal{E}_{cross}. \tag{47}$$

The difference-in-difference coefficient correctly identifies the relative elasticity. Indeed, this setting satisfies Assumptions A1 and A2: because the elasticity matrix is symmetric, substitution is homogeneous and the relative elasticity is constant. Also, because the endowment shock does not have a direct effect on log investors' choice of portfolio shares, the standard exogeneity condition is satisfied. Note that the identified relative elasticity is unbounded when ϵ converges to 0, in line with the economic intuition that the relative demand for near arbitrage assets should react strongly to a change in their relative price. The estimator leads to this limit because the relative change in portfolio remains finite, while the relative change in price goes to zero in this limit due to arbitrage.

This example highlights an important conceptual point: Demand elasticities are well defined regardless of the source of the change in prices. Precisely, the demand curve maps the price vector to the quantity vector through any source of change in price that are not accompanied by changes in other drivers of demand. As a result, it is not surprising that equilibrium spillovers are not a problem for the identification of demand elasticities per se. Instead, the econometrician has to be careful that prices of other assets might introduce omitted variable bias. Assumptions A1 and A2 ensure that this is not the case for a standard difference-in-difference estimator, which is then an unbiased estimator of relative elasticity.

3 Price Impact

The other natural application for causal inference in asset pricing is the estimation of price impact or multipliers. After setting up the corresponding regression framework, we show how our identification results apply to this situation. We then relate demand elasticity estimates and price impact estimates.

3.1 Price impact regression

Simple causal inference of price impact. Price impact measures how much prices change in response to an exogenous shift in demand. Because in equilibrium, aggregate demand does not change if assets are in fixed supply, the empirical setup differs from that of demand estimation. To understand the basic intuition, start with one asset to put aside issues of substitution. While equilibrium demand is fixed, it is possible for demand curves to shift; and we are interested in measuring the impact of such a shift. An idealized example

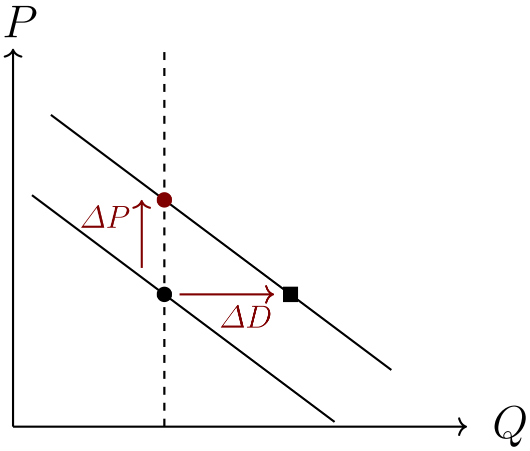


Figure 3: Equilibrium Effect of an Upward Shift in Demand Curve

would be an investor waking up in the morning and deciding to buy one share of Apple for no specific reason. Then, the aggregate demand curve for the asset shifts to the right by one unit. In equilibrium, the price must adjust upwards to satisfy market clearing. Figure 3 illustrates this process. Similarly supply shocks can be viewed as the negative of a demand shock and be treated likewise.

In practice, the econometrician starts from a shift of the demand curve Z_i (measured as the actual amount); examples of such shocks from the literature include asset purchases by central banks or rebalancing due to flows in and out of mutual funds (Lou, 2012). Armed with this shift in demand, we run the regression:

$$\Delta P_i = \widehat{\mathcal{M}} Z_i + \epsilon_i, \tag{48}$$

$$Z_i \perp \epsilon_i. \tag{49}$$

Of course, the shock Z_i is not the only source of variations in prices. Hence, there is still a stringent exclusion restriction in (49). In words, the change in demand under consideration must be orthogonal to any other demand shifts in the economy. For example, if a group of investors systematically mimicks the Fed's asset purchases, exogeneity is violated and the regression will be biased; the measured shock undercounts the actual change in demand, and overestimates the price impact.

There is no first stage because Z_i directly measures the magnitude of the shift in the demand curve. This shift does not materialize in equilibrium quantity demanded: prices adjust so that the total quantity demanded stays equal to the fixed supply. Equivalently, this identification condition corresponds to assuming that, if one could measure quantities before prices adjust — the out-of-equilibrium square in Figure 3 — the first-stage coefficient would be one.[23]

Handling substitution. The same issue as for the estimation of demand elasticity arises: all prices are determined together in equilibrium. All of the considerations discussed in Section 1 also apply. There is no such thing as the multiplier but instead a matrix $\mathcal{M}$ of own-demand and cross-demand multipliers such that $\mathit{\Delta} P = \mathcal{M} \mathit{\Delta} D$. This implies that in response to a vector of shocks Z, price changes will be:

$$\mathit{\Delta} P = \mathcal{M} Z + \epsilon, \tag{50}$$

where ϵ captures the impact of all other demand shocks. The following proposition highlights how assumptions A1 and A2 applied to the matrix $\mathcal{M}$ allow us to reach a valid estimation.

Proposition 4 *If the matrix $\mathcal{M}$ satisfies assumptions A1 and A2, and the demand shocks Z satisfy the exclusion restriction $Z_i \perp \epsilon_i | X_i$, the regression*

$$\mathit{\Delta} P_i = \widehat{\mathcal{M}} Z_i + \theta' X_i + u_i \tag{51}$$

identifies the relative multiplier:

$$\widehat{\mathcal{M}} = \mathcal{M}_{relative}. \tag{52}$$

The relative multiplier measures how the price of one asset relative to another comparable one changes if the relative demand for these assets shifts. How much does the price of Ford change relative to the price of General Motors if the demand for Ford changes relative to the demand for General Motors? To apply Proposition 4, the econometrician should argue that assumptions A1 and A2 are plausible for their experiment. In the next section we show that much of the intuition on the validity of the assumptions for elasticities translates directly to multipliers.

3.2 Link with elasticity estimation

Beyond the symmetry between the price impact regression and the demand elasticity regression, the two problems are intimately connected economically. Let us write the aggregate

[23] If the researcher is willing to make stronger assumptions on how or which investors are affected by the demand shock, they can run a first-stage converting an instrument into a demand shock for that group. However, the stronger assumptions correspond to assuming that the demand shock for the subgroup coincides with the aggregate demand shock, what we refer to as assuming a first-stage coefficient of 1.

demand curve $D(P)$, the sum of the demand curves of all agents in the economy. The corresponding elasticity matrix is $\mathcal{E} = \partial D / \partial P$. In equilibrium, prices have to be such that aggregate demand equals the aggregate supply S, such that $D(P) = S$. If demand curves shift by an amount $\varDelta D$, the new equilibrium price $P + \varDelta P$ satisfies $D(P + \varDelta P) + \varDelta D = S$. Using the implicit function theorem we obtain

$$\mathcal{M} = -\left(\frac{\partial D}{\partial P}\right)^{-1} = -\mathcal{E}^{-1}; \tag{53}$$

that is, the multiplier matrix is the inverse of the elasticity matrix.[24]

The next proposition connects estimation in the world of elasticities to the world of multipliers.

Proposition 5 *If the elasticity matrix $\mathcal{E}$ satisfies assumptions A1 and A2, then the multiplier matrix $\mathcal{M} = -\mathcal{E}^{-1}$ satisfies them as well. Furthermore, the relative elasticity and relative multiplier (both being scalars) are the inverse of each other:*

$$\widehat{\mathcal{M}} = -\widehat{\mathcal{E}}^{-1}. \tag{57}$$

The proposition has two parts, each proved in Appendix A.1. First, it states equivalence of assumptions A1 and A2 for $\mathcal{M}$ and for $\mathcal{E}$. This implies that the arguments of Section 2.3 for their validity also apply to the estimation of multipliers.[25]

Second, under these conditions the relative multiplier coincides with the inverse of the relative elasticity.[26] The two types of regression reveal the same information about demand. This conjunction occurs despite neither own-price and the cross-price elasticities being stable by inversion ($\mathcal{M}_{ij} \neq -1/\mathcal{E}_{ij}$); inverting a matrix is different from inverting each of its elements.

[24] As we describe in Section 2.4.1, it is sometimes more suitable to estimate demand elasticities in different units (logarithms, portfolio shares instead of quantities, ...). The inversion result of Proposition 5 applies to these different cases but with slightly adjusted formulas:

$$\mathcal{M}_{\{\log P, \log Q\}} = -\mathcal{E}^{-1}_{\{\log Q, \log P\}}, \tag{54}$$

$$\mathcal{M}_{\{\log P, \log Q\}} = -\left[\mathcal{E}_{\{\log \omega, \log P\}} - \left(\mathbf{I} - \mathbf{1}\omega'\right)\right]^{-1}, \tag{55}$$

$$\mathcal{M}_{\{\log P, \log Q\}} = -\left[\operatorname{diag}(\omega)^{-1}\mathcal{E}_{\{\omega, \log P\}} - \left(\mathbf{I} - \mathbf{1}\omega'\right)\right]^{-1}. \tag{56}$$

For example in the case of logit where demand elasticity is measured by regressing the log portfolio share on log price, equation (55) gives us the multiplier in log units: by how many percents do prices move in response to a one percent change in aggregate demand. Similarly, equation (56) is useful for the case of CRRA.

[25] Appendix A.2.4 shows that if the assumptions apply to each individual demand curve, it applies to the aggregate demand curve as well.

[26] Gabaix and Koijen (2021) derive this inversion result for a setting without observables shaping substitution.

3.3 Example: Relative multipliers in corporate bonds

To make things concrete, we consider a practical application. The goal is not to convince the reader that the assumptions hold perfectly. Instead, we sketch out the discussions one must go through at each step when conducting causal inference:

1. choose a source of variation,
2. assess exogeneity,
3. assess assumptions A1 and A2 and select observables,
4. implement the regression analysis.

Consider the market for investment-grade corporate bonds, broadly following Chaudhary et al. (2022). We obtain data on returns from the WRDS Bond Returns database between 2010 and 2022. To estimate the price impact of demand shocks, we need an exogenous source of variation in demand. One such source of variation is flow-induced trading from mutual funds as in Lou (2012). Flow-induced trading is the predicted demand shock from mutual funds' mechanical scaling of positions in response to flows:

$$Z_{it} = \sum_k \frac{A_{k,t-1} w_{i,k,t-1}}{P_{i,t-1} S_{i,t-1}} f_{kt}, \tag{58}$$

where $A_{k,t-1} w_{i,k,t-1}$ are the holdings fund k has of bond i at time $t-1$ (in dollars), the product of the fund's assets under management $A_{k,t-1}$ and portfolio weight $w_{i,k,t-1}$, f_{kt} denotes relative flows into fund k at time t, and $P_{i,t-1} S_{i,t-1}$ is the total bond supply for corporate bond i at $t-1$, the product of the bond's price and quantity outstanding. The instrument is constructed from mutual fund bond holdings and flows from the CRSP Survivor-Bias-Free US Mutual Fund Database.

The basic idea behind this instrument is that flows in and out of mutual funds are not related to the underlying details of the holdings of the fund. Aggregating these flows across all funds for a specific bond creates variation in demand for this bond. Weighting the flows by past portfolio shares removes the potential endogeneity due to selective trading by mutual funds. Furthermore, for the exclusion restriction to be respected, the flows into mutual funds should not be coming from investors who were already buying similar assets. If households are replacing portfolios held directly by similar portfolios inside mutual funds, there is actually no net shift in demand. Finally, the measured demand shocks should not be related to unobserved demand shocks. For example, the exclusion restriction would be violated if another type of institution, say insurance companies, would direct their investments to similar strategies as mutual fund investors.[27] To support the exogeneity condition of Proposition 4,

[27] Chaudhry (2025) shows that fund flows correlate with fund characteristics like growth, size, and income,

the empiricist should present empirical evidence and argue that these concerns are not in their data.

The next step is to gauge assumptions A1 and A2. Of course, these assumptions must be made jointly with a choice of observables.[28] Consider Assumption A1 first. It is clear that homogeneous substitution across bonds is unlikely to hold unconditionally; if demand for many long-term bonds rises, this will likely affect the price of other long-term bonds differently from the price of short-term bonds. As discussed in Section 2.3.1, a simple diagnostic for the plausibility of assumption A1 is a test of balance on covariances. One can ask: do the treated bonds comove in the same way with broad portfolios as control bonds?

Figure 4 suggests that this is not the case. For a given date, we form a long-short portfolio based on whether Z_{it} is above or below median on that date and compute the beta of this portfolio on a series of broad indices in a 2-year range around the date of the sort — the blue dot for that date. Each panel corresponds to a different index: a broad bond index, long-short portfolios based on credit ratings and maturity, and a broad stock index. Bonds with a high instrumented inflow appear to differ systematically from their low-inflow counterparts: they comove more strongly with the credit-rating-sorted portfolio and more weakly with the broad bond index and the duration-sorted portfolio. Such behavior is not surprising if investors choose their fund flows based on these dimensions. However, there is no meaningful difference in terms of exposure to the stock index.

Once we control for duration and credit rating as observables, covariances are much better balanced. Specifically, we construct a conditional instrument $Z_{idio,it}$ by residualizing Z_{it} with respect to duration and credit rating for each time period before sorting portfolios. This corresponds to the orange dots in Figure 4, which are much closer to 0. While this evidence bolsters Assumption A1, the empiricist should ask themselves whether other variables are likely to drive substitution across bonds before moving on. Relatedly, they should also ensure that treated and control bonds have similar properties, such as their idiosyncratic variance, to support Assumption A2.[29]

Provided the empiricist is convinced that the exclusion restriction and assumptions A1 and A2 hold, they can move on to the estimation of the relative multiplier. When facing repeated cross-sections, as in this setting, it is important to include time fixed effects in order to focus on cross-sectional variation. As such, column 2 of Table 2 estimates the relative

which suggests violation of the exclusion restriction. He also shows how to adjust the identification strategy to avoid this source of endogeneity.

[28] Note that while we have not emphasized it in the discussion above, the exclusion restriction is also conditional on observables.

[29] Appendix Figure 7 shows a similar picture for idiosyncratic volatilities as for the balance-on-covariance tests; without controlling for duration and credit rating, treated and control bonds have different idiosyncratic volatilities, but with controlling, they are close to identical.

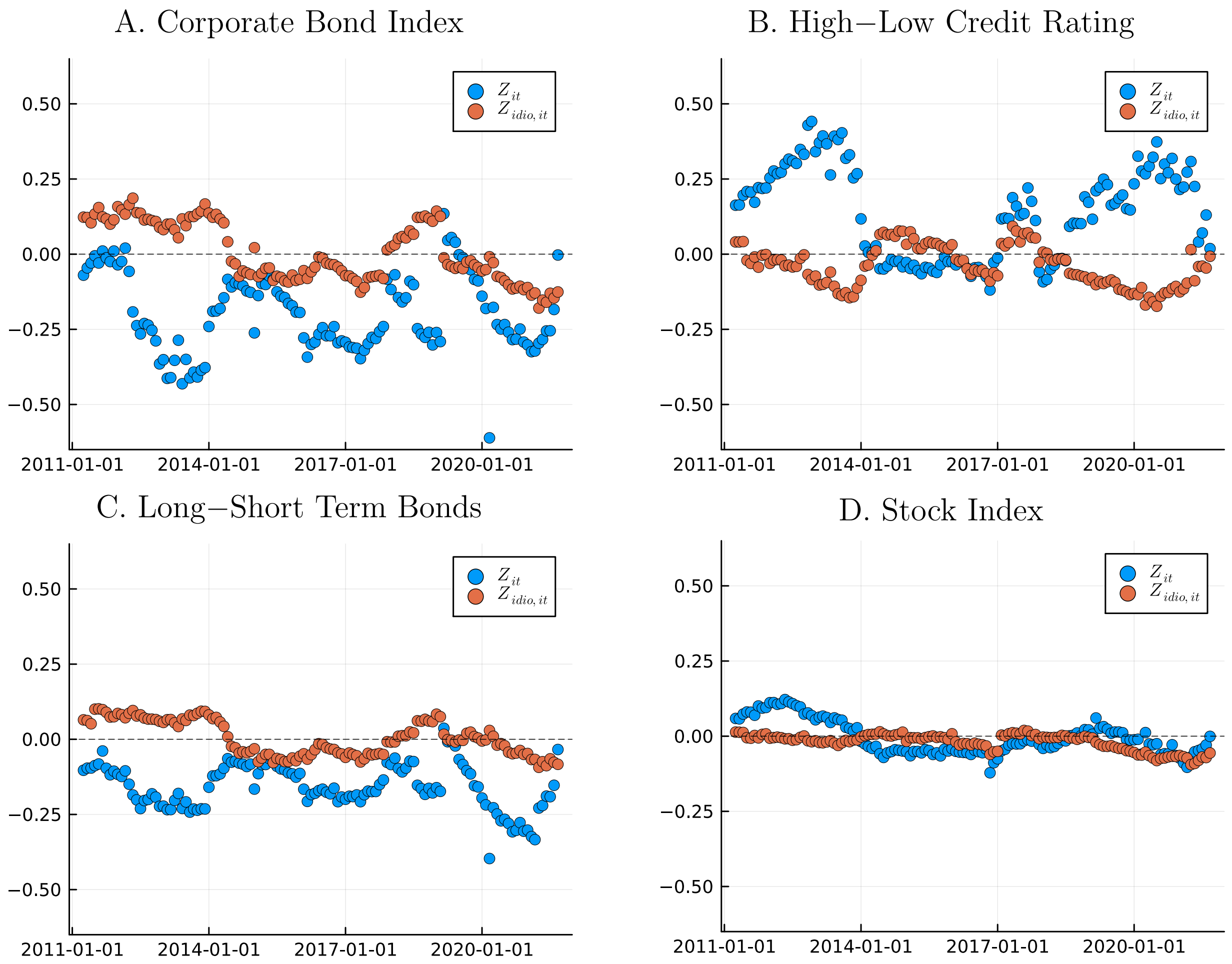


Figure 4: **Balance on covariances: exposure of long-short portfolios sorted on demand shocks to various factors.** Figure 4 reports regression coefficients from balance-on-covariance regressions based on both the raw demand shock Z_{it} (blue) and the demand shock $Z_{idio,it}$ (orange) that is cross-sectionally orthogonalized to duration and S&P credit ratings at each point in time. At each date, we form long–short equal-weighted portfolios based on whether Z_{it} (or $Z_{idio,it}$) is above or below the median. We compute the returns of these portfolios over two years centered around t, excluding t, and regress these returns on four aggregate factors. Panel A shows the time-series of coefficients for regressions on an aggregate investment-grade corporate bond factor, the ICE BofA US Corporate Index Total Return. Panel B uses the difference between aggregate high-yield and investment-grade corporate bond factors, the ICE BofA US High Yield Index Total Return and the ICE BofA US Corporate Index Total Return. Panel C uses the difference between the ICE BofA 15+ Year US Corporate Index Total Return and the ICE BofA 1-3 Year US Corporate Index Total Return. Panel D uses the Fama and French (1993) excess stock market return. The data for factors in panels A to C is from FRED, while the data for the excess market return in Panel D is from the Kenneth French data library. We exclude the bottom-quintile smallest bonds based on outstanding bond supply. The time series is from 2011:04 to 2021:09.

multiplier under the (implausible) assumptions of homogeneous substitution and constant relative elasticity without any observables. In column 3, we include controls for duration

Table 2: **Relative multiplier $\widehat{\mathcal{M}}$ in corporate bonds**

	Return $\mathit{\Delta} P_{it}/P_{i,t-1}$				
	(1)	(2)	(3)	(4)	(5)
Demand shock:					
Z_{it}	1.541*	-0.254	0.019		
	(0.637)	(0.229)	(0.065)		
$Z_{idio,it}$				0.019	0.019
				(0.065)	(0.065)
Date Fixed Effects		Yes	Yes	Yes	Yes
Duration × Date Fixed Effects			Yes	Yes	
Credit Rating × Date Fixed Effects			Yes	Yes	
N	646,335	646,335	646,335	646,335	646,335
R^2	0.010	0.415	0.632	0.632	0.415

Table 2 reports the results of relative multiplier regressions of bond returns $\mathit{\Delta} P_{it}/P_{i,t-1}$ on demand shocks Z_{it} and $Z_{idio,it}$ for U.S. investment-grade corporate bonds. Specifications (1)–(3) use the flow-induced trading demand shock Z_{it} defined in Equation (58). Specification (1) includes a common intercept, specification (2) uses date fixed effects, and specification (3) adds controls for a continuous duration variable and S&P credit rating dummies for each date. Specifications (4)–(5) use the demand shock $Z_{idio,it}$ orthogonalized to duration and credit rating each period, with and without controlling for duration and credit rating in the regression. We exclude the bottom-quintile smallest bonds based on outstanding bond supply. The sample period is 2010:04 to 2022:09. Standard errors are clustered by date and bond.

and credit-risk interacted with a time dummy; here again the idea is to control for these variables within each cross-section. In this specific case, the inclusion of these controls lead to a positive but insignificant price impact, unlike in the previous specifications. Concretely, the point estimate of 0.019 suggests that if the demand for one bond relative to another one with same credit rating and duration increases by 1%, this bond's price increases by 1.9bps relative to the other one. Columns 4 and 5 regress directly the change in price on the residualized instrument $Z_{idio,it}$, with and without controls for the characteristics. This leads to the exact same estimate of relative elasticity, a mathematical property independent of the specific dataset. This observation highlights that the source of variation for the estimates is variation in the residual component of the instrument $Z_{idio,it}$.

Again, the point of this section is not that these values constitute the best possible estimates in this setting, but merely to illustrate the process of using causal inference.

4 Beyond the Relative Elasticity

Standard cross-sectional causal inference can estimate the relative elasticity, a useful moment for answering micro questions comparing individual assets. There are many other interesting questions concerning more aggregated levels; here, aggregation is across assets. For instance, how do investors rebalance when the price of all small stocks changes relative to all big stocks, or when the price of long-duration bonds changes relative to short-duration ones? At the most aggregated level, what is the price impact of a demand shock for all assets at the same time?

This section aims to address these questions. Doing so hinges on estimating cross multipliers separately of the own multipliers. Estimating these dimensions of the multiplier matrix must rely on sources of variation in the time series, one for each characteristic driving substitution plus one for the overall aggregate. We focus on price impact estimation in the text; similar ideas apply to elasticity estimation.

4.1 A simple case of micro vs. macro

Consider the simple case of a symmetric multiplier matrix (as in Section 2.3.1): a constant own-price elasticity $\mathcal{M}_{own}$ and cross-price elasticity $\mathcal{M}_{cross}$. This simple configuration follows closely the analysis in Gabaix and Koijen (2021) to illustrate the basic principles behind answering aggregate questions before we can move on to a discussion of richer substitution with observables in the next subsection.

We first show that in this situation, the response of prices to demand can be decomposed into two distinct components: a composition effect and an aggregate effect. The composition effect corresponds to the relative comparisons obtained with standard cross-sectional inference. In contrast the aggregate effect is a component we have not explored yet.

Proposition 6 (Multiplier decomposition in symmetric case) *Take a multiplier matrix $\mathcal{M}$ with constant own-multiplier $\mathcal{M}_{own}$ and cross-multiplier $\mathcal{M}_{cross}$. Consider a generic change in demand and price connected by this matrix, so that $\Delta P = \mathcal{M}\Delta D$. Define the aggregate and idiosyncratic components of the changes in price and demand:*

$$\Delta P_{agg} = \frac{1}{N}\sum_i \Delta P_i, \qquad \Delta D_{agg} = \frac{1}{N}\sum_i \Delta D_i, \tag{59}$$

$$\Delta P_{idio,i} = \Delta P_i - \Delta P_{agg}, \qquad \Delta D_{idio,i} = \Delta D_i - \Delta D_{agg}. \tag{60}$$

The response of prices to a change in demand can be decomposed into univariate components:

$$\text{Micro:} \qquad \mathit{\Delta} P_{idio,i} = \widehat{\mathcal{M}}\, \mathit{\Delta} D_{idio,i}, \tag{61}$$

$$\text{Macro:} \qquad \mathit{\Delta} P_{agg} = \overline{\mathcal{M}}\, \mathit{\Delta} D_{agg}, \tag{62}$$

where $\widehat{\mathcal{M}} = \mathcal{M}_{own} - \mathcal{M}_{cross}$ *is the relative multiplier and* $\overline{\mathcal{M}} = \mathcal{M}_{own} + (N-1)\mathcal{M}_{cross}$ *is the aggregate multiplier.*

Aggregate and composition effects. Equation (61) captures the relative comparison between assets. If an asset experiences a higher demand shock relative to the average, its price increases relative to the average. The magnitude of this response is determined by the relative multiplier $\widehat{\mathcal{M}}$. In contrast, equation (62) captures the change in aggregate price, which only depends on the change in aggregate demand. There, the strength of the response is captured by the aggregate multiplier $\overline{\mathcal{M}}$. Interestingly, those two components are separate. The composition of the demand shock has no bearing on the aggregate price. Conversely, the aggregate shift in demand affects the price of all assets equally.

This setting is consistent with the common approach used in macroeconomics to focus only on relations between aggregates. For example, Gabaix and Koijen (2021) present a model of the aggregate multiplier where the only asset is "the stock market" without tracking individual stocks. Proposition 6 shows that such a model generalizes to an arbitrary composition of demand shocks to individual stocks under the symmetry assumption of this section.

Separating different multipliers. The second takeaway from Proposition 6 is that the aggregate multiplier $\overline{\mathcal{M}}$ cannot be calculated from the relative multiplier $\widehat{\mathcal{M}}$. It is not only that the two multipliers have different magnitudes, but also that one cannot be recovered from the other one. It is immediate to see this result in such a simple setting: the two multipliers represent different linear combinations of own- and cross-multipliers $\mathcal{M}_{own}$ and $\mathcal{M}_{cross}$. We will show this distinction remains in richer settings.

Estimating the aggregate multiplier. This distinction also has important implications for estimation. We have already shown that the cross-section allows to recover $\widehat{\mathcal{M}}$. Equations (61) and (62) show that it is impossible to recover anything else than $\widehat{\mathcal{M}}$ from the cross-section alone. The aggregate component $\overline{M}$ is only contained in the intercept of the regression, which cannot be guaranteed to be exogenous. This is the classic missing intercept problem. Because relative and aggregate multiplier are transformations of own- and cross-multipliers, this observation also implies that one cannot separate own- and cross-multipliers from the cross-section alone.

Then, how can we estimate $\overline{\mathcal{M}}$? The only way is to have a series of observations over time, and then use exogenous variation in demand in the time series. Concretely, one needs a variable Z_t such that other demand shifters are orthogonal to this instrument: $Z_t \perp (D_{agg,t} - Z_t)$.[30] Then a time series regression of P_t on Z_t correctly recovers $\overline{\mathcal{M}}$:

$$\Delta P_{agg,t} = \overline{\mathcal{M}}\, Z_t + v_t \tag{63}$$

$$\text{with } Z_t \text{ such that } \Delta D_{agg,t} = Z_t + \epsilon_t, \text{ with } \epsilon_t \perp Z_t. \tag{64}$$

Equation (62) highlights that not only is a single time-series necessary for this regression but also that no additional information comes from observing the entire panel. Arguments about exogeneity of the source of variation must be about time-series variation.

4.2 Decomposition into micro, meso, and macro levels

In practice, substitution across assets is unlikely to be symmetric. For example when investors substitute across bonds, they care about the maturity profile of their portfolio. The observables in our framework (the variables X in assumptions A1 and A2) capture this dimension of heterogeneity beyond the relative elasticity. In this case, there is still a decomposition between micro and aggregated multipliers, with one important distinction: there are multiple aggregated multipliers corresponding to each source of substitution and the overall aggregate.

To make this point concrete, consider a single observable X which is standardized. The following proposition gives the decomposition at micro, meso, and macro levels; we discuss the general case later in Section 4.4.

Proposition 7 (Multiplier decomposition with observables) *Take a multiplier matrix $\mathcal{M}$ satisfying Assumptions A1 and A2 with an observable X which has mean zero and unit variance in the cross-section. Consider a generic change in demand and price connected by this matrix: $\Delta P = \mathcal{M}\Delta D$. Define the aggregate, idiosyncratic, and X-based components of*

[30] For example Gabaix and Koijen (2021) construct such shifters using granular instruments (Gabaix and Koijen, 2024) across investors.

the changes in price and demand:

$$\mathit{\Delta} P_{agg} = \frac{1}{N}\sum_i \mathit{\Delta} P_i, \qquad \mathit{\Delta} D_{agg} = \frac{1}{N}\sum_i \mathit{\Delta} D_i, \tag{65}$$

$$\mathit{\Delta} P_X = \frac{1}{N}\sum_i X_i \mathit{\Delta} P_i, \qquad \mathit{\Delta} D_X = \frac{1}{N}\sum_i X_i \mathit{\Delta} D_i, \tag{66}$$

$$\mathit{\Delta} P_{idio,i} = \mathit{\Delta} P_i - \mathit{\Delta} P_{agg} - X_i \mathit{\Delta} P_X, \qquad \mathit{\Delta} D_{idio,i} = \mathit{\Delta} D_i - \mathit{\Delta} D_{agg} - X_i \mathit{\Delta} D_X. \tag{67}$$

The response of prices to a change in demand can be decomposed into three univariate components:

$$\text{Micro:} \qquad \mathit{\Delta} P_{idio,i} = \widehat{\mathcal{M}} \mathit{\Delta} D_{idio,i} \tag{68}$$

$$\text{Meso:} \qquad \mathit{\Delta} P_X = \widetilde{\mathcal{M}}_{agg} \mathit{\Delta} D_{agg} + \widetilde{\mathcal{M}}_X \mathit{\Delta} D_X, \tag{69}$$

$$\text{Macro:} \qquad \mathit{\Delta} P_{agg} = \overline{\mathcal{M}}_{agg} \mathit{\Delta} D_{agg} + \overline{\mathcal{M}}_X \mathit{\Delta} D_X, \tag{70}$$

where the scalar coefficients $\widehat{\mathcal{M}}, \widetilde{\mathcal{M}}_{agg}, \widetilde{\mathcal{M}}_X, \overline{\mathcal{M}}_{agg}$, and $\overline{\mathcal{M}}_X$ map one-to-one to the matrix $\mathcal{M}$.

Proposition 7 shows that the presence of the observable X breaks down the dichotomy between a micro- and macro-multiplier explaining all the impact of the price. It adds an intermediate "meso" layer, with a distinct role for fluctuations along this variable. This is captured by the aggregates $\mathit{\Delta} P_X$ and $\mathit{\Delta} D_X$. These quantities measure how the price and the demand for assets with larger values of X change relative to those with lower values. Indeed, because we assume the observable has mean zero in the cross-section, $\mathit{\Delta} P_X$ is the change in price of a long-short portfolio sorted on this characteristic.

However, this is in general not just a third, intermediate, layer. Idiosyncratic asset-level changes in prices and demand remain autonomous — equation (68). In contrast, meso and macro price impacts — equations (69) and (70) — are linked together. The shift in demand along X, $\mathit{\Delta} D_X$, affects the aggregate price $\mathit{\Delta} P_{agg}$. Conversely, the aggregate shift in demand $\mathit{\Delta} D_{agg}$ affects the relative price of assets along X, $\mathit{\Delta} P_X$. This connection yields additional challenges for estimation and interpretation of multipliers. We first discuss the implications for estimation of the macro multiplier, then turn to the meso multiplier.

4.3 Estimating the macro multiplier

Defining the macro multiplier. The first observation is that a more precise definition of the macro multiplier is required in this setting. Equation (70) highlights that the aggregate change in demand $\mathit{\Delta} D_{agg}$ is not the only determinant of changes in the aggregate price.

The composition of this shift matters as well. This is the second term, proportional to ΔD_X. Going back to the example of bonds of different maturity, a shift in the supply of all Treasuries might have a different effect on the total value of government debt than a disproportionate shift in the supply of long-term treasuries.

There is still a natural definition of the macro multiplier, the coefficient $\overline{\mathcal{M}}_{agg}$. This number represents the price impact of a parallel shift in demand for all bonds. More generally, it represents the price impact of a shift in demand with composition orthogonal to the observable X. Another version of this result is that one can focus only on aggregate price and demand if they assume that all shocks are proportional across assets.

Conditions for identification of the macro multiplier. The impact of composition effects creates a potential additional omitted variable in the estimation of the macro multiplier relative to the simple case of Section 4.1. Therefore, an additional condition is required for identification. A candidate shock Z_t must be orthogonal to both other aggregate demand shocks and all composition-based shocks. This corresponds to the regression specification:

$$\Delta P_{agg,t} = \overline{\mathcal{M}}_{agg} Z_t + v_t, \tag{71}$$

$$\text{with } Z_t \text{ such that } \begin{cases} \Delta D_{agg,t} = Z_t + \epsilon_t, \text{ with } \epsilon_t \perp Z_t; \\ Z_t \perp \Delta D_{X,t}. \end{cases} \tag{72}$$

Verifying the identification conditions. In practice, how can we ensure this additional condition is satisfied? Instruments for aggregate demand often come from specific investors or groups of investors, for which we know the reason behind their trading. This feature makes it plausible that any shift in the demand curve of other investors is orthogonal to the instrument. Without composition effects, this property is only needed for the demand of other investors for the aggregate portfolio. In our current setting with composition effects, it also needs to apply to their demand for the long-short portfolio based on X.

However, this is not enough. It is also necessary that the shock to the initial investors' aggregate demand is orthogonal to their own shock of demand for the long-short portfolio. Concretely, the econometrician should look at their shock and evaluate whether it only creates a parallel shift in portfolios. For example, a central bank can suddenly decide to intervene and purchase corporate bonds, like the Federal Reserve in 2020 (Haddad et al. (2021)). If the purchase policy is tilted towards a certain category of bonds, such as investment grade, the shock also creates variation in the demand along that characteristic, $\Delta D_X \neq 0$. More broadly, one should always ask whether the composition of aggregate shocks is related to important observables.

If the econometrician cannot confirm the condition that Z_t is orthogonal to $\Delta D_{X,t}$, the alternative way to make progress is to find two separate sources of variations with known impact on aggregate demand and on the demand along X:

$$\Delta D_{agg,t} = Z_t^{(1)} + \lambda Z_t^{(2)} + \epsilon_{agg,t}, \tag{73}$$

$$\Delta D_{X,t} = \mu Z_t^{(1)} + Z_t^{(2)} + \epsilon_{X,t}, \tag{74}$$

$$\text{with} \quad (Z_t^{(1)}, Z_t^{(2)}) \perp (\epsilon_{agg,t}, \epsilon_{X,t}), \tag{75}$$

where the cross-impact of the shocks, λ and μ, are known and $\lambda \neq \mu^{-1}$ so that there is independent variation between ΔD_{agg} and ΔD_X.

It is tempting to follow the traditional asset pricing approach and simply control for the factor return based on the characteristic X. For example, when estimating the aggregate multiplier for the stock market, one could control for the returns on the factors of Fama and French (1993) say HML and SMB. Unfortunately this path is flawed because factor returns might also respond to aggregate demand shocks, as shown in equation (69).

4.4 Estimating meso multipliers

The direct meso multiplier. Variation along the observables is also interesting for its own sake. How does a change in demand for green stocks relative to brown stocks affects the relative price of these two groups of assets? Answering this question corresponds to estimating $\widetilde{\mathcal{M}}_X$ for X being a variable measuring the greenness of a firm. How does a quantitative easing operation purchasing long-term bonds by issuing short-term reserves lowers the term premium? Again, this is the coefficient $\widetilde{\mathcal{M}}_X$, this time for X measuring duration.

Proposition 7 demonstrates clearly that the answer to these meso questions are not provided by the micro multiplier $\widehat{\mathcal{M}}$ or macro multiplier $\overline{\mathcal{M}}_{agg}$, nor by combining them. Instead, it reflects how investors substitute across assets precisely along the characteristic of interest.

The symmetry between the meso and aggregate multipliers— equations (69) and (70)— implies that all the discussion above for the macro multiplier applies to the estimation of $\widetilde{M}_X$. One must first find a shock in the time series that shifts the demand for high-X assets relative to low-X assets. This shock must be orthogonal to other demand shifts along this characteristic and the shift in aggregate demand:

$$\Delta P_{X,t} = \widetilde{\mathcal{M}}_X Z_{X,t} + v_{X,t} \tag{76}$$

$$\text{with } Z_{X,t} \text{ such that } \begin{cases} \Delta D_{X,t} = Z_{X,t} + \epsilon_{X,t}, \ \text{with } \epsilon_{X,t} \perp Z_{X,t} \\ Z_{X,t} \perp \Delta D_{agg,t}. \end{cases} \tag{77}$$

For example a shock that increases the demand for long-term bonds and reduces the demand for short-term bond by the same amount could be a valid $Z_{X,t}$, because it results in no shift in aggregate demand. In contrast, a shock to the demand for long-term bonds only would both create a shift along the observable and a shift in overall demand, violating the exclusion restriction.

By measuring the impact of changes in prices along the observables, the meso multiplier $\widetilde{\mathcal{M}}_X$ captures the substitution that was missing in the cross-sectional regression, as we discussed in Section 2.2. The limitation of the cross-sectional regression is immediate in the context of the identification conditions we just developed: within one period in the cross-section, there is no variation in the demand $\Delta D_{X,t}$. Analogous to the fact that the constant in the regression does not reveal the macro multiplier (Gabaix and Koijen, 2021), this issue is an incarnation of the well-known missing intercept problem for cross-sectional identification.[31]

Cross-multipliers. The connection between the meso and macro level also implies the existence of interesting cross-multipliers. The coefficient $\widetilde{\mathcal{M}}_{agg}$ captures how an aggregate shock to demand affects the relative price of assets with different values of the observable X. Conversely, the coefficient $\overline{\mathcal{M}}_X$ measures the aggregate effect of relative demand shock along the characteristic X. For example, it measures how an "operation twist" affects the total valuation of all debt.

The estimation of these cross-terms requires having two separate demand shocks that hit the two dimensions, as in equations (73)–(75). By including both shocks in the macro and meso price impact regressions, one can separate the direct effect of each type of shocks from their cross-effects.

With multiple observables. In Appendix A.4, we generalize Proposition 7 to the case of an arbitrary set of observables. In this case, one must track a greater set of aggregate price and demand indices: an overall aggregate price, and an index along each dimension of the observables. These indices appear as the coefficient of a regression of prices and demands on the observables. For the demand indices, this corresponds to defining

$$\Delta D_i = \Delta D_{agg} + \sum_{k=2}^{K} X_i^{(k)} \Delta D_{X,k} + \Delta D_{idio}, \tag{78}$$

$$\text{with} \quad \Delta D_X = (X'X)^{-1} X' \Delta D, \tag{79}$$

[31] See for example Wolf (2023) in the context of macroeconomics.

and the first component of ΔD_X is ΔD_{agg}.

Then, the decomposition result is that each of the price indices responds to all of the demand indices: $\Delta P_X = \breve{\mathcal{M}} \Delta D_X$, where $\breve{\mathcal{M}}$ is a $K \times K$ matrix. All of the discussion above regarding identification for the single observable case generalize immediately. For example, to completely identify the matrix $\breve{\mathcal{M}}$, one needs a set of K demand shocks in the time series.

Using demeaned observables leads to easier interpretation. In this case, the aggregate index is the average change in demand $\Delta D_{agg} = N^{-1} \sum_i \Delta D_i$.[32] The coefficient $\breve{\mathcal{M}}_{11}$ represents a well-defined notion of macro multiplier: the response of the overall level of prices to a parallel shift in demand. For the observables, the other components $\Delta P_{X,k}$ represent the change in price of a long-short portfolio formed along the k-th characteristic with no tilt along the other characteristics. This is analogous to the coefficient in a Fama and MacBeth (1973) regression. The diagonal term $\breve{\mathcal{M}}_{kk}$ measures the direct meso multiplier: how this relative price responds to a shift in demand along this characteristic only. The off-diagonal components measure spillovers between the various meso components and the aggregate component.

An alternative approach: assuming symmetric groups. A simple way to make assets comparable at the individual level is to classify them in disjoint groups, as in Section 2.3.2. This approach corresponds to the observables X being dummy variables for each group — in this case, the superfluous constant should be removed. The price and demand aggregates of equation (79) are the averages for each group. This is equivalent to working directly with data that is aggregated at the group-level, and estimating a multiplier matrix $\breve{\mathcal{M}}$ of size $K \times K$, with K the number of groups.

To fully estimate $\breve{\mathcal{M}}$ requires an instrument for each group. Instead of estimating the matrix of multipliers with one source of time-series variation for each group, one can go back to causal inference as in Section 2, by making assumptions A1 and A2 about $\breve{\mathcal{M}}$. Specifically, one can assume homogeneous substitution and constant relative elasticity *across* groups. Then, a single source of exogenous variation in the cross-section of groups allows to estimate the relative multiplier across groups $\breve{\mathcal{M}}_{relative}$. A special case of such structure is the nested logit model, used for example in Fang (2023) and Koijen and Yogo (2024).

Importantly, making assumptions A1 and A2 in the context of more aggregated data is more restrictive than at the micro level. Consider for instance a setting where investors manage the duration of their portfolio, hence duration affects substitution across bonds. In this context, if one groups bonds by duration, assumptions A1 and A2 hold at the asset level (with the group dummies as observables) but not at the group level.

[32] This occurs because the constant in a regression on mean zero variables (equation (79)) is simply the average of the dependent variable.

Table 3: **Macro- and meso multipliers in corporate bonds**

	Return $\Delta P_{agg,t}/P_{agg,t-1}$		Return $\Delta P_{X,t}/P_{X,t-1}$	Return $\Delta P_{it}/P_{i,t-1}$	
	(1)	(2)	(3)	(4)	(5)
$Z_{agg,t}$	14.231***	12.347**	7.294**	12.347**	12.347**
	(3.643)	(3.985)	(2.423)	(3.959)	(3.958)
$Z_{X,t}$		-6.170	0.817	-6.170	-6.170
		(7.810)	(4.591)	(7.757)	(7.757)
$Z_{agg,t} \times X_{it}$					7.294**
					(2.407)
$Z_{X,t} \times X_{it}$					0.817
					(4.558)
$Z_{idio,it}$				0.090	0.090
				(0.055)	(0.054)
Duration X_{it}				0.001	-0.001
				(0.001)	(0.001)
N	150	150	150	646,335	646,335
R^2	0.242	0.250	0.135	0.101	0.125

Table 3 reports the results of macro- and meso multiplier regressions of bond returns on demand shocks for U.S. investment-grade corporate bonds. Specification (1) follows equation (63) in estimating the macro multiplier by regressing aggregate bond returns $\Delta P_{agg,t}/P_{agg,t-1}$ on the aggregated instrument $Z_{agg,t}$ in the time series. Specification (2) jointly estimates the macro multiplier $\overline{\mathcal{M}}_{agg}$ and a cross-multiplier $\overline{\mathcal{M}}_X$ from equation (70) by adding the aggregated duration-tilted shock $Z_{X,t}$. Conversely, specification (3) jointly estimates the meso multiplier $\widetilde{\mathcal{M}}_X$ and cross-multiplier $\widetilde{\mathcal{M}}_{agg}$ from equation (69). Specifications (4) and (5) estimate the mechanically identical macro- and meso-level multipliers as in specifications (2) and (3) using disaggregated, repeated cross-sectional regressions, while adding the relative multiplier $\widehat{\mathcal{M}}$. We exclude the bottom-quintile smallest bonds based on outstanding bond supply. The sample period is 2010:04 to 2022:09. Robust standard errors are used for specifications (1) to (3). For specifications (4) and (5), standard errors are clustered by date and bond, and regressions are weighted such that each date receives equal weight.

4.5 Example: Duration-based multipliers in corporate bonds

To illustrate concretely the process for estimating the aggregated multipliers, we return to the setting of Section 3.3 where we have focused on price impact in investment-grade corporate bonds, using flow-induced trading from mutual funds as the instrument $Z_{i,t}$.

We start by considering the macro multiplier. In the simple approach of Section 4.1, Proposition 6 decomposes the multiplier matrix into only two distinct components: a micro and macro multiplier. The macro multiplier $\overline{\mathcal{M}}$ measures the response of aggregate prices to a change in aggregate demand. Thus, to estimate $\overline{\mathcal{M}}$ one needs a source of variation in aggregate demand which can only come from the time series. If one would like to parallel the micro estimation, they would use aggregate flows from mutual funds. This corresponds to aggregating the micro-level instrument: $Z_{agg,t} = N^{-1} \sum_i Z_{it}$. While this aggregate instrument is based off the micro-level one, the exclusion restriction is different. To satisfy this condition, the aggregate mutual fund flow must not be driven by a response to prices, and cannot be related to shifts in the demand curves of other investors such as banks. Supporting the exclusion restriction would be a tall order for this shock. Column 1 of Table 3 estimates the macro multiplier by regressing changes in aggregate price on this shock to aggregate demand. The estimate suggests that a 1% increase in demand leads to a 14% increase in bond prices.[33]

A potential concern for this estimation of the macro multiplier is that meso-level demand shocks also affect aggregate prices, and that these shocks are correlated with the aggregate instrument. For the sake of simplicity we narrow the analysis down to shocks along one observable, duration, and abstract from variation in credit risk. The concern is that the instrument for aggregate demand, $Z_{agg,t}$, is correlated with shocks to the demand for long-term bonds relative to short-term bonds. If you have already argued that the instrument is independent of changes in the aggregate demand of other investors, then it is natural to assume that the instrument is also uncorrelated with their demand for long-term bonds relative to short-term bonds. But aggregate flows from mutual funds themselves are in general not uniform across bonds; instead they tilt towards either short- or long-term bonds. In fact, constructing the demand shock along duration, $Z_{X,t} = N^{-1} \sum_i X_{i,t} D_{i,t}$, with $X_{i,t}$ being each bond's duration demeaned and standardized for each time period, we find a correlation of -0.59 between $Z_{agg,t}$ and $Z_{X,t}$ in the data.

In light of this substantial relation, it is necessary for researchers to account for the role of the meso-level shock as in equations (73) to (75). Of course, doing so raises the bar on exogeneity because $Z_{X,t}$ must also be unrelated to other demand shocks, and a defense of this assumption along the same lines as for $Z_{agg,t}$ must be provided. Column (2) presents the result of the estimation accounting for both meso and macro demand shocks. In this case, the estimate of the macro multiplier does not change much, because meso-level shocks appear to not have a strong effect on aggregate prices.

The response of the price of long-term bonds relative to short-term bonds, $\Delta P_{X,t}$, to shifts in aggregate demand $\Delta D_{agg,t}$ and demand along durations $\Delta D_{X,t}$ is interesting in its own right. We regress $\Delta P_{X,t}$ on the two demand shocks $Z_{agg,t}$ and $Z_{X,t}$ in column (3) of Table 3. The identification condition is the same as for the previous regression. Across bonds of different durations, a 1% increase in the aggregate demand for bonds leads to a 7.3% higher return for each standard deviation. A shift in demand away from short-term bonds towards long-term bonds of 1% leads to a positive but insignificant increase in the relative price of

[33] This estimate for the macro multiplier in corporate bonds is unusually large compared to the literature studying these multipliers carefully and find values between 2.3 and 6 (e.g., Bouveret and Yu, 2021; Chaudhary et al., 2022; Darmouni et al., 2023); see Haddad et al. (2025) for a review.

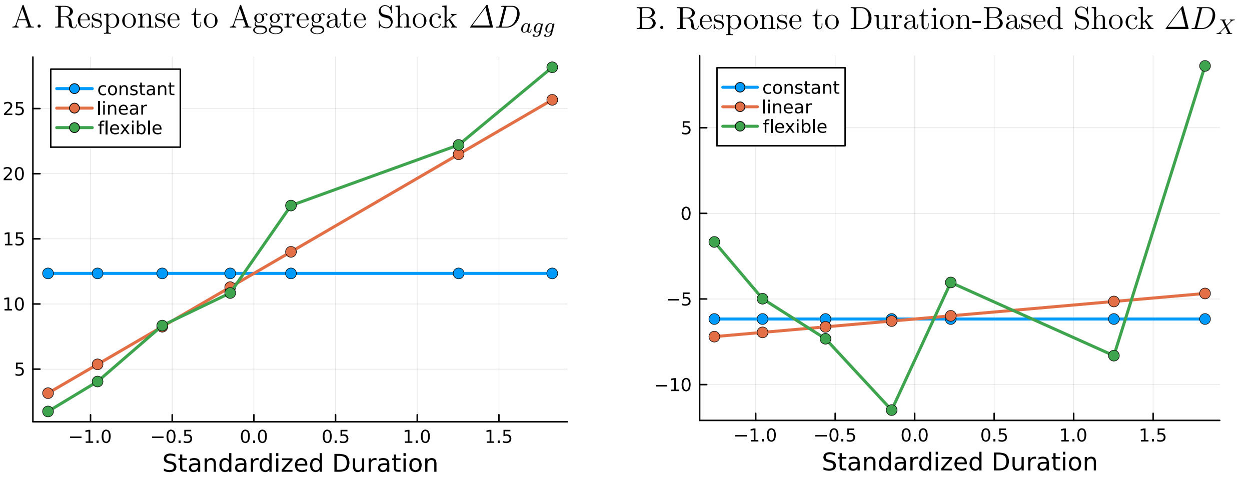


Figure 5: **Macro- and meso multipliers across durations.** Figure 5 reports the response of portfolios of corporate bonds to aggregate demand shocks ΔD_{agg} (Panel A) and shocks along duration ΔD_X (Panel B). Bonds are grouped in seven buckets based on duration: <1 year, 1–3 years, 3–5 years, 5–7 years, 7–10 years, 10–15 years, and 15+ years. The blue lines correspond to the estimates from column (4) of Table 3, which assume identical responses. The red lines are based on column (5), which includes linear interaction terms with duration X_{it}. The green line estimates these multipliers separately each duration-based portfolio in a pooled panel regression. The sample period is 2010:04 to 2022:09.

long-term bonds.

Columns (4) and (5) illustrate that the panel does not provide additional information about the macro and meso-multipliers relative to the time-series regression. These specifications correspond to panel regressions of individual bond returns on the idiosyncratic shock, $Z_{idio,it}$, the macro and meso demand shocks, $Z_{agg,t}$ and $Z_{X,t}$, as well as their interaction with duration X_{it} (for column (5)), and a control for X_{it}. Mechanically, the coefficient on $Z_{agg,t}$ and $Z_{X,t}$ coincide with the estimates of column (2); the coefficients on their interaction with X_{it} coincide with those of column (4).[34]

Stepping outside of our framework, one can support the linear specification for the role of duration by examining the impact of the meso and macro shocks on portfolios sorted on duration. Each time period, we form bond portfolios based on seven buckets of duration.[35] The blue lines in Figure 5 correspond to the response of each portfolio to macro (Panel A) and meso (Panel B) shocks predicted by the estimates of column (4) of Table 3 which assume that all bonds have the same response. The red lines entertain responses that depend on duration in a linear way as in column (5). Instead, the green lines estimate the coefficients on meso

[34] The number of corporate bonds present in the data varies across dates, creating an unbalanced panel. In such a situation, the coefficients in columns (4)-(5) mechanically coincide with those in columns (2)-(3) only if the repeated cross-sectional regressions in (4)-(5) are weighted so that each date receives equal weight.

[35] We follow the classification of the ICE BofA US Corporate indices: <1 year, 1–3 years, 3–5 years, 5–7 years, 7–10 years, 10–15 years, and 15+ years.

and macro shocks for each bucket separately. In this case, the estimate based on portfolio sorts suggest that heterogeneity is necessary. While this heterogeneity is well captured by a linear specification for the response to aggregate demand shocks (Panel A), it appears that the response to meso shocks (Panel B) is more subtle.[36]

5 Concluding Remarks

This paper provides a framework for using causal inference with asset prices and quantities. Specifically, we provide conditions for valid estimation in presence of the natural spillovers that exist between assets when making portfolio choices. The two conditions are constant relative elasticity and homogeneous substitution conditional on observables. The latter implies that two assets with the same observables are comparable if the demand for them responds in the same way to the price of every other asset. We show that the two conditions map naturally to restrictions often imposed in standard asset pricing models, and also provide guidelines to design experiments satisfying these conditions and assess their plausibility in the data.

When these conditions hold, the standard cross-sectional difference-in-difference or instrumental variable approach identifies the relative elasticity between comparable assets—that is, the difference between their own-price and cross-price elasticity. Other dimensions of substitutions such as separating own-price and cross-price elasticity, the macro elasticity, or responses to shocks across broad categories of assets, must be jointly estimated by a set of time series regressions. These simple tools and principles offer a straightforward package for researchers wanting to use natural experiments to better understand investment decisions and their equilibrium impact.

Because our conditions are flexible, they can guide empirical design without having to take a strong stance on a specific structural model. Still, these causal estimates should only be a first step towards a deeper understanding of how investors and institutions make portfolio decisions, and how those decisions shape equilibrium prices.

[36] Appendix E revisits the estimation focusing on changes in yields instead of returns, and finds more regularity in the estimates.

References

Aghaee, Alireza, "The Flattening Demand Curves," *Working paper*, 2024.

Allen, Jason, Jakub Kastl, and Milena Wittwer, "Estimating demand systems with bidding data," *Available at SSRN 5171755*, 2018.

An, Yu and Amy W. Huber, "Demand Propagation Through Traded Risk Factors," Technical Report 2025.

_ , Yinan Su, and Chen Wang, "Quantity, Risk, and Return," *Working Paper*, 2024.

Anderson, Simon P, André De Palma, and J-F Thisse, "A representative consumer theory of the logit model," *International Economic Review*, 1988, pp. 461–466.

Angrist, Joshua D. and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, "Ratings-driven demand and systematic price fluctuations," *The Review of Financial Studies*, 2022, *35* (6), 2790–2838.

Berg, Tobias, Markus Reisinger, and Daniel Streitz, "Spillover effects in empirical corporate finance," *Journal of Financial Economics*, 2021, *142* (3), 1109–1127.

Berry, Steven, James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, *63* (4), 841–890.

Berry, Steven T. and Philip A. Haile, "Foundations of demand estimation," in Kate Ho, Ali Hortaçsu, and Alessandro Lizzeri, eds., *Handbook of Industrial Organization*, Vol. 4 (1) of *Handbook of Industrial Organization*, Elsevier, 2021, pp. 1–62.

Bouveret, Antoine and Jie Yu, "Risks and vulnerabilities in the US bond mutual fund industry," Technical Report 2021.

Bretscher, Lorenzo, Lukas Schmid, Ishita Sen, and Varun Sharma, "Institutional corporate bond pricing," *Swiss Finance Institute Research Paper*, 2022, *21-07.*

Campbell, John Y., *Financial Decisions and Markets: A Course in Asset Pricing*, Princeton University Press, 2017.

_ and Luis Viceira, *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*, first ed., Oxford, UK: Oxford University Press, 2002.

Chang, Yen-Cheng, Harrison Hong, and Inessa Liskovich, "Regression Discontinuity and the Price Effects of Stock Market Indexing," *The Review of Financial Studies*, 07 2014, *28* (1), 212–246.

Chaudhary, Manav, Zhiyu Fu, and Jian Li, "Corporate bond multipliers: Substitutes matter," Technical Report 2022.

Chaudhry, Aditya, "The Impact of Prices on Analyst Cash Flow Expectations: Reconciling Subjective Beliefs Data with Rational Discount Rate Variation," *Journal of Financial Economics*, 2025, *forthcoming.*

Chen, Hui, Zhuo Chen, Zhiguo He, Jinyu Liu, and Rengming Xie, "Pledgeability and asset prices: Evidence from the Chinese corporate bond markets," *The Journal of Finance*, 2023, *78* (5), 2563–2620.

Chodorow-Reich, Gabriel, Plamen T Nenov, and Alp Simsek, "Stock market wealth and the real economy: A local labor market approach," *American Economic Review*, 2021, *111* (5), 1613–1657.

Cochrane, John H, *Asset pricing*, Princeton University Press, 2005.

Coppola, Antonio, "In safe hands: The financial and real impact of investor composition over the credit cycle," *The Review of Financial Studies*, 2025, *forthcoming.*

Coval, Joshua and Erik Stafford, "Asset fire sales (and purchases) in equity markets," *Journal of Financial Economics*, 2007, *86* (2), 479–512.

Darmouni, Olivier, Kerry Siani, and Kairong Xiao, "Nonbank Fragility in Credit Markets: Evidence from a Two-Layer Asset Demand System," Technical Report 2023.

Davis, Carter, "The Elasticity of Quantitative Investment," *The Review of Financial Studies*, 2024, *forthcoming.*

_ , Mahyar Kargar, and Jiacui Li, "Why Do Portfolio Choice Models Predict Inelastic Demand?," *Journal of Financial Economics*, 2025, *forthcoming.*

Deaton, Angus and John Muellbauer, "An almost ideal demand system," *The American economic review*, 1980, *70* (3), 312–326.

der Beck, Philippe Van, "Flow-driven ESG returns," *Swiss Finance Institute Research Paper*, 2021, *21-71.*

Du, Wenxin, Alexander Tepper, and Adrien Verdelhan, "Deviations from covered interest rate parity," *The Journal of Finance*, 2018, *73* (3), 915–957.

Duffie, Darrell, *Dynamic asset pricing theory*, Princeton University Press, 2010.

Fama, Eugene F and James D MacBeth, "Risk, return, and equilibrium: Empirical tests," *Journal of political economy*, 1973, *81* (3), 607–636.

_ and Kenneth R French, "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, 1993, *33* (1), 3–56.

Fang, Chuck, *Monetary policy amplification through bond fund flows*, University of Pennsylvania, 2023.

_ **and Kairong Xiao**, “Dissecting Bond Market Transmission of Monetary Policy,” *Available at SSRN 5025417*, 2024.

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, “Demand-System Asset Pricing: Theoretical Foundations,” *Available at SSRN 4672473*, 2025.

Gabaix, Xavier and Ralph SJ Koijen, “In search of the origins of financial fluctuations: The inelastic markets hypothesis,” Technical Report, National Bureau of Economic Research 2021.

_ **and** _ , “Granular instrumental variables,” *Journal of Political Economy*, 2024, *132* (7), 000–000.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár, “Contamination bias in linear regressions,” *American Economic Review*, 2024, *114* (12), 4015–4051.

Gompers, Paul A. and Andrew Metrick, “Institutional Investors and Equity Prices*,” *The Quarterly Journal of Economics*, 02 2001, *116* (1), 229–259.

Greenwood, Robin and Marco Sammon, “The Disappearing Index Effect,” *Working paper*, 2024.

Greenwood, Robin Marc and Annette Vissing-Jorgensen, “The impact of pensions and insurance on global yield curves,” Technical Report, Harvard Business School 2018.

Greenwood, Robin, Samuel G Hanson, and Gordon Y Liao, “Asset Price Dynamics in Partially Segmented Markets,” *The Review of Financial Studies*, 04 2018, *31* (9), 3307–3343.

Guren, Adam, Alisdair McKay, Emi Nakamura, and Jón Steinsson, “What Do We Learn from Cross-Regional Empirical Estimates in Macroeconomics?,” *NBER Macroeconomics Annual*, 2021, *35*, 175–223.

Haddad, Valentin, Alan Moreira, and Tyler Muir, “When selling becomes viral: Disruptions in debt markets in the COVID-19 crisis and the Fed’s response,” *The Review of Financial Studies*, 2021, *34* (11), 5309–5351.

_ **,** _ **, and** _ , “Whatever it takes? The impact of conditional policy promises,” *American Economic Review*, 2025, *115* (1), 295–329.

_ **and Tyler Muir**, “Do intermediaries matter for aggregate asset prices?,” *The Journal of Finance*, 2021, *76* (6), 2719–2761.

_ **, Paul Huebner, and Erik Loualiche**, “How competitive is the stock market? theory, evidence from portfolios, and implications for the rise of passive investing,” *Working paper*, 2024.

Hansen, Lars Peter and Kenneth J Singleton, “Generalized instrumental variables estimation of nonlinear rational expectations models,” *Econometrica: Journal of the Econometric Society*, 1982, pp. 1269–1286.

Harris, Lawrence and Eitan Gurel, "Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures," *The Journal of Finance*, 1986, *41* (4), 815–829.

Hartzmark, Samuel M and David H Solomon, "Predictable price pressure," Technical Report, National Bureau of Economic Research 2022.

He, Zhiguo, Paymon Khorrami, and Zhaogang Song, "Commonality in credit spread changes: Dealer inventory and intermediary distress," *The Review of Financial Studies*, 2022, *35* (10), 4630–4673.

_ , Peter Kondor, and Jessica Shi Li, "Demand Elasticity in Dynamic Asset Pricing," Technical Report, Working paper 2025.

Huber, Kilian, "Estimating general equilibrium spillovers of large-scale shocks," *The Review of Financial Studies*, 2023, *36* (4), 1548–1584.

Huebner, Paul, "The Making of Momentum: A Demand-System Perspective," *Working paper*, 2024.

Jansen, Kristy A.E., Wenhao Li, and Lukas Schmid, "Granular Treasury Demand with Arbitrageurs," *Available at SSRN 4940397*, 2024.

Jiang, Zhengyang, Robert J Richmond, and Tony Zhang, "A portfolio approach to global imbalances," *The Journal of Finance*, 2024, *79* (3), 2025–2076.

Koijen, Ralph S. J. and Motohiro Yogo, "A Demand System Approach to Asset Pricing," *Journal of Political Economy*, 2019, *127* (4), 1475–1515.

_ and _ , "Exchange Rates and Asset Prices in a Global Demand System," *Working paper*, 2024.

Koijen, Ralph S J, Robert J Richmond, and Motohiro Yogo, "Which Investors Matter for Equity Valuations and Expected Returns?," *The Review of Economic Studies*, 08 2023, *91* (4), 2387–2424.

Krishnamurthy, Arvind and Annette Vissing-Jorgensen, "The effects of quantitative easing on interest rates: channels and implications for policy," Technical Report, National Bureau of Economic Research 2011.

Li, Jiacui and Zihan Lin, "Price Multipliers are Larger at More Aggregate Levels," *Available at SSRN 4038664*, 2022.

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

Lu, Xu and Lingxuan Wu, "Monetary Transmission and Portfolio Rebalancing: A Cross-Sectional Approach," Technical Report 2023.

Markowitz, Harry M, "Portfolio selection," *The journal of finance*, 1952, *7*, 77–91.

Olea, José Luis Montiel and Carolin Pflueger, "A robust test for weak instruments," *Journal of Business & Economic Statistics*, 2013, *31* (3), 358–369.

Pavlova, Anna and Taisiya Sikorskaya, "Benchmarking Intensity," *The Review of Financial Studies*, 08 2022, *36* (3), 859–903.

Peng, Cameron and Chen Wang, "Factor Demand and Factor Returns," *Working paper*, 2023.

Petajisto, Antti, "Why do demand curves for stocks slope down?," *Journal of Financial and Quantitative Analysis*, 2009, *44* (5), 1013–1044.

Selgrad, Julia, "Testing the Portfolio Rebalancing Channel of Quantitative Easing," *Working paper*, 2023.

Shleifer, Andrei, "Do Demand Curves for Stocks Slope Down?," *The Journal of Finance*, 1986, *41* (3), 579–590.

Stock, James H and Motohiro Yogo, "Testing for weak instruments in Linear Iv regression," in "Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg," Cambridge University Press, 2005, pp. 80–108.

Vayanos, Dimitri and Jean-Luc Vila, "A preferred-habitat model of the term structure of interest rates," *Econometrica*, 2021, *89* (1), 77–112.

Wolf, Christian K, "The missing intercept: A demand equivalence approach," *American Economic Review*, 2023, *113* (8), 2232–2269.

Appendix

Contents

A Proofs and Derivations

A.1 Identifying the relative elasticity – Proposition 2

Start from the general demand equation with demand shocks:

$$\Delta D_i = \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{ij}\Delta P_j + \Delta\bar{D}_i. \tag{80}$$

We recall the two assumptions necessary for identification:

- **Assumption A1.** $X_i = X_j \Rightarrow \mathcal{E}_{il} = \mathcal{E}_{jl} = \mathcal{E}_{\text{cross}}(X_i, X_l) = X_i'\mathcal{E}_X X_l, \quad \forall i,j \in \mathcal{S}, l \neq i,j,$ where X_i is a $K \times 1$ vector of observables, and $\mathcal{E}_X$ is a $K \times K$ matrix.
- **Assumption A2.** $\mathcal{E}_{ii} - \mathcal{E}_{\text{cross}}(X_i, X_i) = \mathcal{E}_{jj} - \mathcal{E}_{\text{cross}}(X_j, X_j) = \widehat{\mathcal{E}}, \quad \forall i,j \in \mathcal{S}$

Proposition 2 shows that under assumptions A1 and A2 and the exogeneity condition, the IV estimator, conditioning on X_i, identifies coefficient $\widehat{\mathcal{E}}$.

Proof. Starting from equation (80), we can rewrite the demand equation as a cross-sectional regression:

$$\begin{aligned}
\Delta D_i &= \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{ij}\Delta P_j + \Delta\bar{D}_i && (81)\\
&= \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \Delta\bar{D}_i && (82)\\
&= \left(\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i)\right)\Delta P_i + \sum_{j}\mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \Delta\bar{D}_i && (83)\\
&= \widehat{\mathcal{E}}\Delta P_i + \sum_{j}\mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \Delta\bar{D}_i && (84)\\
&= \widehat{\mathcal{E}}\Delta P_i + \sum_{j}X_i'\mathcal{E}_X X_j\Delta P_j + \Delta\bar{D}_i && (85)\\
&= \widehat{\mathcal{E}}\Delta P_i + X_i'\underbrace{\left(\sum_{j}\mathcal{E}_X X_j\Delta P_j\right)}_{\theta} + \Delta\bar{D}_i && (86)\\
&= \widehat{\mathcal{E}}\Delta P_i + \theta' X_i + \Delta\bar{D}_i && (87)
\end{aligned}$$

Equation (83) adds and subtracts $\mathcal{E}_{cross}(X_i, X_i)\Delta P_i$. Equations (84) and (85) use assumptions 2 and 1, respectively. Equation (86) pulls out X_i' from the sum. The remaining part of the sum gets absorbed into θ, a $K \times 1$ vector of cross-sectional constants. These θ are K regression coefficients on the K observables, X_{ik}.

Given the exclusion restriction that $Z_i \perp \varDelta \bar{D}_i | X_i$ and the relevance condition that $cov(\varDelta P_i, Z_i | X_i) \neq 0$, this is the standard IV setting, and the regression estimates $\widehat{\mathcal{E}}$. ■

A.2 Properties of elasticity under assumptions A1 and A2

A.2.1 A matrix representation.

First, we derive a simple matrix representation for an elasticity matrix under our two assumptions.

Lemma 8 *Let $\mathcal{E}$ be an elasticity matrix that satisfies assumptions A1 and A2. Then it can be written as:*

$$\mathcal{E} = \widehat{\mathcal{E}}\mathbf{I} + X\mathcal{E}_X X', \tag{88}$$

where $\widehat{\mathcal{E}}$ is a scalar equal to the relative elasticity and $\mathcal{E}_X$ is a $K \times K$ matrix.

Proof. Write out the elasticity matrix $\mathcal{E}$:

$$\mathcal{E} = \begin{pmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} & \dots & \mathcal{E}_{1N} \\ \mathcal{E}_{21} & \mathcal{E}_{22} & \dots & \mathcal{E}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}_{N1} & \mathcal{E}_{N2} & \dots & \mathcal{E}_{NN} \end{pmatrix} = \begin{pmatrix} \widehat{\mathcal{E}} + X_1'\mathcal{E}_X X_1 & X_1'\mathcal{E}_X X_2 & \dots & X_1'\mathcal{E}_X X_N \\ X_2'\mathcal{E}_X X_1 & \widehat{\mathcal{E}} + X_2'\mathcal{E}_X X_2 & \dots & X_2'\mathcal{E}_X X_N \\ \vdots & \vdots & \ddots & \vdots \\ X_N'\mathcal{E}_X X_1 & X_N'\mathcal{E}_X X_2 & \dots & \widehat{\mathcal{E}} + X_N'\mathcal{E}_X X_N \end{pmatrix} \tag{89}$$

The (i, j) element of matrix $\mathcal{E}$ is $[\mathcal{E}]_{ij} = X_i'\mathcal{E}_X X_j = \mathcal{E}_{cross}(X_i, X_j)$, as defined by Assumption 1, for $i \neq j$. The diagonal elements are $[\mathcal{E}]_{ii} = \widehat{\mathcal{E}} + X_i'\mathcal{E}_X X_i = \widehat{\mathcal{E}} + \mathcal{E}_{cross}(X_i, X_i)$, as in Assumption 2. Since each element in $\mathcal{E}$ directly corresponds to the respective $\mathcal{E}_{ij}$ defined by assumptions A1 and A2, the assumptions are equivalent to the elasticity matrix in (88). ■

A.2.2 Transforming the observables.

The following lemma shows that observables can be recombined in a linear way. In particular they could be demeaned, standardized, or orthogonalized.

Lemma 9 *Let $\mathcal{E}$ be an elasticity matrix that satisfies assumptions A1 and A2 with respect to a set of observables X. If A is a $K \times K$ invertible matrix, $\mathcal{E}$ satisfies assumptions A1 and A2 with respect to the recombined observables $\tilde{X} = XA$.*

Proof. Insert AA^{-1} judiciously into the decomposition of Lemma 8.

$$\mathcal{M} = \widehat{\mathcal{M}}\mathbf{I} + XAA^{-1}\mathcal{M}_X(A')^{-1}A'X' = \widehat{\mathcal{M}}\mathbf{I} + \tilde{X}\mathcal{M}_{\tilde{X}}\tilde{X}', \tag{90}$$

$$\text{with } \mathcal{M}_{\tilde{X}} = A^{-1}\mathcal{M}_X(A')^{-1}. \tag{91}$$

■

For example, if the first observable is a constant and the other ones have mean $\bar{X}_1, \ldots, \bar{X}_{K-1}$, the following matrix demeans them:

$$A_{\text{demean}} = \mathbf{I}_K - \begin{pmatrix} 0 & \bar{X}_1 & \cdots & \bar{X}_{K-1} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{92}$$

Importantly, note that there is no reason that orthogonalizing the characteristics makes the substitution matrix $\mathcal{M}_{\tilde{X}}$ diagonal.

A.2.3 Stability by inversion — Proposition 5.

Proposition 5 shows that under assumptions 1 and 2, the multiplier matrix $\mathcal{M} = -\mathcal{E}^{-1}$ also satisfies assumptions 1 and 2, with $\widehat{\mathcal{M}} = -1/\widehat{\mathcal{E}}$.
Proof. Start from equation (88), and apply the Woodbury matrix identity:

$$\begin{aligned} -\mathcal{E}^{-1} &= -\left(\widehat{\mathcal{E}}\mathbf{I} + X\mathcal{E}_X X'\right)^{-1} & (93) \\ &= -\widehat{\mathcal{E}}^{-1}\mathbf{I} + X\left(\widehat{\mathcal{E}}^2\mathcal{E}_X^{-1} + \widehat{\mathcal{E}}X'X\right)^{-1} X' & (94) \\ &= \widehat{\mathcal{M}}\mathbf{I} + X\mathcal{M}_X X'. & (95) \end{aligned}$$

This corresponds exactly to assumptions A1 and A2 applied to $\mathcal{M}$ with $\widehat{\mathcal{M}} = -1/\widehat{\mathcal{E}}$. ■

A.2.4 Stability by aggregation

We show that that assumptions A1 and A2 are stable by aggregation across investors.

Lemma 10 *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two elasticity matrices that satisfy assumptions A1 and A2, and (λ_1, λ_2) two scalars. Then the matrix $\lambda_1\mathcal{E}_1 + \lambda_2\mathcal{E}_2$ satisfies assumptions A1 and A2.*

Proof. From lemma 8 we decompose both elasticities which leads to:

$$\lambda_1\mathcal{E}_1 + \lambda_2\mathcal{E}_2 = \left(\lambda_1\widehat{\mathcal{E}}_1 + \lambda_2\widehat{\mathcal{E}}_2\right)\mathbf{I} + X\left(\lambda_1\mathcal{E}_{X,1} + \lambda_2\mathcal{E}_{X,2}\right)X' \tag{96}$$

The decomposition and the equivalence from lemma 8 concludes the proof. ■

A.3 Heterogeneous relative elasticities based on observables

We maintain assumption A1. We relax assumption A2 by allowing the relative elasticity to depend linearly on observables:

$$\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i) = \mathcal{E}_{relative}(X_i) = \mathcal{E}_r' X_i. \tag{97}$$

From section A.1 and the proof of Proposition 2 we obtain:

$$\begin{aligned}
\Delta D_i &= \mathcal{E}_{ii} \Delta P_i + \sum_{j \neq i} \mathcal{E}_{ij} \Delta P_j + \epsilon_i && (98)\\
&= \mathcal{E}_{ii} \Delta P_i + \sum_{j \neq i} \mathcal{E}_{cross}(X_i, X_j) \Delta P_j + \epsilon_i && (99)\\
&= \left(\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i)\right) \Delta P_i + \sum_{j} \mathcal{E}_{cross}(X_i, X_j) \Delta P_j + \epsilon_i && (100)\\
&= \mathcal{E}_{relative}(X_i) \Delta P_i + \sum_{j} X_i' \mathcal{E}_X X_j \Delta P_j + \epsilon_i && (101)\\
&= \underbrace{\mathcal{E}_{relative}(X_i)}_{\mathcal{E}_r' X_i} \Delta P_i + X_i' \underbrace{\left(\sum_{j} \mathcal{E}_X X_j \Delta P_j \right)}_{\theta} + \epsilon_i && (102)\\
&= \mathcal{E}_r' X_i \Delta P_i + \theta' X_i + \epsilon_i && (103)
\end{aligned}$$

In this case we want to identify the vector $\mathcal{E}_r$ which characterizes the relative elasticity with respect to observables. Identification must rely on a vector of instruments. It is natural to construct those instruments from a single instrument Z_i for the price interacted with the observables. The set of identification conditions is:

$$Z_i X_i \perp \epsilon_i | X_i \tag{104}$$

Under these conditions the two-stage least squares regression proceeds as follows. First, regress each component of $X_i \Delta P_i$ on the vector of instruments $X_i Z_i$ and the observables X_i. The relevance condition is that the matrix of coefficients on the instruments is full-rank. This leads to predicted values of the change in price interacted with observables $\widehat{X_i \Delta P_i}$. Second, regress the change in demand on these predicted values and the observables. The coefficients on the predicted values recovers $\mathcal{E}_r$. Finally, the relative elasticity for each asset is simply $\mathcal{E}_{relative}(X_i) = \mathcal{E}_r' X_i$.

A.4 Identification beyond the relative elasticity.

We consider aggregation for the generic case of a multiplier matrix $\mathcal{M}$ that satisfies assumptions A1 and A2 for an arbitrary set of observables X. Remember that the first observable is the constant in most cases.

Using Lemma 8, we can represent $\mathcal{M}$ as

$$\mathcal{M} = \widehat{\mathcal{M}}\mathbf{I} + X\mathcal{M}_X X'. \tag{105}$$

To define price and quantity aggregates along the various dimensions of the observables, we regress these vectors on X. We will see that this the natural generalization of the aggregation presented in Proposition 7.

Proposition 11 (Multiplier decomposition with observables in the general case) *Take a multiplier matrix $\mathcal{M}$ satisfying assumptions A1 and A2. Consider generic changes in demand and price connected by this matrix: $\Delta P = \mathcal{M}\Delta D$. Define the change in demand and price aggregated along observables and the idiosyncratic component:*

$$\Delta P_X = (X'X)^{-1}X'\Delta P \qquad \Delta D_X = (X'X)^{-1}X'\Delta D. \tag{106}$$

$$\Delta P_{idio,i} = \Delta P_i - X_i'\Delta P_X \qquad \Delta D_{idio,i} = \Delta D_i - X_i'\Delta D_X. \tag{107}$$

The response of prices to a change in demand can be decomposed into two sets of components:

$$\text{Micro:} \qquad \Delta P_{idio,i} = \widehat{\mathcal{M}}\Delta D_{idio,i} \tag{108}$$

$$\text{Meso-Macro:} \qquad \Delta P_X = \widetilde{\mathcal{M}}\Delta D_X, \tag{109}$$

where $\widetilde{\mathcal{M}} = \widehat{\mathcal{M}}\mathbf{I}_K + \mathcal{M}_X X'X$.

Proof. Using the relation $\Delta P = \mathcal{M}\Delta D$ and the decomposition of $\mathcal{M}$ under the two assumptions, we obtain

$$(X'X)^{-1}X'\Delta P = \left(\widehat{\mathcal{M}}(X'X)^{-1}X'\mathbf{I}_N + (X'X)^{-1}X'X\mathcal{M}_X X'\right)\Delta D \tag{110}$$

$$= \left(\widehat{\mathcal{M}}(X'X)^{-1} + \mathcal{M}_X\right)X'\Delta D \tag{111}$$

$$= \left(\widehat{\mathcal{M}}\mathbf{I}_K + \mathcal{M}_X(X'X)\right)(X'X)^{-1}X'\Delta D \tag{112}$$

$$\Delta P_X = \left(\widehat{\mathcal{M}}\mathbf{I}_K + \mathcal{M}_X(X'X)\right)\Delta D_X \tag{113}$$

This implies that ΔP_X can be expressed as a linear combination of the K elements of ΔD_X, as opposed to the whole N components of the changes in demand ΔD.

From the definition of the idiosyncratic change in demand:

$$\begin{aligned}\varDelta P_{idio} &= \varDelta P - X\varDelta P_X && (114)\\ &= \widehat{\mathcal{M}}\varDelta D + X\mathcal{M}_X X'\varDelta D - \left(X\widehat{\mathcal{M}} + X\mathcal{M}_X X'X\right)(X'X)^{-1}X'\varDelta D && (115)\\ &= \widehat{\mathcal{M}}(\varDelta D - X\varDelta D_X) + X\mathcal{M}_X X'\varDelta D - X\mathcal{M}_X(X'X)(X'X)^{-1}X'\varDelta D && (116)\\ &= \widehat{\mathcal{M}}\varDelta D_{idio}. && (117)\end{aligned}$$

Because $\widehat{\mathcal{M}}$ is scalar, the idiosyncratic component is determined asset by asset, which concludes the proof. ■

Simple case with no characteristic. We can recover the simpler cases studied in the paper. Proposition 6 corresponds to a single variable X which is constant equal to one. In this case $\varDelta P_X$ has only one component equal to $\varDelta P_{agg} = N^{-1}\sum_i \varDelta P_i$, and $\widetilde{\mathcal{M}} = \widehat{\mathcal{M}} + N\mathcal{M}_X$ is a scalar equal to the macro multiplier.

With one normalized characteristic. Proposition 7 corresponds to observables that include a constant and a single standardized characteristic that we call X in a slight abuse of notation. There, the regression gives two aggregate prices and quantities: the aggregate component $\varDelta P_{agg}$ defined as before (the constant of the regression); the meso component $\varDelta P_X = N^{-1}\sum_i X_i \varDelta P_i$. Then the matrix $\widetilde{\mathcal{M}}$ is 2 × 2 and equal to

$$\widetilde{\mathcal{M}} = \begin{pmatrix} \widehat{\mathcal{M}} + N(\mathcal{M}_X)_{11} & N(\mathcal{M}_X)_{12} \\ N(\mathcal{M}_X)_{21} & \widehat{\mathcal{M}} + N(\mathcal{M}_X)_{22} \end{pmatrix} = \begin{pmatrix} \overline{\mathcal{M}}_{agg} & \overline{\mathcal{M}}_X \\ \widetilde{\mathcal{M}}_{agg} & \widetilde{\mathcal{M}}_X \end{pmatrix}. \tag{118}$$

When observables are group dummies. Consider the case when the observables are dummy variables expressing the belonging to disjoint groups. In this situation, there is no need for a constant. The aggregate price and demand indices have a simple interpretation: they measure the average change in price and demand for each group k:

$$\varDelta D_{X,k} = \frac{1}{N_k}\sum_{i\in k} \varDelta D_i \tag{119}$$

This implies that price impact has a nested structure. First, there is a relative multiplier within each group $\widehat{\mathcal{M}}$ capturing the impact of changes in relative demand within a group. Then individual assets can be replaced by the aggregate portfolio of each group, and there is a multiplier matrix across these aggregate portfolios, $\widetilde{\mathcal{M}}$.

A.5 Lack of identification of substitution from the cross-section

We show that without additional restrictions, substitution cannot be identified from a single cross section. Start with the structural relation: $\varDelta D = \mathcal{E}\varDelta P + \epsilon$, and impose our assumptions: $\mathcal{E} = \hat{\mathcal{E}}I + X\mathcal{E}_X X'$.

Recall that the demand shift ϵ measures all demand changes that are not caused by a price change. In particular, it can have a relation with the observables X (e.g., if beliefs about relative payoffs of assets with different values of X change) and a non-zero mean (e.g., if the investor demands more assets overall). We can separate ϵ across those components:

$$\begin{aligned} \epsilon_X &= (X'X)^{-1}X'\epsilon & (120) \\ \epsilon_{idio,i} &= \epsilon_i - X_i'\epsilon_X, & (121) \end{aligned}$$

with all components of ϵ_X potentially different from 0 even in the limit of many assets (large N) and imposing no constraints on all other model quantities.

Plugging into demand, we obtain:

$$\begin{aligned} \varDelta D_i &= \hat{\mathcal{E}}\varDelta P_i + X_i'\mathcal{E}_X X'\varDelta P + X_i'\epsilon_X + \epsilon_{idio,i} & (122) \\ &= \hat{\mathcal{E}}\varDelta P_i + \epsilon_{idio,i} + X_i' \underbrace{(\mathcal{E}_X X'\varDelta P + \epsilon_X)}_{K\times 1} & (123) \end{aligned}$$

Proposition 12 *No free parameter of the matrix $\mathcal{E}_X$ can be identified from a single cross section, even under the restriction that some of coefficients of $\mathcal{E}_X$ are 0.*

Proof. Denote $\mathcal{E}_X^{true}$ and ϵ_X^{true} the true values of $\mathcal{E}_X$ and ϵ_X. For any other guess $\mathcal{E}_X^{false}$, the model with $\mathcal{E}_X = \mathcal{E}_X^{false}$ and $\epsilon_X = \epsilon_X^{true} + (\mathcal{E}_X^{true} - \mathcal{E}_X^{false})X'\varDelta P$ is observationally equivalent to the true model. As long as such guesses exist, that is, as long as $\mathcal{E}_X$ has at least one free parameter, this concludes the proof of no identification. ■

Simplest case: only a constant. Consider the simplest possible case, where X is a constant. Writing cross-sectional means at date t with a bar, and noting that $\mathcal{E}_X$ is a scalar in this case, we have:

$$\begin{aligned} \varDelta D_{i,t} &= \hat{\mathcal{E}}\varDelta P_{i,t} + \mathcal{E}_X\overline{\varDelta P}_t + \overline{\epsilon}_t + \epsilon_{idio,i,t} & (124) \\ &= (\mathcal{E}_X\overline{\varDelta P}_t + \overline{\epsilon}_t) + \hat{\mathcal{E}}\varDelta P_{i,t} + \epsilon_{idio,i,t} & (125) \end{aligned}$$

Both substitution $\mathcal{E}_X\overline{\varDelta P}_t$ and the aggregate demand shift $\overline{\epsilon}_t$ contribute to the constant of a cross-sectional regression, and there is no way to separate them. This is a version of the missing intercept problem.

Obtaining partial identification with symmetry across observables. One way to obtain some partial identification of substitution is to impose that the same parameter drives substitution across many observables. Then, if one also assumes that the number of observables grows as the number of assets increases in the cross section, one can identify part of the substitution matrix.

We illustrate this approach when the observables are dummy variables belonging to a given group. A priori, substitution across groups could follow any matrix $\mathcal{E}_X$. But one might want to assume that all groups substitute symmetrically, that is, $\mathcal{E}_X = \mathcal{E}_{own-g}I + \mathcal{E}_{cross-g}11'$, with $\mathcal{E}_{own-g}$ an own-group elasticity and $\mathcal{E}_{cross-g}$ a cross-group elasticity. In this case, we have:

$$\varDelta D_i = \hat{\mathcal{E}}\varDelta P_i + \mathcal{E}_{own-g} N_g \varDelta P_g + \mathcal{E}_{cross-g} N \overline{\varDelta P} \text{ if } i \in g \tag{126}$$

A cross-sectional regression with an instrument for $\varDelta P_g$ across groups allow to recover $\mathcal{E}_{own-g}$. Notice that $\mathcal{E}_{cross-g}$ remains unidentified. This approach corresponds to repeating the relative elasticity estimation at a higher level of aggregation: the matrix $\mathcal{E}_X$ (in contrast to $\mathcal{E}$) satisfies assumptions A1 and A2 with only a constant as observable.

The key assumption here is not that the observables correspond to disjoint groups, but instead that there is a common substitution parameter that affects each of the observables separately. Its plausibility depends on context: for example in a simple mean variance setting, it does not apply if the groups are based on levels of factor loadings.

The nested logit model assumes such a symmetry across groups (the "nests") and additionally imposes that the missing intercept $\mathcal{E}_{cross-g}$ is pinned down by the other parameters.

A.6 Estimating a LATE — Proposition 3

The data-generating process under heterogeneous treatment effects is:

$$\varDelta D_i = \mathcal{E}_{ii}\varDelta P_i + \sum_{j\neq i} \mathcal{E}_{ij}\varDelta P_j + \epsilon_i \tag{127}$$

$$\varDelta P_i = \lambda_i Z_i + u_i \tag{128}$$

The instrument Z_i, with constant variance $var(Z_i) = var(Z), \forall i$, is randomly assigned and independent of everything else:

$$Z_i \perp\!\!\!\perp Z_j \quad \forall i \neq j \tag{129}$$

$$Z_i \perp\!\!\!\perp \mathcal{E}_{kl} \quad \forall i, k, l \tag{130}$$

$$Z_i \perp\!\!\!\perp \lambda_j \quad \forall i, j \tag{131}$$

$$Z_i \perp\!\!\!\perp u_j \quad \forall i, j \tag{132}$$

$$Z_i \perp\!\!\!\perp \epsilon_j \quad \forall i, j \tag{133}$$

After substituting (128) into (127), we derive the estimate from the demand equation

$$\varDelta D_i = \mathcal{E}_{ii}\lambda_i Z_i + \sum_{j\neq i} \mathcal{E}_{ij}\lambda_j Z_j + \mathcal{E}_{ii} u_i + \sum_{j\neq i} \mathcal{E}_{ij} u_j + \epsilon_i \tag{134}$$

Definitions and preliminaries. Without loss of generality, define a centered instrument $\tilde{Z}_i$ as

$$\tilde{Z}_i \equiv Z_i - \frac{1}{N}\sum_j Z_j, \tag{135}$$

such that we have the following properties:

$$\sum_{j\neq i}\tilde{Z}_j = -\tilde{Z}_i \tag{136}$$

$$cov(\tilde{Z}_i,\tilde{Z}_j) = \underbrace{\text{cov}(Z_i,Z_j)}_{=0} - \frac{1}{N}\underbrace{\sum_k \text{cov}(Z_k,Z_j)}_{=\text{var}(Z)} - \frac{1}{N}\underbrace{\sum_l cov(Z_i,Z_l)}_{=\text{var}(Z)} + \frac{1}{N^2}\underbrace{\text{cov}(\sum_k Z_k,\sum_l Z_l)}_{=N\,\text{var}(Z)} \tag{137}$$

$$= -\frac{1}{N}\,\text{var}(Z) \tag{138}$$

Next, define $\bar{\lambda}$ and $\tilde{\lambda}_i$ such that:

$$\lambda_i = \bar{\lambda} + \tilde{\lambda}_i \tag{139}$$

$$\sum_j \lambda_j = N\bar{\lambda} \tag{140}$$

$$\sum_{j\neq i}\tilde{\lambda}_j = -\tilde{\lambda}_i \tag{141}$$

Finally, define $\mathcal{E}_{i,cross}$ as the λ_j weighted average of $\mathcal{E}_{ij}$:

$$\mathcal{E}_{i,cross} = \frac{\sum_{j\neq i}\lambda_j\mathcal{E}_{ij}}{\sum_{j\neq i}\lambda_j} \tag{142}$$

Proof. Based on the definitions above, rewrite $\sum_{j\neq i}\mathcal{E}_{ij}\lambda_j\tilde{Z}_j$ from equation (134) as:

$$\sum_{j\neq i}\mathcal{E}_{ij}\lambda_j\tilde{Z}_j = \mathcal{E}_{i,cross}\sum_{j\neq i}\lambda_j\tilde{Z}_j + \sum_{j\neq i}(\mathcal{E}_{ij}-\mathcal{E}_{i,cross})\lambda_j\tilde{Z}_j \tag{143}$$

$$= \mathcal{E}_{i,cross}\bar{\lambda}\underbrace{\sum_{j\neq i}\tilde{Z}_j}_{=-\tilde{Z}_i} + \mathcal{E}_{i,cross}\sum_{j\neq i}\tilde{\lambda}_j\tilde{Z}_j + \sum_{j\neq i}(\mathcal{E}_{ij}-\mathcal{E}_{i,cross})\lambda_j\tilde{Z}_j \tag{144}$$

$$= -\mathcal{E}_{i,cross}\bar{\lambda}\tilde{Z}_i + \mathcal{E}_{i,cross}\sum_{j\neq i}\tilde{\lambda}_j\tilde{Z}_j + \sum_{j\neq i}(\mathcal{E}_{ij}-\mathcal{E}_{i,cross})\lambda_j\tilde{Z}_j \tag{145}$$

Plugging into equation (134):

$$\mathit{\Delta}D_i = \left(\mathcal{E}_{ii}\lambda_i - \mathcal{E}_{i,cross}\bar{\lambda}\right)\tilde{Z}_i + \mathcal{E}_{i,cross}\sum_{j\neq i}\tilde{\lambda}_j\tilde{Z}_j + \sum_{j\neq i}(\mathcal{E}_{ij}-\mathcal{E}_{i,cross})\lambda_j\tilde{Z}_j + \mathcal{E}_{ii}u_i + \sum_{j\neq i}\mathcal{E}_{ij}u_j + \epsilon_i \tag{146}$$

We are interested in $cov(\mathit{\Delta}D_i,\tilde{Z}_i)$ and $cov(\mathit{\Delta}P_i,\tilde{Z}_i)$. Since $\tilde{Z}_i$ is mean-zero, by the law of iterated expectations we have:

$$cov(\mathit{\Delta} D_i, \tilde{Z}_i) = \mathbb{E}\left[\mathit{\Delta} D_i \tilde{Z}_i\right] = \mathbb{E}\left[\mathbb{E}\left[\mathit{\Delta} D_i \tilde{Z}_i | \mathit{\Theta}\right]\right], \tag{147}$$

where $\mathit{\Theta}$ is a set that contains all $\mathcal{E}_{ij}$ and λ_i.

We have:

$$\begin{aligned}
\mathbb{E}\left[\left(\mathcal{E}_{ii}\lambda_i - \mathcal{E}_{i,cross}\bar{\lambda}\right)\tilde{Z}_i^2 | \mathit{\Theta}\right] &= \left(\mathcal{E}_{ii}\lambda_i - \mathcal{E}_{i,cross}\bar{\lambda}\right) var(\tilde{Z}) && (148)\\
\mathbb{E}\left[\mathcal{E}_{i,cross}\sum_{j\neq i}\tilde{\lambda}_j \tilde{Z}_i \tilde{Z}_j | \mathit{\Theta}\right] &= \mathcal{E}_{i,cross}\sum_{j\neq i}\tilde{\lambda}_j \mathbb{E}\left[\tilde{Z}_i, \tilde{Z}_j\right] && (149)\\
&= -\frac{var(Z)}{N}\mathcal{E}_{i,cross}\sum_{j\neq i}\tilde{\lambda}_j && (150)\\
&= \frac{var(Z)}{N}\mathcal{E}_{i,cross}\tilde{\lambda}_i && (151)\\
&= \frac{N var(\tilde{Z})}{(N-1)^2}\mathcal{E}_{i,cross}\tilde{\lambda}_i && (152)\\
\mathbb{E}\left[\sum_{j\neq i}(\mathcal{E}_{ij} - \mathcal{E}_{i,cross})\lambda_j \tilde{Z}_i \tilde{Z}_j | \mathit{\Theta}\right] &= \sum_{j\neq i}(\mathcal{E}_{ij} - \mathcal{E}_{i,cross})\lambda_j \mathbb{E}\left[\tilde{Z}_i, \tilde{Z}_j\right] && (153)\\
&= -\frac{var(Z)}{N}\sum_{j\neq i}(\mathcal{E}_{ij} - \mathcal{E}_{i,cross})\lambda_j && (154)\\
&= -\frac{var(Z)}{N}\left(\frac{\sum_{j\neq i}\lambda_j \mathcal{E}_{ij}}{\sum_{j\neq i}\lambda_j}\sum_{j\neq i}\lambda_j - \mathcal{E}_{i,cross}\sum_{j\neq i}\lambda_j\right) && (155)\\
&= -\frac{var(Z)}{N}\left(\mathcal{E}_{i,cross}\sum_{j\neq i}\lambda_j - \mathcal{E}_{i,cross}\sum_{j\neq i}\lambda_j\right) && (156)\\
&= 0 && (157)\\
\mathbb{E}\left[\mathcal{E}_{ii}\tilde{Z}_i u_i + \sum_{j\neq i}\mathcal{E}_{ij}\tilde{Z}_i u_j + \tilde{Z}_i \epsilon_i | \mathit{\Theta}\right] &= 0 && (158)\\
\mathbb{E}\left[\mathit{\Delta} P_i \tilde{Z}_i | \mathit{\Theta}\right] &= \lambda_i var(\tilde{Z}) && (159)
\end{aligned}$$

Then:

$$
\begin{aligned}
cov\left(\Delta P_i, \tilde{Z}_i\right) &= \mathbb{E}\left[\mathbb{E}\left[\Delta P_i \tilde{Z}_i | \Theta\right]\right] = \mathbb{E}\left[\lambda_i var(\tilde{Z})\right] = var(\tilde{Z})\mathbb{E}\left[\lambda_i\right] && (160)\\
cov\left(\Delta D_i, \tilde{Z}_i\right) &= \mathbb{E}\left[\mathbb{E}\left[\Delta D_i \tilde{Z}_i | \Theta\right]\right] && (161)\\
&= var(\tilde{Z})\left(\mathbb{E}\left[\mathcal{E}_{ii}\lambda_i\right] - \mathbb{E}\left[\lambda_i\right]\mathbb{E}\left[\mathcal{E}_{i,cross}\right] + \frac{1}{(N-1)^2}\mathbb{E}\left[\mathcal{E}_{i,cross}\tilde{\lambda}_i\right]\right) && (162)
\end{aligned}
$$

The instrumental variable regression with heterogenous treatment effects identifies:

$$
\begin{aligned}
\widehat{\mathcal{E}} &= \frac{\text{cov}\left(\Delta D_i, \tilde{Z}_i\right)}{\text{cov}\left(\Delta P_i, \tilde{Z}_i\right)} && (163)\\
&= \frac{\mathbb{E}\left[\lambda_i \mathcal{E}_{ii}\right]}{\mathbb{E}\left[\lambda_i\right]} - \mathbb{E}\left[\mathcal{E}_{i,cross}\right] + \frac{N}{(N-1)^2}\mathbb{E}\left[\left(\frac{\lambda_i}{\mathbb{E}\left[\lambda_i\right]} - 1\right)\mathcal{E}_{i,cross}\right] && (164)
\end{aligned}
$$

We can rewrite the middle term of equation (164) as:

$$
\begin{aligned}
\mathbb{E}_i\left[\mathcal{E}_{i,cross}\right] &= \mathbb{E}_i\left[\frac{\mathbb{E}_{j\neq i}\left[\lambda_j \mathcal{E}_{ij}\right]}{\mathbb{E}_{j\neq i}\left[\lambda_j\right]}\right] && (165)\\
&= \mathbb{E}_{j\neq i}\left[\frac{\mathbb{E}_{i}\left[\lambda_j \mathcal{E}_{ij}\right]}{\mathbb{E}_{j\neq i}\left[\lambda_j\right]}\right] && (166)\\
&= \mathbb{E}_j\left[\frac{\lambda_j}{\mathbb{E}_j\left[\lambda_j\right]}\mathbb{E}_{i\neq j}\left[\frac{\mathbb{E}_j\left[\lambda_j\right]}{\mathbb{E}_{j\neq i}\left[\lambda_j\right]}\mathcal{E}_{ij}\right]\right] && (167)\\
&= \mathbb{E}_j\left[\frac{\lambda_j}{\mathbb{E}_j\left[\lambda_j\right]}\bar{\mathcal{E}}_{.j}\right] && (168)
\end{aligned}
$$

The term $\bar{\mathcal{E}}_{.j}$ resembles an equal-weighted average over rows of $\mathcal{E}_{ij}$ excluding the diagonal element.

The right term of equation (164) is a small-sample term that goes to zero when $N \to +\infty$:

$$
\begin{aligned}
\operatorname*{plim}_{N\to+\infty} \widehat{\mathcal{E}} &= \mathbb{E}\left[\omega_i\left(\mathcal{E}_{ii} - \bar{\mathcal{E}}_{.i}\right)\right], && (169)\\
\text{with}\quad \omega_i &= \frac{\lambda_i}{\mathbb{E}\left[\lambda_i\right]}. && (170)
\end{aligned}
$$

The limit is an average of the own- and cross elasticities weighted by λ_i. ■

B Robustness

B.1 Robustness to the assumptions

We assess the robustness of the identification result of Proposition 2 with respect to deviations from the assumptions. For simplicity, we focus on a case with two assets and deviations from Assumption A2. The argument generalizes to other types of deviations.

We first show that the conclusions are robust when the first stage is strong. Then we illustrate potential issues in presence of weak instruments in the context of a model.

B.1.1 Robustness to deviations from Assumption A2

Consider a setting with two assets and an arbitrary elasticity matrix. In response to an exogenous shock, the relative change in demand is:

$$\begin{aligned} \mathit{\Delta} D_1 - \mathit{\Delta} D_2 &= \mathcal{E}_{11}\mathit{\Delta} P_1 + \mathcal{E}_{12}\mathit{\Delta} P_2 - \mathcal{E}_{22}\mathit{\Delta} P_2 - \mathcal{E}_{21}\mathit{\Delta} P_1 && (171)\\ &= (\mathcal{E}_{11} - \mathcal{E}_{21})\mathit{\Delta} P_1 + (\mathcal{E}_{22} - \mathcal{E}_{12})\mathit{\Delta} P_2 && (172)\end{aligned}$$

We denote the relative elasticities by $\mathcal{E}_{rel,1} = \mathcal{E}_{11} - \mathcal{E}_{21}$ and $\mathcal{E}_{rel,2} = \mathcal{E}_{22} - \mathcal{E}_{12}$. Rearranging the terms leads to

$$\mathit{\Delta} D_1 - \mathit{\Delta} D_2 = \frac{\mathcal{E}_{rel,1} + \mathcal{E}_{rel,2}}{2}\left(\mathit{\Delta} P_1 - \mathit{\Delta} P_2\right) + \frac{\mathcal{E}_{rel,1} - \mathcal{E}_{rel,2}}{2}\left(\mathit{\Delta} P_1 + \mathit{\Delta} P_2\right)/ \tag{173}$$

Dividing by the relative change in price in response to the shock, we obtain the estimator:

$$\frac{\mathit{\Delta} D_1 - \mathit{\Delta} D_2}{\mathit{\Delta} P_1 - \mathit{\Delta} P_2} = \frac{\mathcal{E}_{rel,1} + \mathcal{E}_{rel,2}}{2} + \frac{\mathcal{E}_{rel,1} - \mathcal{E}_{rel,2}}{2}\frac{\mathit{\Delta} P_1 + \mathit{\Delta} P_2}{\mathit{\Delta} P_1 - \mathit{\Delta} P_2} \tag{174}$$

The first term is the average relative elasticity. The second term is the potential bias: the heterogeneity of relative elasticities times the ratio of sum of changes in prices to their difference. The denominator $\mathit{\Delta} P_1 - \mathit{\Delta} P_2$ is the first stage of the regression, and the relevance condition is $\mathit{\Delta} P_1 - \mathit{\Delta} P_2 \neq 0$.

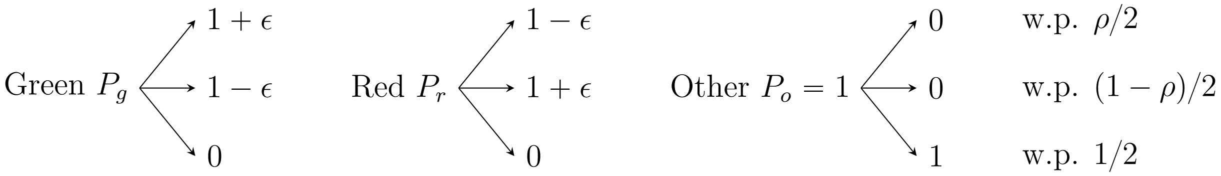
It is straightforward to see that if the relevance condition is satisfied, the identification result of relative elasticity is robust to small deviations from assumption A2. Formally, consider a family of experiments (or models) indexed by a variable x, such that the experiment for $x = 0$ satisfies assumption A2 but it does not otherwise. If elasticities and changes in prices are continuous in x, then the IV estimator is also continous in x as long as $\mathit{\Delta} P_1 - \mathit{\Delta} P_2 \neq 0$.

In presence of a weak instrument, when the relevance condition is not satisfied, the last term of equation (174) becomes infinite. The bias becomes large relative to the actual coefficient. We show so in a model example next. Of course, in practice, one can simply assess the first stage empirically and should not proceed anyways without a strong instrument.

B.1.2 An example with discontinuous estimates and a weak first stage

Setting. We consider a variation from the model of Section 2.4.2, where the probability of the different states is asymmetric. Furthermore, we measure elasticities in the "wrong" units, portfolio share on price rather than portfolio share on log price. These two features create deviations from the Assumption A2. The experiment is a shock to the endowment of the green asset E_g.

The setting is the same as before except for a change in the probability of the states:



The optimal portfolio shares are:

$$\omega_g\left(P_g, P_r\right) = \frac{P_g\left(\left(\epsilon^2-1\right)P_g + P_r\left(4\rho\epsilon + (\epsilon-1)^2\right)\right)}{4\left(\epsilon^2+1\right)P_gP_r + 2\left(\epsilon^2-1\right)P_g^2 + 2\left(\epsilon^2-1\right)P_r^2} \tag{175}$$

$$\omega_r\left(P_g, P_r\right) = \frac{P_r\left(P_g\left((\epsilon+1)^2 - 4\rho\epsilon\right) + \left(\epsilon^2-1\right)P_r\right)}{4\left(\epsilon^2+1\right)P_gP_r + 2\left(\epsilon^2-1\right)P_g^2 + 2\left(\epsilon^2-1\right)P_r^2} \tag{176}$$

Equilibrium and elasticities. Assume that the endowments are $E_g = E_r = 1/2$ and $E_o = 1$. Then equilibrium prices are

$$P_g = 1 - \epsilon(1-2\rho), \qquad P_r = 1 + \epsilon(1-2\rho). \tag{177}$$

At this equilibrium, the elasticity matrix of the portfolio shares with respect to the level of prices for the green and red asset is:

$$\mathcal{E}_{gg} = \frac{\partial \omega_g}{\partial P_g} = \frac{\left(\epsilon^2-1\right)\left((2\rho-1)\epsilon-1\right)}{32(\rho-1)\rho\epsilon^2}, \tag{178}$$

$$\mathcal{E}_{rr} = \frac{\partial \omega_r}{\partial P_r} = -\frac{\left(\epsilon^2-1\right)\left((2\rho-1)\epsilon+1\right)}{32(\rho-1)\rho\epsilon^2}, \tag{179}$$

$$\mathcal{E}_{gr} = \frac{\partial \omega_g}{\partial P_r} = \frac{\left(\epsilon^2-1\right)\left((2\rho-1)\epsilon+1\right)}{32(\rho-1)\rho\epsilon^2}, \tag{180}$$

$$\mathcal{E}_{rg} = \frac{\partial \omega_r}{\partial P_g} = -\frac{\left(\epsilon^2-1\right)\left((2\rho-1)\epsilon-1\right)}{32(\rho-1)\rho\epsilon^2}. \tag{181}$$

This leads to the two relative elasticities:

$$\mathcal{E}_{rel,g} = \mathcal{E}_{gg} - \mathcal{E}_{rg} = \frac{\left(\epsilon^2-1\right)\left((2\rho-1)\epsilon-1\right)}{16(\rho-1)\rho\epsilon^2}, \tag{182}$$

$$\mathcal{E}_{rel,r} = \mathcal{E}_{rr} - \mathcal{E}_{gr} = \frac{\left(\epsilon^2-1\right)\left(-(2\rho-1)\epsilon-1\right)}{16(\rho-1)\rho\epsilon^2} \tag{183}$$

Then the terms from the difference-in-difference estimator are:

$$\frac{\mathcal{E}_{rel,g}+\mathcal{E}_{rel,r}}{2} = \frac{1-\epsilon^2}{16(\rho-1)\rho\epsilon^2} \tag{184}$$

$$\frac{\mathcal{E}_{rel,g}-\mathcal{E}_{rel,r}}{2} = \frac{(\epsilon^2-1)(2\rho-1)}{16(\rho-1)\rho\epsilon^2} \tag{185}$$

Note that constant relative elasticity, assumption A2, holds only if $\rho = \frac{1}{2}$. We work in a neighborhood of assumption A2, where $\rho \sim \frac{1}{2}$. To weaken the first stage and make the assets perfect substitute, we take ϵ to zero. These expressions become approximately

$$\frac{\mathcal{E}_{rel,g}+\mathcal{E}_{rel,r}}{2} \approx \frac{-1}{4\epsilon^2} \tag{186}$$

$$\frac{\mathcal{E}_{rel,g}-\mathcal{E}_{rel,r}}{2} \approx \frac{2\rho-1}{4\epsilon^2} \tag{187}$$

Equilibrium prices as a function of the endowments are

$$P_g\left(E_o,E_g,E_r\right) = \frac{E_o\left(\left(\epsilon^2-1\right)E_g - E_r\left(4\rho\epsilon+(\epsilon-1)^2\right)\right)}{-2\left(\epsilon^2+1\right)E_gE_r+\left(\epsilon^2-1\right)E_g^2+\left(\epsilon^2-1\right)E_r^2} \tag{188}$$

$$P_r\left(E_o,E_g,E_r\right) = \frac{E_o\left(\left(\epsilon^2-1\right)E_r - E_g\left((\epsilon+1)^2-4\rho\epsilon\right)\right)}{-2\left(\epsilon^2+1\right)E_gE_r+\left(\epsilon^2-1\right)E_g^2+\left(\epsilon^2-1\right)E_r^2}. \tag{189}$$

Around the initial equilibrium, the changes in prices are:

$$\Delta P_g = \frac{\partial P_g}{\partial E_g} = -4\epsilon\rho-(\epsilon-1)^2 \tag{190}$$

$$\Delta P_r = \frac{\partial P_g}{\partial E_g} = \epsilon^2-1 \tag{191}$$

The term controlling the bias is:

$$\frac{\Delta P_g+\Delta P_r}{\Delta P_g-\Delta P_r} = \frac{2\epsilon\rho-\epsilon+1}{\epsilon(2\rho+\epsilon-1)} \tag{192}$$

Putting it all together. We will study what happens around $\rho = 1/2$, so, echoing our general setup, we call $x = 2\rho-1$. When $x=0$, assumption A2 is satisfied, and the estimator is unbiased.

We plug all the expressions above in equation (174):

$$\frac{\Delta D_1-\Delta D_2}{\Delta P_1-\Delta P_2} = \frac{\mathcal{E}_{rel,1}+\mathcal{E}_{rel,2}}{2} + \frac{\mathcal{E}_{rel,1}-\mathcal{E}_{rel,2}}{2}\frac{\Delta P_1+\Delta P_2}{\Delta P_1-\Delta P_2} \tag{193}$$

$$\approx \frac{-1}{4\epsilon^2} + \frac{x}{4\epsilon^2}\frac{1}{\epsilon(x+\epsilon)} \tag{194}$$

Both the average relative elasticity and the difference in relative elasticity go to infinity at the same pace ($1/\epsilon^2$). However, the weak first stage amplifies the bias by another order of magnitude. To visualize this issue, it is more natural to compute the relative bias of the estimator:

$$\frac{\frac{\Delta D_1 - \Delta D_2}{\Delta P_1 - \Delta P_2} - \frac{\mathcal{E}_{rel,1} + \mathcal{E}_{rel,2}}{2}}{\frac{\mathcal{E}_{rel,1} + \mathcal{E}_{rel,2}}{2}} \approx -\frac{x}{\epsilon(x+\epsilon)} \tag{195}$$

The bias term present when $x \neq 0$ is an order of magnitude large than the correct estimate in the limit of a weak instrument.

C Demand beyond risk-based motives for substitution

Consider the problem of investors combining risk-based mean-variance demand where the covariance matrix Σ is characterized by a set of characteristics $X^{(3)}$ with a cost of holding assets that is quadratic in another set of characteristics $X^{(1)}$ and a portfolio constraint linear in yet another set characteristics $X^{(2)}$. Section 2.3.4 is a special case of this that assumes that $X^{(1)}$ and $X^{(2)}$ each only contain one observable: carbon intensity and a bank's liquidity ratio. The proposition below generalizes this.

Proposition 13 (Mean-variance demand with quadratic cost and linear constraint) *Assume that investors choose their demand according to the problem*

$$\max_{D} \quad D'(M-P) - \frac{\gamma}{2} D' \Sigma D - \frac{\kappa}{2} D' X^{(1)} X^{(1)\prime} D \tag{196}$$

$$\textit{such that} \quad D' X^{(2)} \leqslant \Theta, \tag{197}$$

where D, M, and P are the $N \times 1$ vectors of investor demand, expected payoffs, and prices, γ is risk aversion, Σ the $N \times N$ covariance matrix, κ controls the quadratic cost function, $X^{(1)}$ and $X^{(2)}$ are the $N \times K_1$ and $N \times K_2$ matrices of stock characteristics, and Θ is a $1 \times K_2$ vector that controls the linear constraint.

Further assume that the risk-based component of investor demand satisfies assumptions A1 and A2, i.e.,

$$-\frac{1}{\gamma} \Sigma^{-1} = \widehat{\mathcal{E}}^{(3)} I + X^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime}, \tag{198}$$

where $X^{(3)}$ is another $N \times K_3$ matrix of stock characteristics, and $\mathcal{E}_{X^{(3)}}$ the $K_3 \times K_3$ matrix of substitution between observables.

Then the resulting demand curve satisfies assumptions A1 and A2 conditional on the stacked observables $\mathbb{X} = [X^{(1)}, X^{(2)}, X^{(3)}]$.

Proof. By Lemma 8, to proof the proposition, we need to show that the elasticity matrix $\mathcal{E}$ can be expressed as

$$\mathcal{E} = \widehat{\mathcal{E}} I + \mathbb{X} \mathcal{E}_{\mathbb{X}} \mathbb{X}'. \tag{199}$$

Start by putting together equations (196) and (197) in the Lagrangian

$$\mathcal{L}(D, \lambda) = D'(M - P) - \frac{\gamma}{2} D' \varSigma D - \frac{\kappa}{2} D' X^{(1)} X^{(1)\prime} D - \lambda(D' X^{(2)} - \varTheta), \tag{200}$$

where λ is the $K_2 \times 1$ Lagrange multiplier on the linear constraint.
Setting the first-order condition with respect to D to zero, and solving for D, yields

$$\begin{aligned} D &= \Bigg(\underbrace{\gamma \varSigma + \kappa X^{(1)} X^{(1)\prime}}_{\equiv \varOmega}\Bigg)^{-1} \left(M - P - X^{(2)} \lambda\right) && (201) \\ &= \varOmega^{-1} \left(M - P - X^{(2)} \lambda\right), && (202) \end{aligned}$$

where

$$\varOmega = \gamma \varSigma + \kappa X^{(1)} X^{(1)\prime}. \tag{203}$$

Plugging into the linear constraint to solve for λ:

$$\begin{aligned} & \left(M - P - X^{(2)} \lambda\right)' \varOmega^{-1} X^{(2)} && = \varTheta && (204) \\ \Longrightarrow\ & (M - P)' \varOmega^{-1} X^{(2)} - \lambda' X^{(2)\prime} \varOmega^{-1} X^{(2)} && = \varTheta && (205) \\ \Longrightarrow\ & \lambda = \left[X^{(2)\prime} \varOmega^{-1} X^{(2)}\right]^{-1} \left[X^{(2)\prime} \varOmega^{-1} (M - P) - \varTheta'\right]^{+} && && (206) \end{aligned}$$

Plugging the Lagrange multipliers back into optimal investor demand gives:

$$D = \varOmega^{-1} (M - P) - \varOmega^{-1} X^{(2)} \left[X^{(2)\prime} \varOmega^{-1} X^{(2)}\right]^{-1} \left[X^{(2)\prime} \varOmega^{-1} (M - P) - \varTheta'\right]^{+} \tag{207}$$

The elasticity matrix therefore is:

$$\frac{dD}{dP} = -\varOmega^{-1} + \varOmega^{-1} X^{(2)} S_b \left[S_b' X^{(2)\prime} \varOmega^{-1} X^{(2)} S_b\right]^{-1} S_b' X^{(2)\prime} \varOmega^{-1} \tag{208}$$

Here, S_b is the binding constraint selection matrix, which for the first-order condition selects the columns of $X^{(2)}$ for which constraints are binding.

Start now with the part for when the inequality constraints are all non-binding:

$$\begin{aligned}
-\Omega^{-1} =& \widehat{\mathcal{E}}^{(3)} I + X^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime} \qquad (209)\\
&+ \left(\widehat{\mathcal{E}}^{(3)} I + X^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime}\right) X^{(1)} \underbrace{\left[\frac{1}{\kappa} I - X^{(1)\prime} \left(\gamma \Sigma\right)^{-1} X^{(1)}\right]^{-1}}_{\equiv \mathcal{H}} X^{(1)\prime} \left(\widehat{\mathcal{E}}^{(3)} I + X^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime}\right) \qquad (210)\\
=& \widehat{\mathcal{E}}^{(3)} I + X^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime} + \left(\widehat{\mathcal{E}}^{(3)}\right)^2 X^{(1)} \mathcal{H} X^{(1)\prime} + \widehat{\mathcal{E}}^{(3)} X^{(1)} \mathcal{H} X^{(1)\prime} X^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime} \qquad (211)\\
&+ \widehat{\mathcal{E}}^{(3)} X^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime} X^{(1)} \mathcal{H} X^{(1)\prime} + X^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime} X^{(1)} \mathcal{H} X^{(1)\prime} X^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime} \qquad (212)\\
=& \widehat{\mathcal{E}}^{(3)} I + \underbrace{[X^{(1)}, X^{(3)}]}_{\equiv X^{(1,3)}} \underbrace{\begin{bmatrix} \left(\widehat{\mathcal{E}}^{(3)}\right)^2 \mathcal{H} & \widehat{\mathcal{E}}^{(3)} \mathcal{H} X^{(1)\prime} X^{(3)} \mathcal{E}_{X^{(3)}} \\ \widehat{\mathcal{E}}^{(3)} \mathcal{E}_{X^{(3)}} X^{(3)\prime} X^{(1)} \mathcal{H} & \mathcal{E}_{X^{(3)}} + \mathcal{E}_{X^{(3)}} X^{(3)\prime} X^{(1)} \mathcal{H} X^{(1)\prime} X^{(3)} \mathcal{E}_{X^{(3)}} \end{bmatrix}}_{\equiv \mathcal{F}} [X^{(1)}, X^{(3)}]' \qquad (213)\\
=& \widehat{\mathcal{E}}^{(3)} I + X^{(1,3)} \mathcal{F} X^{(1,3)\prime} \qquad (214)
\end{aligned}$$

When the linear constraints are not binding, the elasticity matrix satisfies assumptions A1 and A2 conditional on the stacked observables $[X^{(1)}, X^{(3)}]$.

For the case that some constraints are not binding, define:

$$\mathcal{G} \equiv S_b \left[S_b' X^{(2)\prime} \Omega^{-1} X^{(2)} S_b\right]^{-1} S_b' \tag{215}$$

The elasticity matrix is

$$\begin{aligned}
\frac{dD}{dP} &= \widehat{\mathcal{E}}^{(3)} I + X^{(1,3)} \mathcal{F} X^{(1,3)\prime} + \left(\widehat{\mathcal{E}}^{(3)} I + X^{(1,3)} \mathcal{F} X^{(1,3)\prime}\right) X^{(2)} \mathcal{G} X^{(2)\prime} \left(\widehat{\mathcal{E}}^{(3)} I + X^{(1,3)} \mathcal{F} X^{(1,3)\prime}\right) \qquad (216)\\
&= \widehat{\mathcal{E}} I + \mathbb{X} \mathcal{E}_{\mathbb{X}} \mathbb{X}', \qquad (217)
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\mathcal{E}} &= \widehat{\mathcal{E}}^{(3)} \qquad (218)\\
\mathbb{X} &= \left[X^{(1)}, X^{(3)}, X^{(2)}\right] \qquad (219)\\
\mathcal{E}_{\mathbb{X}} &= \begin{bmatrix} \mathcal{F} + \mathcal{F} X^{(1,3)\prime} X^{(2)} \mathcal{G} X^{(2)\prime} X^{(1,3)} \mathcal{F} & \widehat{\mathcal{E}}^{(3)} \mathcal{F} X^{(1,3)\prime} X^{(2)} \mathcal{G} \\ \widehat{\mathcal{E}}^{(3)} \mathcal{G} X^{(2)\prime} X^{(1,3)} \mathcal{F} & \left(\widehat{\mathcal{E}}^{(3)}\right)^2 \mathcal{G} \end{bmatrix}. \qquad (220)
\end{aligned}$$

The elasticity matrix satisfies assumptions A1 and A2 conditional on the stacked observables $\mathbb{X}$.

■

D A non-linear framework

We derive properties for a family of non-linear demand functions which satisfy locally our assumption of homogeneous substitution conditional on observables. Doing so provides more general intuition behind our results in linear structures.

Because the non-linear structural models considered in Koijen and Yogo (2019) also belong to this family of demand functions, this framework also allows us to better understand the connection of our results with properties of those models. In particular, we explain the restrictions imposed by the logit form relative to arbitrary factor models, simple factor models (with constant variance and expected payoffs), and more general demand functions.

D.1 Basic concepts

We consider a setting with an investor, N assets indexed by i, and K observables for each asset. We start with a general demand function defined as a mapping from the vector of (log) prices $\boldsymbol{p}$ and the $N \times K$ matrix of observables $\boldsymbol{x}$ to a vector of positions D (portfolio shares in our applications):

$$D(\boldsymbol{p}, \boldsymbol{x}) : \mathbb{R}^N \times \mathbb{R}^{N \times K} \to \mathbb{R}^N$$

It will be helpful to define the following property.

Definition 14 (HCO functions) *A function $F : \mathbb{R}^N \times \mathbb{R}^{N \times K} \to \mathbb{R}^N$ is homogenous-conditional-on-observables (HCO) if $\forall i, [F(\boldsymbol{p}, \boldsymbol{x})]_i = f(p_i, x_i; \boldsymbol{p}, \boldsymbol{x})$ for a function $f : (\mathbb{R} \times \mathbb{R}^K) \times (\mathbb{R}^N \times \mathbb{R}^{N \times K}) \to \mathbb{R}$ for each i.*

That is, for a fixed overall price and observables vector, the value for each element is given by the same (scalar-valued) function of its own price and observables.

Then, in the spirit of the discussion of properties of factor models of Koijen and Yogo (2019), we can restrict attention to a subset of general demand functions as follows.

Definition 15 (HCO demand) *A demand function is a homogenous-conditional-on-observables demand if it is a HCO function.*

With HCO demand functions, individual positions can be written as:

$$\left[D\left(\boldsymbol{p}, \boldsymbol{x}\right)\right]_i = d\left(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}\right).$$

This notation emphasizes the dual role of prices and observables. On the one hand, the same function $d(\cdot, \cdot; \boldsymbol{p}, \boldsymbol{x})$ describes how the demand of each asset depends on its own price and own observables only. On the other hand, this mapping varies with the vector $(\boldsymbol{p}, \boldsymbol{x})$. Thinking of this vector as the state of the economy, a HCO demand describes a mapping which is possibly state-dependent, but identical across assets.

Naturally, the choice of observables is what gives meaningful restrictions to this definition. For example, if the observables $\boldsymbol{x}$ includes each asset's "name", i, then all demand functions are also HCO demand.

An example of HCO demand is logit:

$$[D^{logit}(\boldsymbol{p}, \boldsymbol{x})]_i = \frac{\exp\left(-\alpha p_i + \beta' x_i\right)}{1 + \sum_{j=1}^{N} \exp\left(-\alpha p_j + \beta' x_j\right)},$$

because the numerator is a function of p_i and x_i only, while the denominator is a fixed function (i.e. that does not depend on i) of $\boldsymbol{p}$ and $\boldsymbol{x}$.

D.2 Relative elasticity vs. substitution and identification.

HCO demand leads to a natural decomposition of the elasticity matrix between a relative elasticity and a substitution matrices. HCO demand implies an elasticity matrix:

$$\mathcal{E} = \frac{\partial D}{\partial p} = \underbrace{diag\left(\frac{\partial d}{\partial p_i}\left(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}\right)\right)}_{\text{relative elasticity, } N \times N} + \underbrace{\begin{bmatrix} \frac{\partial d}{\partial p}\left(p_1, x_1; \boldsymbol{p}, \boldsymbol{x}\right)' \\ \vdots \\ \underbrace{\frac{\partial d}{\partial p}\left(p_N, x_N; \boldsymbol{p}, \boldsymbol{x}\right)'}_{1 \times N} \end{bmatrix}}_{\text{substitution, } N \times N}$$

where the derivative in the second term is with respect to the third argument of d, not a total derivative. We call the first term relative elasticity and the second one substitution. To understand why the first one is a relative elasticity, notice that if two assets have the same price and observables, this term is equal to the difference between their own-price and cross-price elasticity:

$$\begin{aligned} \mathcal{E}_{ii} - \mathcal{E}_{ji} &= \underbrace{\left(\frac{\partial d}{\partial p_i}\left(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}\right) + \left[\frac{\partial d}{\partial p}\left(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}\right)\right]_i\right)}_{\mathcal{E}_{ii}} - \underbrace{\left[\frac{\partial d}{\partial p}\left(p_j, x_j; \boldsymbol{p}, \boldsymbol{x}\right)\right]_i}_{\mathcal{E}_{ji}} \\ &= \frac{\partial d}{\partial p_i}\left(p_i, x_i; \boldsymbol{p}, \boldsymbol{x}\right) \text{ if } \left(p_i, x_i\right) = \left(p_j, x_j\right) \end{aligned}$$

The substitution matrix captures how investor reallocate between assets when their price change. Any HCO demand satisfies homogeneous substitution conditional on all observables and the price:

$$\mathcal{E}_{il} = \mathcal{E}_{jl} = \left[\frac{\partial d}{\partial p}(p_i, x_i, \boldsymbol{p}, \boldsymbol{x})\right]_l \text{ if } \left(p_i, x_i\right) = \left(p_j, x_j\right)$$

Indeed, this corresponds to assumption A1 in the text when the observables $\boldsymbol{x}$ are variables that the econometrician can measure.

Identification. The cross-section can allow to identify relative elasticity by comparing demand for two assets with the same observables but nearby prices. However, because $(\boldsymbol{p}, \boldsymbol{x})$ are fixed in a given cross-section, identification of substitution is generally impossible with the cross-section.

This limitation of the cross-section can be overcome by imposing additional restrictions. For example, when there is no substitution, that is the demand function does not depend on the price vector per se (its third argument), we have: $d\left(p_i, x_i, \boldsymbol{p}, \boldsymbol{x}\right) = d(p_i, x_i; \boldsymbol{x})$. Elasticity is relative elasticity, and hence can be estimated from the cross-section alone.

Another case is logit. Even though this model has non-zero substitution, it can be estimated from the cross-section because parameters determining substitution can be identified by measuring relative elasticity. Specifically the relative elasticity vector is $-\alpha\omega$ and the substitution matrix is $\alpha\omega\omega'$ where $\omega = D(\boldsymbol{p}, \boldsymbol{x})$ is the realized vector of portfolio weights. This calculation also highlights that the structure of substitution is very restricted in logit: the substitution matrix of rank 1, and the effects must be proportional to portfolio weights. To better understand how limiting these restrictions are, we compare logit to other demand models in the following section.

D.3 Logit, log utility and factor models

Our main interest in this section is to what extent the demand of an investor with a standard utility function and particular views on the dynamics of expected returns might be represented with logit demand.

For this purpose, we consider the demand of a log investor with log-normal returns (like in Section 2.4.1) as the simplest example of a standard utility.

$$D^{log}(\boldsymbol{p}, \boldsymbol{x}) = \mathit{\Sigma}(\boldsymbol{p}, \boldsymbol{x})^{-1}\mu(\boldsymbol{p}, \boldsymbol{x})$$

where $\mu(\boldsymbol{p}, \boldsymbol{x})$ is the expected return vector and $\mathit{\Sigma}(\boldsymbol{p}, \boldsymbol{x})$ is the return covariance matrix.

For the representation of views of the investor on the structure of asset returns, we define first a class of factor models where factor loadings might be state-dependent.

Definition 16 (General factor models) *A general M-factor model is defined by functions* $\mu(\boldsymbol{p}, \boldsymbol{x}) : \mathbb{R}^N \times \mathbb{R}^{N\times K} \to \mathbb{R}^N$ *and* $\mathit{\Sigma}(\boldsymbol{p}, \boldsymbol{x}) : \mathbb{R}^N \times \mathbb{R}^{N\times K} \to \mathbb{R}^{N\times N}$ *describing the expected return vector and covariance matrix respectively such that*

$$\mathit{\Sigma}(\boldsymbol{p}, \boldsymbol{x}) = diag(\underbrace{\sigma^2_\epsilon(\boldsymbol{p}, \boldsymbol{x})}_{N\times 1}) + \underbrace{\beta(\boldsymbol{p}, \boldsymbol{x})}_{N\times M}\underbrace{\mathit{\Sigma}_F}_{M\times M}\beta(\boldsymbol{p}, \boldsymbol{x})'$$

, where $\mu(\boldsymbol{p}, \boldsymbol{x})$*,* $\sigma^2_\epsilon(\boldsymbol{p}, \boldsymbol{x})$ *and each column of* $\beta(\boldsymbol{p}, \boldsymbol{x})$ *are HCO functions and* $\mathit{\Sigma}_F$ *is a covariance matrix.*

A general M-factor model lets the functions mapping prices and observables to expected payoffs, to factor loadings and to idosyncratic risk for each asset to freely vary with the state of the economy $(\boldsymbol{p}, \boldsymbol{x})$. As we show below, this is a rich enough set that for any HCO demand $D^{HCO}(\boldsymbol{p}, \boldsymbol{x})$ one can always find a particular factor model that the log investor's demand exactly corresponds to the choosen HCO demand: $D^{HCO}(\boldsymbol{p}, \boldsymbol{x}) = D^{log}(\boldsymbol{p}, \boldsymbol{x})$. This is a generalized version of Corrollary 1 of Koijen and Yogo (2019) who specialize the function $D^{HCO}(\boldsymbol{p}, \boldsymbol{x})$ to be the logit demand.

Proposition 17 *Fix two functions $\sigma_\epsilon^2(\boldsymbol{p},\boldsymbol{x})$ and an $N \times 1$ $\beta(\boldsymbol{p},\boldsymbol{x})$ which are HCO. For any HCO demand $D^{HCO}(\boldsymbol{p},\boldsymbol{x})$, there exists a general 1-factor model with the corresponding covariance matrix $\Sigma(\boldsymbol{p},\boldsymbol{x}) = \mathrm{diag}(\sigma_\epsilon^2(\boldsymbol{p},\boldsymbol{x})) + \sigma_F^2 \underbrace{\beta(\boldsymbol{p},\boldsymbol{x})}_{N\times 1} \beta(\boldsymbol{p},\boldsymbol{x})'$, such that log utility demand with this factor model yields the same demand function.*

Proof. Choose $\mu(\boldsymbol{p},\boldsymbol{x}) = \Sigma(\boldsymbol{p},\boldsymbol{x}) D^{HCO}(\boldsymbol{p},\boldsymbol{x})$ then clearly $D^{log}(\boldsymbol{p},\boldsymbol{x}) = D^{HCO}(\boldsymbol{p},\boldsymbol{x})$. Therefore, we have to show only that $\mu(\boldsymbol{p},\boldsymbol{x})$ is HCO. For this, note that

$$\left[\Sigma(\boldsymbol{p},\boldsymbol{x}) D^{HCO}(\boldsymbol{p},\boldsymbol{x})\right]_i = \\ \left[\sigma_\epsilon^2(\boldsymbol{p},\boldsymbol{x})\right]_i d^{HCO}(p_i,\boldsymbol{x}_i;\boldsymbol{p},\boldsymbol{x}) + \left[\beta(\boldsymbol{p},\boldsymbol{x}) \underbrace{\beta'(\boldsymbol{p},\boldsymbol{x}) D^{HCO}(\boldsymbol{p},\boldsymbol{x})}_{scalar}\right]_i$$

which, given that $\beta(\boldsymbol{p},\boldsymbol{x})$ is HCO gives the proof. ■

This result shows that for any HCO demand, and a fortiori for logit demand, there exist factor models that microfound it. However, the reverse is clearly not true: an arbitrary factor model does not give rise to logit demand. To move closer to common finance intuition, we consider a restricted class of models often used to think about portfolio choice with stable variance and expected payoffs.

Definition 18 (Stable factor model) *A stable M-factor model (based on observables) is a factor model where expected payoff idiosyncratic risk and factor loadings depend on observables only: $\mu(\boldsymbol{p},\boldsymbol{x}) = M(\boldsymbol{x}) - \boldsymbol{p}$, $\sigma_\epsilon^2(\boldsymbol{p},\boldsymbol{x}) = \sigma_\epsilon^2(\boldsymbol{x})$, and $\beta(\boldsymbol{p},\boldsymbol{x}) = \beta(\boldsymbol{x})$ where $M(\boldsymbol{x})$ and $\sigma_\epsilon^2(\boldsymbol{x})$ and the columns of $\beta(\boldsymbol{x})$ are all HCO.*

One might wonder if there are stable factor models that the demand of a log investor can be represented with the logit form. This cannot hold overall due to different functional forms: the stable factor model is linear in log prices. To make the comparison more meaningful we ask if such an equivalence holds locally. To do so, we define first-order equivalence.

Definition 19 *Two demand functions $D^1(\boldsymbol{p},\boldsymbol{x})$ and $D^2(\boldsymbol{p},\boldsymbol{x})$ are first-order equivalent around a point $(\boldsymbol{p}_0,\boldsymbol{x}_0)$ if they have same value and Jacobian with respect to the price matrix at that point:*

$$\begin{aligned} D^1(\boldsymbol{p}_0,\boldsymbol{x}_0) &= D^2(\boldsymbol{p}_0,\boldsymbol{x}_0) \\ \underbrace{\frac{\partial D^1}{\partial p}(\boldsymbol{p}_0,\boldsymbol{x}_0)}_{N\times N} &= \frac{\partial D^2}{\partial p}(\boldsymbol{p}_0,\boldsymbol{x}_0) \end{aligned}$$

We obtain that in general the answer is negative.

Proposition 20 *Out of the set of all stable M-factor models, there is only one under which a logit demand model with $\alpha > 0$ can be first-order equivalent to the demand of the log investor. This specific model has 1 factor with identical factor loadings and idiosyncratic variance inversely proportional to demand. Under any other stable factor model, logit demand is not a valid approximation of the demand of log investor.*

Proof. For a stable factor model, we have $D^{log}(\boldsymbol{p}, \boldsymbol{x}) = \Sigma(x)^{-1}(M(x) - p)$, so $\partial D^{log}/\partial p = -\Sigma(x)^{-1}$. For logit we have: $\partial D^{logit}/\partial p = -\alpha diag(\omega)(I - \mathbf{1}\omega') = -\alpha diag(\omega) + \alpha\omega\omega'$, where ω is the investor's portfolio share vector. We can invert it with the Sherman-Morrison formula and identify with Σ:

$$\begin{aligned} -\left(\partial D^{logit}/\partial p\right)^{-1} &= \alpha^{-1}\left(I - \mathbf{1}\omega'\right)^{-1} diag(\omega)^{-1} \\ &= \alpha^{-1}\left(I + \frac{1}{1-\omega'\mathbf{1}}\mathbf{1}\omega'\right) diag(\omega)^{-1} \\ \Sigma &= \underbrace{\alpha^{-1} diag(\omega)^{-1}}_{\text{idiosyncratic risk}} + \underbrace{\alpha^{-1}\frac{1}{1-\omega'\mathbf{1}}\mathbf{1}\mathbf{1}'}_{\text{single factor}} \end{aligned}$$

Comparing this to the covariance matrix under a generic stable factor model,

$$\Sigma(\boldsymbol{p}, \boldsymbol{x}) = diag(\underbrace{\sigma_\epsilon^2(\boldsymbol{x})}_{N\times 1}) + \underbrace{\beta(\boldsymbol{x})}_{N\times M}\underbrace{\Sigma_F}_{M\times M}\beta(\boldsymbol{x})',$$

concludes the statement. ■

The proof also illustrates immediately that logit can never be the approximation of a stable multi-factor model that cannot be reduced to a single factor. In such a model, the substitution matrix is of rank equal to the number of factors. Intuitively investors substitute along portfolios corresponding to the various risk factors, differently for assets with different loading on those factors. More broadly, HCO demands include models where, following a price increase for a given position, the investor would substitute disproportionately with assets with similar observables.

A two-asset example. We illustrate that this limitation arises even in the simplest possible 2×2 example with the same variance. Fix the vectors $\boldsymbol{p}$ and $d(p_i; \boldsymbol{p}) = \omega_i$ that we are looking for first-order equivalence around. Such position could come from a factor model with covariance matrix for any correlation ρ and variance σ^2:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

If we have a valid approximating logit model, it would feature:

$$-\left(\partial D^{logit}/\partial p\right)^{-1} = \alpha^{-1}\begin{pmatrix} \omega_1^{-1} + \frac{1}{1-\omega_1-\omega_2} & \frac{1}{1-\omega_1-\omega_2} \\ \frac{1}{1-\omega_1-\omega_2} & \omega_2^{-1} + \frac{1}{1-\omega_1-\omega_2} \end{pmatrix}$$

Clearly the two matrix can never be identical if $\omega_1 \neq \omega_2$ because the diagonal terms must be equal. Even if we assume that our point of approximation has a given value $\omega_1 = \omega_2 = \bar{\omega}$, the models are identical only if it matches both on-diagonal and off-diagonal elements:

$$\begin{aligned}\sigma^2 &= \alpha^{-1}\left(\bar{\omega}^{-1} + \frac{1}{1-2\bar{\omega}}\right)\\ \sigma^2\rho &= \alpha^{-1}\frac{1}{1-2\bar{\omega}}\end{aligned}$$

We can already see the issue: there is only one degree of freedom in logit (α) but two degrees of freedom for the covariance matrix (σ^2 and ρ). Let us construct the corresponding contradiction. Subtracting the second equation from the first one gives:

$$\sigma^2\left(1-\rho\right)\bar{\omega} = \alpha^{-1}$$

Plugging this expression for α^{-1} in the second equation leads to:

$$\begin{aligned}\sigma^2\rho &= \sigma^2\left(1-\rho\right)\frac{\bar{\omega}}{1-2\bar{\omega}}\\ \frac{\rho}{1-\rho} &= \frac{\bar{\omega}}{1-2\bar{\omega}}\end{aligned}$$

The right-hand-side is fixed, this is our point of approximation. The left-hand-side could take any value as ρ is a free parameter.

E Appendix Tables and Figures

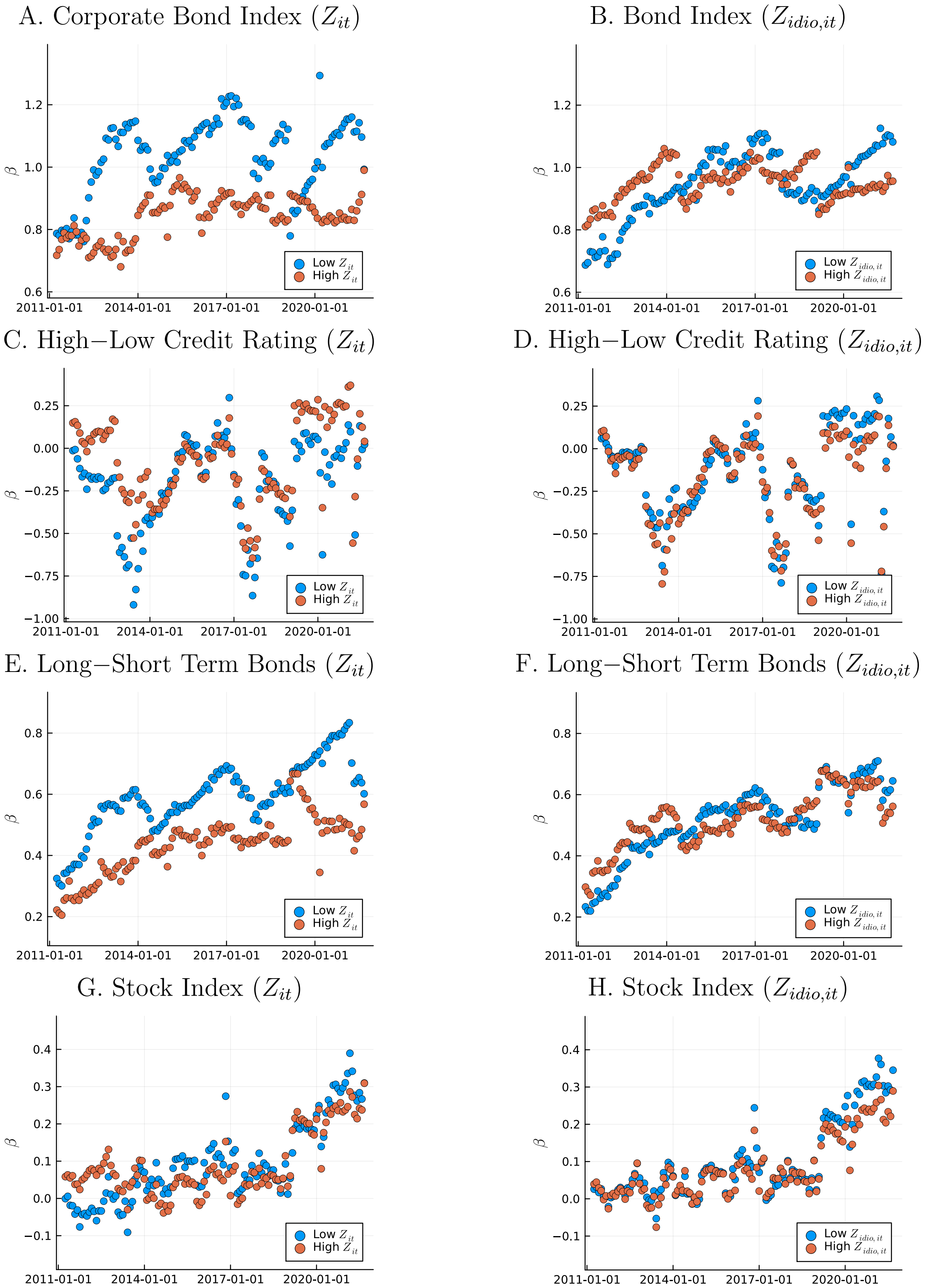
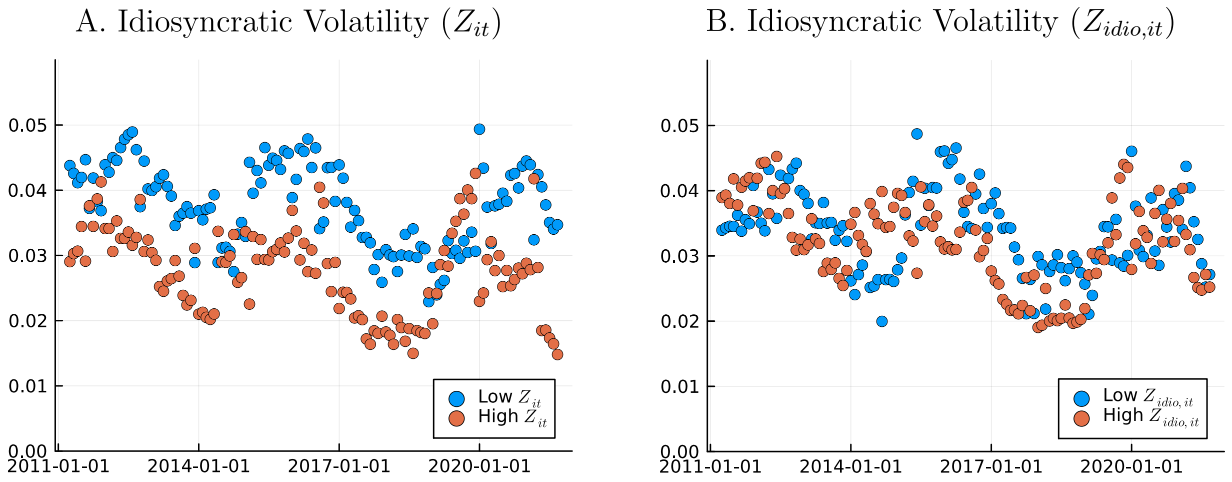


Figure 6: **Balance on covariances: exposure of portfolios sorted on demand shocks to various factors.** Figure 6 follows the exact definitions from Figure 4, but instead of showing the exposure of long-short portfolios to various factors, it shows the exposure for the long (orange) and short (blue) legs separately, sorted based on Z_{it} in the left panels and $Z_{idio,it}$ in the right panels.



A. Idiosyncratic Volatility (Z_{it})

B. Idiosyncratic Volatility ($Z_{idio,it}$)

Figure 7: **Balance on variances: average idiosyncratic volatility sorted on demand shocks.** Figure 7 reports average idiosyncratic volatilities per group sorted on both the raw demand shock Z_{it} (blue) and the demand shock $Z_{idio,it}$ (orange) that is cross-sectionally orthogonalized to duration and S&P credit ratings at each point in time. At each date, we compute idiosyncratic volatilities for each corporate bond over a two-year window centered around t, excluding t, with respect to four factors: the ICE BofA US Corporate Index Total Return, the difference between the ICE BofA US High Yield Index Total Return and the ICE BofA US Corporate Index Total Return, the difference between the ICE BofA 15+ Year US Corporate Index Total Return and the ICE BofA 1-3 Year US Corporate Index Total Return, and the Fama and French (1993) excess stock market return. We present the equal-weighted average idiosyncratic volatility among bonds with above or below median demand shock Z_{it} (or $Z_{idio,it}$). The data for factors is from FRED and the Kenneth French data library. We exclude the bottom-quintile smallest bonds based on outstanding bond supply. The time series is from 2011:04 to 2021:09.

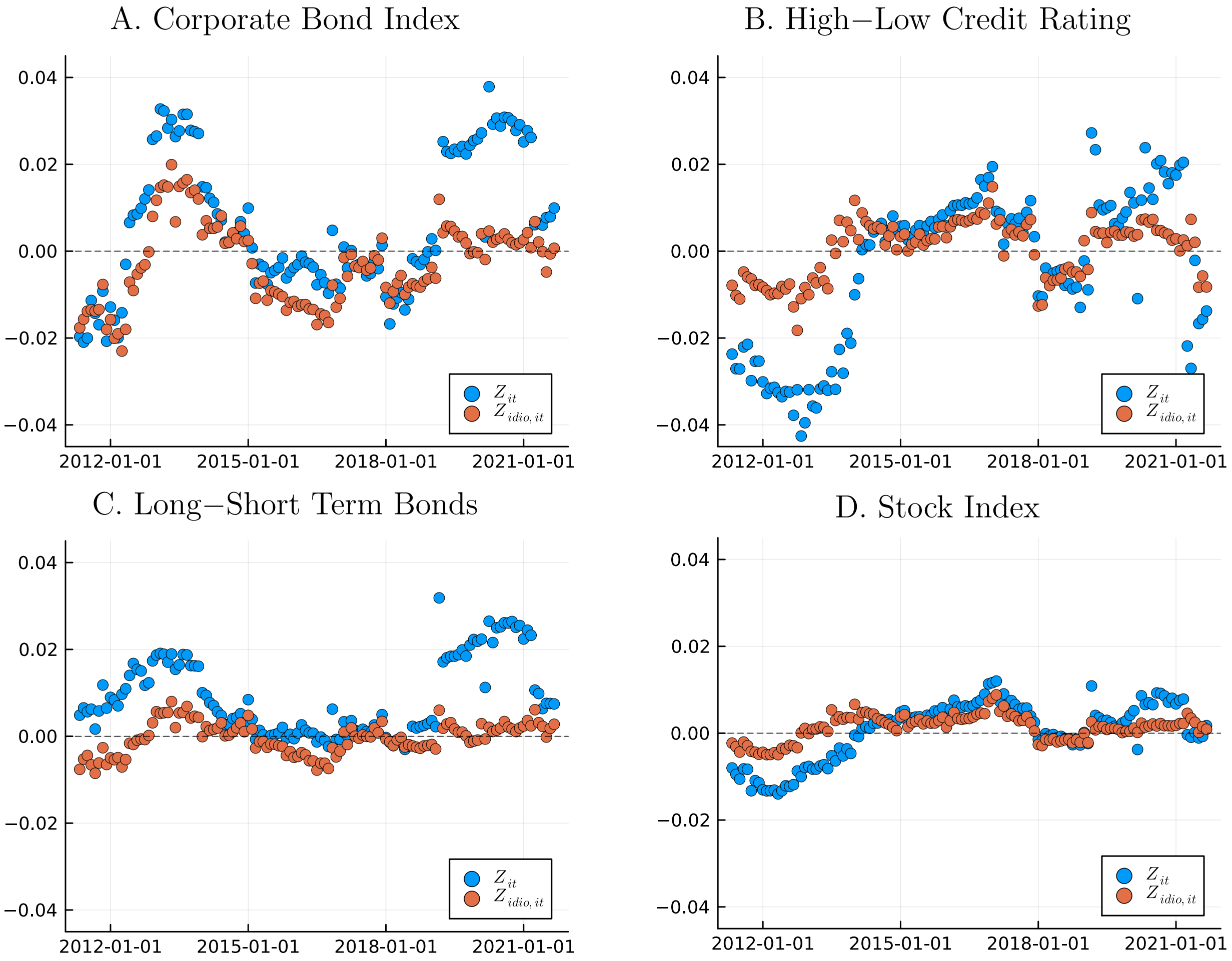


Figure 8: **Balance on covariances: exposure of long-short portfolios sorted on demand shocks to various factors.** Figure 8 reports regression coefficients from balance-on-covariance regressions based on both the raw demand shock Z_{it} (blue) and the demand shock $Z_{idio,it}$ (orange) that is cross-sectionally orthogonalized to duration and S&P credit ratings at each point in time. At each date, we form long–short equal-weighted portfolios based on whether Z_{it} (or $Z_{idio,it}$) is above or below the median. We compute the yield changes of these portfolios over two years centered around t, excluding t, and regress these yield changes on four aggregate factors. Panel A shows the time-series of coefficients for regressions on an aggregate investment-grade corporate bond factor, the ICE BofA US Corporate Index Total Return. Panel B uses the difference between aggregate high-yield and investment-grade corporate bond factors, the ICE BofA US High Yield Index Total Return and the ICE BofA US Corporate Index Total Return. Panel C uses the difference between the ICE BofA 15+ Year US Corporate Index Total Return and the ICE BofA 1-3 Year US Corporate Index Total Return. Panel D uses the Fama and French (1993) excess stock market return. The data for factors in panels A to C is from FRED, while the data for the excess market return in Panel D is from the Kenneth French data library. We exclude the bottom-quintile smallest bonds based on outstanding bond supply. The time series is from 2011:04 to 2021:09.

Table 4: **Relative multiplier $\widehat{\mathcal{M}}$ in corporate bonds**

	Yield change $\mathit{\Delta}Y_{it}$				
	(1)	(2)	(3)	(4)	(5)
Demand shock:					
Z_{it}	-0.384*	-0.104*	-0.072**		
	(0.166)	(0.047)	(0.027)		
$Z_{idio,it}$				-0.072**	-0.072**
				(0.027)	(0.027)
Date Fixed Effects		Yes	Yes	Yes	Yes
Duration × Date Fixed Effects			Yes	Yes	
Credit Rating × Date Fixed Effects			Yes	Yes	
N	630,255	630,255	630,255	630,255	630,255
R^2	0.004	0.071	0.089	0.089	0.070

Table 4 reports the results of relative multiplier regressions of yield changes $\mathit{\Delta}Y_{it}$ on demand shocks Z_{it} and $Z_{idio,it}$ for U.S. investment-grade corporate bonds. Specifications (1)–(3) use the flow-induced trading demand shock Z_{it} defined in Equation (58). Specification (1) includes a common intercept, specification (2) uses date fixed effects, and specification (3) adds controls for a continuous duration variable and S&P credit rating dummies for each date. Specifications (4)–(5) use the demand shock $Z_{idio,it}$ orthogonalized to duration and credit rating each period, with and without controlling for duration and credit rating in the regression. We exclude the bottom-quintile smallest bonds based on outstanding bond supply. The sample period is 2010:04 to 2022:09. Standard errors are clustered by date and bond.

Table 5: **Macro- and meso multipliers in corporate bonds**

	Yield Change $\mathit{\Delta} Y_{agg,t}$		Yield Change $\mathit{\Delta} Y_{X,t}$	Yield Change $\mathit{\Delta} Y_{it}$	
	(1)	(2)	(3)	(4)	(5)
$Z_{agg,t}$	-2.343***	-1.828**	0.137	-1.828**	-1.828**
	(0.627)	(0.621)	(0.105)	(0.616)	(0.616)
$Z_{X,t}$		1.683	-0.966**	1.683	1.683
		(1.295)	(0.364)	(1.286)	(1.286)
$Z_{agg,t} \times X_{it}$					0.137
					(0.105)
$Z_{X,t} \times X_{it}$					-0.966**
					(0.362)
$Z_{idio,it}$				-0.069*	-0.069*
				(0.030)	(0.030)
Duration X_{it}				-0.000***	-0.000***
				(0.000)	(0.000)
N	149	149	149	630,255	630,255
R^2	0.278	0.304	0.202	0.021	0.022

Table 3 reports the results of macro- and meso multiplier regressions of yield changes on demand shocks for U.S. investment-grade corporate bonds. Specification (1) follows equation (63) in estimating the macro multiplier by regressing aggregate yield changes $\mathit{\Delta} Y_{agg,t}$ on the aggregated instrument $Z_{agg,t}$ in the time series. Specification (2) jointly estimates the macro multiplier $\overline{M}_{agg}$ and a cross-multiplier $\overline{M}_X$ from equation (70) by adding the aggregated duration-tilted shock $Z_{X,t}$. Conversely, specification (3) jointly estimates the meso multiplier $\widetilde{M}_X$ and cross-multiplier $\widetilde{M}_{agg}$ from equation (69). Specifications (4) and (5) estimate the mechanically identical macro- and meso-level multipliers as in specifications (2) and (3) using disaggregated, repeated cross-sectional regressions, while adding the relative multiplier $\widehat{\mathcal{M}}$. We exclude the bottom-quintile smallest bonds based on outstanding bond supply. The sample period is 2010:04 to 2022:09. Robust standard errors are used for specifications (1) to (3). For specifications (4) and (5), standard errors are clustered by date and bond, and regressions are weighted such that each date receives equal weight.

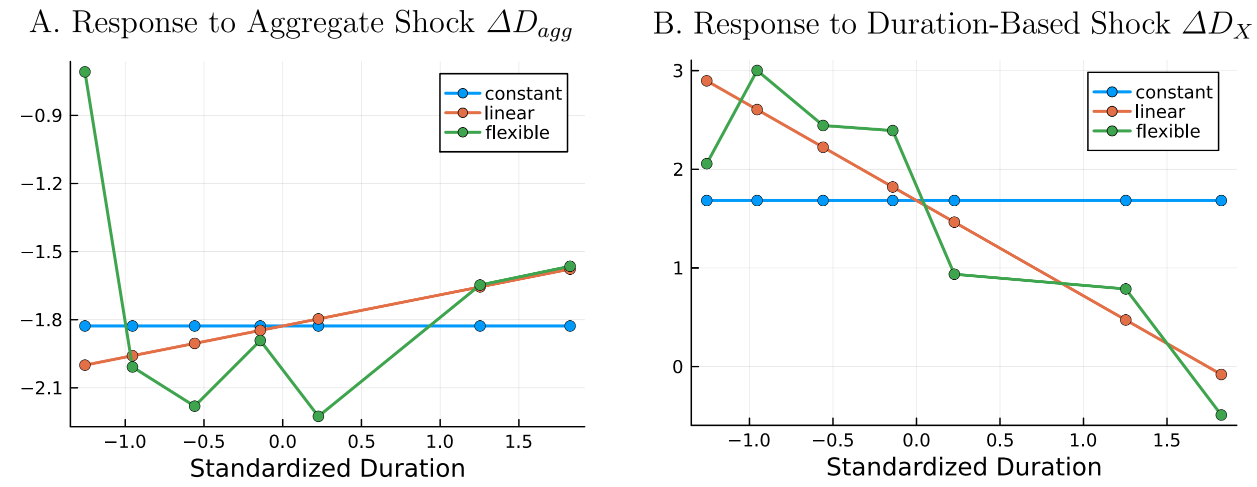


Figure 9: **Macro- and meso multipliers across durations.** Figure 9 reports the response of portfolios of corporate bonds to aggregate demand shocks ΔD_{agg} (Panel A) and shocks along duration ΔD_X (Panel B). Bonds are grouped in seven buckets based on duration: <1 year, 1–3 years, 3–5 years, 5–7 years, 7–10 years, 10–15 years, and 15+ years. The blue lines correspond to the estimates from column (4) of Table 5, which assume identical responses. The red lines are based on column (5), which includes linear interaction terms with duration X_{it}. The green line estimates these multipliers separately each duration-based portfolio in a pooled panel regression. The sample period is 2010:04 to 2022:09.