Title
Estimation and Welfare in Large-Scale Demand Systems

Author
Foley, Conor Patrick

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Estimation and Welfare in Large-Scale Demand Systems

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

by

Conor Patrick Foley

2022

ABSTRACT OF THE DISSERTATION

Estimation and Welfare in Large-Scale Demand Systems

by

Conor Patrick Foley
Doctor of Philosophy in Economics
University of California, Los Angeles, 2022
Professor Ariel T. Burstein, Co-Chair
Professor David R. Baqaee, Co-Chair

In this thesis, I study large-scale demand systems with a focus on characterizing sufficient statistics (welfare-relevant average elasticities) and how allowing for large unstructured heterogeneity affects both welfare and estimation.

In Chapter 1, I make three main contributions. First, I introduce a highly flexible class of demand systems that generalizes the most popular specification in trade and macroeconomics settings. Within this class, I show that we can characterize welfare using a (potentially time- and sample-varying) average elasticity together with an auxiliary aggregate that can be calculated using readily observable data. Second, I introduce a flexible parametric demand system (GSA translog) and adapt recently developed causal machine learning techniques to estimate the key parameters of the demand system. This estimation strategy allows for product-specific price sensitivity parameters without imposing strong ex-ante restrictions on cross-sectional patterns on which products are more or less elastically demanded. Third, I implement my new method to revisit the entry/exit adjustment problem which has been widely studied using a constant elasticity of substitution framework. My new model uncovers a novel interaction between product life-cycle patterns and welfare calculations; products that exit are systematically more

elastically demanded (on exit) relative to entering goods (on entry). This result is driven in part by the recently documented pattern that (at the barcode level) products are systematically more popular on entry than they are on exit.

In Chapter 2, I revisit the identifying assumptions for the popular heteroskedasticity-based identification strategy. While there has been significant attention paid to the statistical assumption of uncorrelated error terms, I turn the focus to the structural assumption of a single common elasticity parameter across groups. I show using Monte Carlo simulations that even minor violations of the common elasticity assumption can lead to extreme divergence between the underlying distribution of price sensitivities and the point estimates yielded by heteroskedasticity-based regression methods. Notably, unlike with linear methods (OLS and IV regression), with the heteroskedasticity-based method when the statistical assumptions hold but there is underlying variation in the product-specific price sensitivities the point estimate is not a weighted average of the underlying parameter values. To test the empirical relevance of this finding, using US trade data I compare heteroskedasticity-based point estimates to set-identification ranges for product-specific elasticities which rely on the same statistical assumptions but do not impose the problematic cross-sectional restriction. I find that in all product categories, the pooled point estimate is outside the set-identified range for some of the included products. The empirical pattern that I find is the point estimate is systematically higher (more elastic) than the set-identified ranges for each product, which provides an explanation for the pattern in the empirical literature whereby heteroskedasticity-based estimates are systematically higher than other estimation techniques.

The dissertation of Conor Patrick Foley is approved.

Jonathan E. Vogel

John W. Asker

David R. Baqaee, Committee Co-Chair

Ariel T. Burstein, Committee Co-Chair

University of California, Los Angeles

2022

For Jeeny, Evelyn, and Neru

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First I must thank my advisors David Baqaee and Ariel Burstein. Without their extraordinary support, this project could not have come to a conclusion. I have drawn consistent inspiration from their guidance and insight. I am sincerely grateful for all their support.

I have also benefited from conversations with faculty at UCLA. Most prominently, my committee members John Asker and Jonathan Vogel have provided support and useful comments. I also benefited tremendously from discussions with Oleg Itskhoki and Hugo Hopenhayn and well as comments from participants in the macroeconomics and industrial organization seminars. The staff of the UCLA economics department have provided consistent support, in particular I thank Chiara Paz for all she does to help keep life for graduate students running. Outside of UCLA, I have also benefited from discussions with Robert Vigfusson and Colin Hottman.

An essential part of the PhD experience is the interactions with classmates past and present. I am particularly grateful to Yongki Hong, Flavien Moreau, Santiago Justel, Ryan Martin, Lucas Zhang, Vitaly Titov, Ksenia Shakhgildyan, Liqiang Shi, Sumit Shinde, and Vladimir Pechu for their friendship and feedback throughout the PhD.

Most importantly, I am eternally grateful for the support of my family - my wife Jeeny, daughter Evelyn, and our beloved dog Neru. They have endured the travails of living with a fledgling researcher with grace and have consistently been my rock. Without these amazing ladies I would not be here today, and I cherish all we do together.

VITA

2016–2017 Staff Economist, Council of Economic Advisers, Washington, D.C.

2014–2016 M.A. Economics, UCLA, Los Angeles, CA

2012–2016 Research Assistant, Federal Reserve Board of Governors, Washington, D.C.

2010–2012 M.A. International Affairs & International Economics, Johns Hopkins SAIS, Washington, D.C.

2007–2011 B.A. International Affairs & East Asian Studies, Johns Hopkins, Baltimore, MD

CHAPTER 1

Flexible Entry/Exit Adjustment for Price Indices

1.1 Introduction

This paper proposes a new method for constructing an entry/exit adjusted price index. I study a restricted translog demand system that remains tractable when there are many goods and products are allowed to enter and exit. This restricted translog places no constraints on own-price elasticities, while imposing mild restrictions on cross-price effects to gain tractability. In turn, this restricted translog yields a sufficient statistic for the entry/exit adjusted price index that only requires calibrating one parameter per product. To estimate the relevant product-specific demand parameter without imposing ex-ante restrictions on the pattern of price effects in the data, I adapt the generalized random forest method of Athey et al. (2019) to my panel data setting. As an application, I apply this method to the ready-to-eat cereal market in Nielsen Consumer Panel data, finding a novel asymmetry between entering and exiting goods that is not admissible with standard CES-based techniques.

Entry and exit is a pervasive feature of the economy with important implications both for measurement and for a variety of equilibrium behaviors. The appearance and disappearance of goods is a prominent feature in highly disaggregated data sets such as retail scanner data, international trade, and detailed administrative data. Entry and exit is also a key mechanism in monopolistic competition models studied in industrial organization, international trade, endogenous growth, and business cycle theory.

Entry and exit is a well-known challenge for price index measurement. Standard price

index formulas take price changes as inputs, but price changes are not observed when goods enter and exit the market. In principle, using observed expenditure changes and the elasticities of the demand system we can impute the missing price changes for entering and exiting goods. These imputed prices correspond to the level of prices that would lead an optimizing consumer to choose to purchase exactly zero units of an unavailable good - also known as the reservation or choke price level. All else equal, entry and exit effects are larger for goods with a higher expenditure share and relatively inelastic demand.

Evaluating the effects of entry and exit requires taking a stand on the consumer demand system. The dominant approaches in the empirical literature rely on restrictive functional forms such as CES and logit. Flexible functional forms with finite choke prices, such as AIDS and homothetic translog, face two major challenges. First, the number of cross-price elasticities grows quickly as more products are added to the demand system leading to a curse of dimensionality. Second, own- and cross-price elasticities may endogenously change due to product availability, leading to a parameter stability problem. Intuitively, this parameter stability problem reflects the fact that the reaction to price changes depends, in part, on what alternatives are available to the consumer.

This paper introduces the group secondary aggregate (GSA) translog demand system, which is highly flexible while overcoming the dimensionality and parameter stability problems of an unrestricted translog function. GSA translog features a modified own price semi-elasticity that is invariant to product entry and exit. There are no restrictions on the distribution of this semi-elasticity parameter, allowing GSA translog to match a wide variety of potential patterns of own-price sensitivity. GSA translog gains tractability by imposing restrictions on cross-price effects, but still allows for a high degree of flexibility including allowing goods to be substitutes, complements, or neutral both within and between groups.

GSA translog yields a sufficient statistic for the price index that tractably incorporates entry and exit effects. This sufficient statistic generalizes the popular Tornqvist index

number formula and only requires calibrating one parameter per product - the stable modified semi-elasticity. The GSA translog price index also allows for a convenient aggregate elasticity term that may be directly compared to the CES benchmark.

Relative to CES-based methods of entry/exit adjustment, GSA translog allows for new effects that influence the net gains from entry and exit. CES imposes that all goods have a single common elasticity value, while GSA translog imposes no restrictions on the cross-section of own price elasticities allowing changes in the composition of available products to affect entry/exit adjustment at the lowest levels of aggregation. CES also imposes that the elasticity is constant for all goods. A constant elasticity implies that the reservation price is unbounded, yielding relatively large entry and exit effects. GSA translog, with finite reservation prices, yields about half the entry and exit effects for a given level of expenditure and elasticity. In addition, a constant elasticity assumption precludes life cycle dynamics for products where the elasticity on entry and on exit may differ. With GSA translog, on the other hand, the elasticity for each product varies endogenously as a product's price, and popularity, adjust over time.

To estimate the relevant product-specific demand parameters for the GSA translog demand system without imposing ex-ante restrictions on the pattern of own-price effects present in the data, I combine a cross-market price instrument common in the industrial organization literature (Hausman, 1996) with the generalized random forest (GRF) methodology of Athey et al. (2019). GRF generates a locally-weighted IV regression function, where "local" is defined in a space of user-specified product features and the weights are adaptively generated from the data using a regression forest algorithm. Athey et al. (2019) establish conditions under which GRF is consistent for causal effects conditional on product features. These consistency conditions are relatively non-restrictive; the main additional assumption is that products with similar features should have similar demand parameters. Advantages of GSA translog in this context are the stability of the parameter to be estimated, the flexibility to accommodate product-specific effects produced by the adaptive weighting procedure, and the computational benefits of a linear functional

form.

As an empirical application, I calculate an entry/exit adjusted price index using the GSA functional form and a random forest calibration for the ready-to-eat cereal market in Nielsen National Consumer Panel data.[1] This data set provides an ideal environment for this methodology because product entry and exit is a prominent feature of the data at the barcode level present in the Nielsen data. In addition, the lowest level of product grouping specified by Nielsen can still include a large number of individual products. For the cereal market in particular, there are over 4000 unique products available over the course of my sample period, providing scope for the GRF estimation strategy to effectively uncover useful variation in the data and demonstrate the effects of incorporating heterogeneity into the estimated gains from entry and exit.

Between 2004 and 2016, a standard continuing goods index using the Nielsen cereal data, which closely tracks the official CPI for cereal, yields an annual average inflation rate of 0.9%. Applying the GSA translog and CES entry/exit adjustments lowers annual average inflation by -2.8 percentage points and -3.3 percentage points, respectively, so that the net entry/exit adjustment with GSA translog is about 2/3 of the CES value. This reflects three offsetting effects. First, as noted earlier, translog demand curves yield half the net entry/exit effects implied by CES for a given level of expenditure and elasticity. Second, the GSA / random forest calibration finds more elastic demand (smaller entry/exit effects) than in the comparable CES calibration. A third offsetting factor, however, is that entering goods are less elastic (on entry) than exiting goods (on exit) over the sample period, so that the losses from exit fall more than the gains from entry and the net effect moves up from the one-half benchmark.

Related Literature This paper connects to a broad literature on entry/exit adjustment in price index measurement. This literature has been dominated by the CES-based entry/exit adjustment methodology first proposed in Feenstra (1994). Relative to the

[1]Formerly, this dataset was also known as Nielsen Homescan Consumer Panel.

CES-based method, the GSA translog studied in this paper embeds more realistic finite choke prices and allows for arbitrary patterns of price sensitivity; as my main results show, asymmetries appear to be present in the data and have substantive implications for entry/exit adjustment. Notable applications of the CES methodology for entry/exit adjustment in retail scanner data include Broda and Weinstein (2010), Braun and Lein (2021), Argente and Lee (2020), Jaravel (2019), Atkin et al. (2018), Hottman et al. (2016), Handbury and Weinstein (2015). Empirical applications in international trade include Broda and Weinstein (2006) and Hsieh et al. (2020). Applications of the CES entry/exit adjustment to productivity measurement include Klenow and Li (2020) and Aghion et al. (2019).

Entry and exit also plays a prominent role in models of monopolistic competition. While this paper does not investigate the firm side of the market (i.e. prices and product availability are taken as given), I provide a tractable framework for calculating welfare effects and estimating a demand system without a strong symmetry assumption. The CES assumption plays a prominent role in this diverse literature: providing consumer love of variety in Dixit and Stiglitz (1977), driving endogenous growth with new varieties in Romer (1990), and defining the gains from trade in Arkolakis et al. (2012). Moves away from the CES assumption typically impose symmetry on the demand system, as in Arkolakis et al. (2019), Feenstra (2018), Zhelobodko et al. (2012). While convenient, symmetry imposes strong cross-sectional restrictions on the distribution of elasticities.[2]

The GSA translog proposed in this paper generalizes the separable translog function (Matsuyama and Ushchev (2017); Kee et al. (2008)).[3] Relative to these papers, I provide a new parameterization that is central to allowing for tractable entry/exit adjustment,

[2]See the discussion in section 1.4.1 (partitcularly footnote 24) and 1.5.2.

[3]As noted in Kee et al. (2008), separable translog is also an application of the semiflexible functional form restrictions proposed in Diewart and Wales (1988) to the translog case. Applications of the Diewart and Wales (1988) semiflexible functional forms have typically focused on settings with a small number of products, as in Diewert and Feenstra (2021) and Neary (2004). Fally (2020) discusses a broad family of demand functions, homothetic and non-homothetic, that also use aggregators to control for cross-price effects.

derive an entry/exit adjusted price index, and provide a flexible estimation procedure when the set of available varieties changes over time. In addition, the grouping structure I introduce allows for greater flexibility in cross-price effects. More distantly related, Feenstra and Shiells (1996) provide a formula for welfare gains with a translog demand when only a single product is entering or exiting.

A more restrictive form of separable translog, the symmetric translog proposed in Feenstra (2003), has been used to study entry and exit in the international trade literature. Like CES, the symmetric translog supposes that all products have a single common price sensitivity parameter. In the symmetric translog case, however, an implication of this single-parameter restriction is that high-expenditure share products will have (relatively) low demand elasticity. My entry/exit adjustment generalizes the symmetric translog case, while my random forest estimation yields results substantially at odds with the single-parameter restriction. Applications using symmetric translog include Feenstra and Weinstein (2017), Novy (2013), and Bergin and Feenstra (2009). Feenstra (2010) and Fajgelbaum and Khandelwal (2016) apply the same symmetry assumption for price effects in an almost ideal demand system (AIDS, Deaton and Muellbauer (1980)), a non-homothetic extension of the homothetic translog.

This paper imposes mild restrictions on cross-price effects to gain tractability, but moves entry and exit closer to fully flexible functional forms. Diewert and Feenstra (2021) studies a quadratic mean of order two (QMOR-2) demand system; relative to that paper, I introduce a flexible and scalable estimation strategy that explicitly addresses endogeneity concerns. To address the problem of missing prices in estimation, Diewert and Feenstra (2021) focuses on the direct utility function rather than the expenditure function-based approach taken here. Hausman (1996) uses an AIDS demand system; to overcome dimensionality and parameter stability concerns that paper uses nests with only a few products and limits estimation to a stable sample of goods.

There is also a broad literature that studies entry and exit using mixed logit demand systems. Like CES, logit demand systems imply unbounded reservation prices. While a

single logit demand system imposes strong restrictions on the cross-section of elasticities, the standard practice is to use nesting together with heterogenous consumers to relax these restrictions for the market demand curves. Relative to this approach, I maintain a single representative consumer and instead rely on product features to drive differences in the estimated parameters of the demand system. Most closely related to my paper are the studies of the cereal market in Nevo (2001). Nevo (2003) compares welfare evaluation using logit-based discrete choice models to cost of living measures prepared by the Bureau of Labor Statistics based on continuing goods indices.

Utilizing product characteristics, hedonic adjustment and matched model methods as discussed in Pakes (2003) provide complements to the demand system based methods of entry/exit adjustment. These methods often imply a high degree of substitutability among the goods being compared. Ueda et al. (2019) combines a matched model approach with a CES price index. Crawford and Neary (forthcoming) uses a CES entry/exit adjustment to allow for changes in the set of characteristics utilized in a hedonic index. Along the same lines as these papers, the translog demand system studied in this paper may be used along with matched model and hedonic index methods. In addition, in principle my random forest calibration allows characteristics to indirectly enter a product-based price index through the features used in calibrating the elasticity.

Finally, this paper is related to applications of machine learning in economics. Surveys of machine learning methods are given in Athey and Imbens (2019) and Varian (2014). In the price index context in particular, Konny et al. (2019) and Groshen et al. (2017) discuss uses of big data and machine learning techniques by U.S. statistical agencies. Notable applications of machine learning techniques for consumer welfare include Chernozhukov et al. (2019) and Hausman and Newey (2017), though these papers take a more non-parametric view. Machine learning methods have also been applied to hedonic price indices as in Bajari et al. (2021) and Ehrlich et al. (2021).

Roadmap The rest of the paper proceeds as follows. Section 1.2 provides a geometric intuition for entry and exit and relates the effects of entry and exit to the Marshallian

consumer surplus. Section 1.3 gives a formal definition of the price index when products enter and exit and reviews some of the difficulties of working with demand systems when the set of consumed products changes over time. Section 1.4 introduces the homothetic with a group secondary aggregate (GSA) class of demand systems, derives a tractable entry/exit adjusted price index, and reviews the special case of CES. Section 1.5 describes the GSA translog demand system, in particular characterizing the key own-price demand parameter and deriving an associated exact entry/exit adjusted price index. Section 1.6 discusses the generalized random forest estimation strategy. Section 1.7 briefly describes the Nielsen data and some of the important features of the cereal market data. Section 1.8 presents results from the empirical application of the GSA translog calibration in the Nielsen Consumer Panel for the cereal market. Section 1.9 concludes. All proofs and some additional details are included in an appendix, available upon request.

1.2 Consumer Surplus and Entry/Exit

In this section, I review the analysis of entry and exit and give a geometric intuition for these effects as the area under the consumer's demand curve otherwise known as the consumer surplus. In addition, I show how the consumer surplus for entry and exit is related to the elasticity of the demand curve at the observed level of expenditure.

Consider the simplest case of the price index term defined in Proposition (2) below, where only a single product exits. In this case, the price index corresponds to the following expression:

$$\Delta \ln P = \int_{p_{it_0}}^{p^*_{it_1}} s_i(p_i) d \ln p_i = \left(\frac{1}{Y}\right) \underbrace{\int_{p_{it_0}}^{p^*_{it_1}} q_i(p_i) dp_i}_{-\Delta \text{ Consumer Surplus}} \tag{1.1}$$

where $s_i(p_i)$ is the consumer's expenditure share function, $q_i(p_i)$ is the consumer's quantity demanded function, Y is nominal income, and p_i is the price of good i. The

observed price when product i is available for purchase is p_{it_0}. For now, assume that $p^*_{it_1}$ is a price set by the firm that happens to exactly induce the consumer to choose to stop purchasing the product. While the firm may have skillfully chosen this price, this poses a problem for an external observer because we only observe the price level for products when an actual transaction occurs. This unknown price, that exactly sets demand equal to zero, is known as the reservation price or choke price level for product i. While it is unobserved, in principle if we knew the consumer's preferences we could simply trace out the demand curve to find this particular price level.

An alternative way of describing the price index can be constructed using a change of variables so that, instead of evaluating the integral in terms of price changes we can instead use share or quantity changes. A particularly convenient way to do this is to use the elasticity[4] of the consumer's demand curve, denoted here as ζ. In this case, we may rewrite the price index expression as:

$$\Delta \ln P = \int_{s_{it_0}}^{0} \frac{-ds_i}{\zeta_i - 1} = \left(\frac{-1}{Y}\right) \int_{q_{it_0}}^{0} \frac{p_i(q_i)}{\zeta_i} dq_i \tag{1.2}$$

where here $p_i(q_i)$ is the inverse demand function. Moving from equation (1.1) to equation (1.2) simply repackages the information in the demand curve needed to evaluate the integral. The advantage, however, is that while we may have little information about $p^*_{it_1}$ we can generally estimate the elasticity using observed data for prices and spending.

For some simple demand curves, solving the integral from equation (1.1) is a straightforward exercise. These solutions can also be written in a compact form that mirrors the elasticity-based expression from (1.2), as shown in the Table 1.1.

Figure 1.1 plots the four demand curves from Table 1.1. To be consistent with the expressions above, the four demand curves pass through a common point and have a common elasticity at the observed level of prices and quantities. Thus, all of the

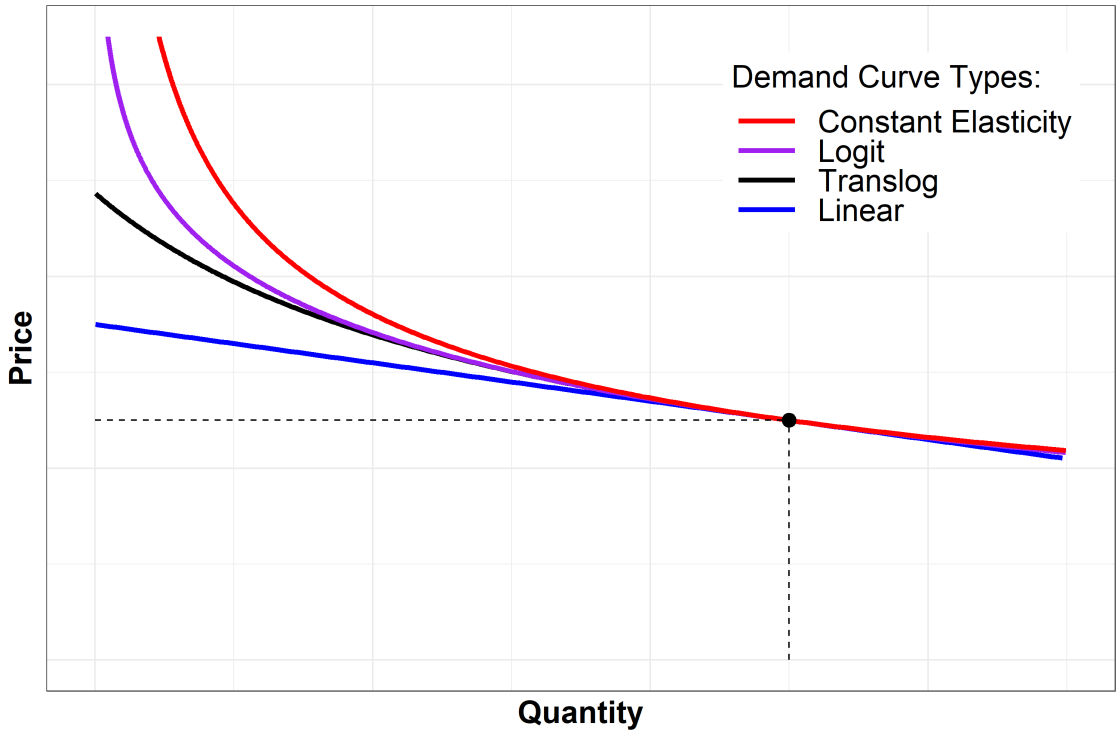
[4] i.e. $\zeta_i = -\partial \ln q_i / \partial \ln p_i$. Since consumers take income and prices as given, it immediately follows that $-d \ln s_i = (\zeta_i - 1) d \ln p_i$ and $(-\zeta_i) q_i dp_i = p_i dq_i$.

Table 1.1: Demand Curves and Consumer Surplus

Demand Type	Formula	Consumer Surplus	Approximation Type
Constant Elasticity	$\ln s_i = \alpha_i - \beta_i \ln p_i$	$\frac{s_i}{\eta_i - 1}$	Left-side (t_0, ds)
Logit	$\ln q_i = \alpha_i - \beta_i p_i$	$\frac{s_i}{\eta_i}$	Left-side (t_0, dq)
Translog	$s_i = \alpha_i - \beta_i \ln p_i$	$\frac{s_i}{2(\eta_i - 1)}$	Trapezoid (t_0, t_1, ds)
Linear	$q_i = \alpha_i - \beta_i p_i$	$\frac{s_i}{2\eta_i}$	Trapezoid (t_0, t_1, dq)

NOTE.—This table reviews the analytical formulas for four common single-product demand curves and shows that for each an exact value for consumer surplus can be written solely as a function of expenditure share and the elasticity at the point of consumption. In addition, these consumer surplus formulas coincide with approximations to the general welfare formula in equation (1.2), where the Trapezoid rule approximations assume finite choke prices.

Figure 1.1: Demand Curves with Common Elasticity



NOTE.—This figure shows four common demand curve specifications (constant elasticity, logit, translog, and linear) all normalized to have the same elasticity at the point of consumption. The different curvature of each demand curve yields different values for the consumer surplus. The formulas for the demand curves and their respective formulas are given in the text.

demand curves are mutually tangent at the point of consumption and are first-order approximations for each other in the neighborhood of (q_{t_0}, p_{t_0}). More elastic demand would correspond to a flatter demand curve, while more inelastic demand would correspond to a steeper demand curve.

As price rises and quantity demanded falls, however, the four demand curves separate reflecting the differing concavity of each functional form. Given the large price changes implied by product entry and exit, the higher-order effects that emerge from different concavity assumptions are able to have an appreciable impact on the implied effects of entry and exit.[5]

The first two demand curves - constant elasticity and logit - both imply relatively large losses from exit for a given level of the expenditure share and elasticity. As noted in the Table 1.1, their associated consumer surplus formulas correspond to first-order approximations of equation (1.2). For both of these demand curves, quantity only falls to zero in the limit as price rises to infinity.[6] This is reflected in Figure 1 as the demand curves asymptote towards, but never actually touch, the vertical axis where price is plotted. In turn, the full effect of product exit on the price index is given by the area bounded from above by the demand curve and from below by the horizontal dashed line at p_{t_0}, normalized by the level of income.

The translog and linear demand curves, with consumer surplus formulas corresponding to second-order (trapezoid rule) approximations for demand curves that actually hit the vertical axis, have relatively low concavity and thus the slopes rise only modestly or not at all as quantity declines. In turn, translog and linear demand curves feature

[5]Elasticity is a first-order feature of the demand function and thus a second-order feature of the expenditure function (with homotheticity, the ideal price index). Concavity of demand curves is a second-order feature of the $q(p, Y)$ function and thus a third-order feature of the expenditure function.

[6]Strictly, being able to push demand to zero requires the denominator to be positive, i.e. $\zeta_i > 1$ for constant elasticity and $\zeta_i > 0$ for logit. Although the reservation price is unbounded, the indefinite integrals still converge given these restrictions. If $\zeta_i \leq 1$ for constant elasticity demand, quantity falls too slowly as price increases and the integral fails to converge. If $\zeta_i < 0$ then the demand curve is upward sloping i.e. the product is a Giffen good.

finite reservation prices as long as the product has a downward sloping demand curve. Hausman (2003) advocates for the use of the linear demand curve consumer surplus formula, as this provides a lower bound for the consumer surplus of any demand curves that are convex to the origin.

Formalizing the price index expression and the treatment of product entry and exit is taken up in section (1.3) and a tractable generalization of demand system inversion from equation (1.2) is taken up in sections (1.4) and (1.5). However, the forces highlighted in the consumer surplus intuition continue to apply. First, for a given level of the elasticity, the entry or exit of a high-expenditure share product has a larger effect on the price index. Second, for a given level of the expenditure share, the effect of entry or exit is larger for products with relatively inelastic (low ζ) demand. Third, even conditional on the elasticity and expenditure share different assumptions about the shape of the demand curve (or types of approximations used) affect the value of the total gains from entry and exit.[7] The most common models for entry and exit adjustment - based on CES and multinomial logit demand systems - embed demand curves (and integrals) very similar to the simple example shown above. In particular, even in the many-goods case both of these demand curves feature unbounded reservation prices.

A limitation of this simple consumer surplus intuition, however, is that it focuses on a snapshot for a single product at a single point in time. Three additional forces that are introduced in the general case are the need to control for cross-price effects, accomodating goods with different elasticities, and allowing the elasticity to vary over time. These issues are taken up in more detail in the discussions in sections (1.4) and (1.5).

[7]See, for example, Behrens et al. (2017), which discusses this issue in the context of "love for variety" effects which are closely related to the general entry/exit adjustment problem.

1.3 The Entry/Exit Adjustment Problem

In this section, I provide a formal treatment of the problem of constructing a price index when products enter and exit the consumption basket. The essence of the exercise is laid out in Hicks (1940), which points out that whenever consumers face availability or rationing constraints we must construct an alternative set of prices that would have induced the observed behavior absent these quantity restrictions.[8] The material is fairly standard, with the caveat that I make clear the distinction between an expenditure function subject to non-negativity and product availability constraints, and an expenditure function defined on a restricted domain that does not directly impose these constraints. Following the Hicks (1940) intuition, I show how a set of virtual prices can be constructed so that the availability-constrained and restricted-domain expenditure functions yield identical outputs and optimal choices.[9] However, the demand system functions associated with these expenditure functions encode different information about the substitution patterns of the underlying preference.

In turn, given the actual and virtual-price demand systems I show how the standard cost of living concept together with the assumption of homotheticity provide an economic justification for a price index term defined over virtual prices. Given this object, the challenge for the economist is solving for unobserved virtual prices that rationalize the observed consumer behavior for the underlying preference. Given a formal statement of the problem, that issue is taken up in sections (1.4) and (1.5).

---

[8] To excerpt: "... the reason why we use prices as weights, when measuring social income as an index of economic welfare, is because prices give us some indication of marginal utilities, because the slope of the price-line is the same as the slope of the indifference curve through that point... [p]rices of commodities on the market only correspond to relative marginal utilities if the consumer's choice is free... [when] goods are not sold in the I situation, it is apparent from the preceding argument what p_1's ought to be introduced... they are those prices which, in the I situation, would just make the demands for these commodities (for the community as a whole) equal to zero."

[9] The "virtual price" terminology is due to Rothbarth (1941).

1.3.1 Observed and Virtual-Price Demand Systems

A representative consumer has preferences over a finite but arbitrarily large set of goods Ω with cardinality $|\Omega|$. Individual varieties are indexed as $i \in \Omega$. A quantity consumed of a specific product is denoted as q_i, and when collected into a vector these are denoted as $\boldsymbol{q} \in \mathbb{R}_+^{|\Omega|}$. The ordinal utility function $U(\boldsymbol{q}) : \mathbb{R}_+^{|\Omega|} \to \mathbb{R}$ represents the consumer's preference over all possible non-negative consumption bundles drawn from this hypothetical set of goods. Preferences are assumed to be locally non-satiated for all consumption bundles and to have strictly positive (albeit potentially vanishingly small) marginal utility for all products when quantities are zero.

Given that one cannot consume negative quantities, the consumer optimization problem includes a non-negativity constraint. In addition, we are concerned with situations in which consumers are unable to purchase some products at any price. Denote the set of goods that are available to consumers as $\Omega_a \subseteq \Omega$. Written out, the non-negativity and availability constraints correspond to

$$q_i \geq 0 \quad \forall i \in \Omega_a \tag{1.3}$$

$$q_i = 0 \quad \forall i \in \Omega \setminus \Omega_a \tag{1.4}$$

The consumer's problem is to maximize utility subject to the non-negativity constraints and availability constraints, together with a budget restriction that limits nominal spending to be less than or equal to nominal income $Y \in \mathbb{R}_{++}$, i.e.

$$\sum_{i \in \Omega} p_i q_i \leq Y \tag{1.5}$$

where the vector of prices is assumed to be strictly positive $\boldsymbol{p} \in \mathbb{R}_{++}^{|\Omega|}$. When goods are missing or otherwise are not purchased, we may not observe their price. For the purposes of theory, however, we allow goods to always have a price regardless of whether the price is directly observed.

Observed consumer behavior is assumed to reflect utility maximization, i.e. given the non-negativity, availability, and budget constraints the consumer choosees their most-preferred bundle, as reflected in a higher value for their utility function. This generates Marshallian demand functions and its associated value function, the indirect utility function.

Definition 1 (Availability-Constrained Marshallian Demand) *The availability-constrained Marshallian demand system, which is assumed to generate the observed data, is characterized by the following functions:*

$$\boldsymbol{q}(\boldsymbol{p}, Y, \Omega_a) \equiv \arg\max_{\boldsymbol{q}} U(\boldsymbol{q}) \quad \textit{subject to (1.3), (1.4), (1.5)}$$

$$V(\boldsymbol{p}, Y, \Omega_a) \equiv U\left(\boldsymbol{q}(\boldsymbol{p}, Y, \Omega_a)\right) = \max_{\boldsymbol{q}} U(\boldsymbol{q}) \quad \textit{subject to (1.3), (1.4), (1.5)}$$

The dual problem to utility maximization is cost minimization.[10] In cost minimization, the objective is to minimize the nominal spending needed to reach a target level of utility. The utility constraint is given by:

$$U(\boldsymbol{q}) \geq V \tag{1.6}$$

where V, written without arguments, is simply a real number that indexes a level of utility. The cost minimization problem yields a set of optimal choices known as the Hicksian demand functions and an associated value function referred to as the expenditure function.

Definition 2 (Availability-Constrained Hicksian Demand) *The availability-constrained Hicksian*

[10]Though I focus on the consumer interpretation of this problem, cost minimization is also relevant for producer purchases. The analysis is identical except that we interpret $\boldsymbol{q}$ as a bundle of input quantities, $U(\boldsymbol{q})$ as the production function, and V as some target level of output.

demand system is characterized by the following functions:

$$\boldsymbol{h}(\boldsymbol{p}, V, \Omega_a) \equiv \arg\min_{\boldsymbol{q}} \boldsymbol{p} \cdot \boldsymbol{q} \quad \textit{subject to (1.3), (1.4), (1.6)}$$

$$E(\boldsymbol{p}, V, \Omega_a) \equiv \boldsymbol{p} \cdot \boldsymbol{h}(\boldsymbol{p}, V, \Omega_a) = \min_{\boldsymbol{q}} \boldsymbol{p} \cdot \boldsymbol{q} \quad \textit{subject to (1.3), (1.4), (1.6)}$$

Given local non-satiation, the budget constraint and the utility constraint bind for each problem. I assume throughout that preferences and circumstances are such that a solution to these optimization problems exists[11], and that the optimal quantity choices are single-valued. It is well known that, under these circumstances, the Marshallian and Hicksian demand systems are dual to each other.

The non-negativity and product availability constraints are difficult to work with. Instead, I will consider an alternative set of value functions and choice functions that do not directly impose these constraints. Since the utility function is only defined for non-negative quantities, however, these functions are only defined for constraint sets that yield non-negative values for the optimal quantity vector, and will only be evaluated within this valid domain. I refer to these valid sets of constraints as the virtual price domain[12] and to emphasize the distinction between observed prices and those used to evaluate the domain-constrained functions I denote the relevant prices as $\boldsymbol{p}^*$. In turn, there is a hypothetical demand system characterized by these constraints.

Definition 3 (Virtual Price Demand System) *The virtual price demand system are the set of choice functions and associated value functions, defined over the virtual price domain, that solve*

[11]This places some restrictions on the combinations of preferences, income, prices, and availability that may be entertained. For example, we cannot define preferences that make a good essential for the consumer but then also make it impossible to procure that good. However, this does allow for goods which, in some circumstances the consumer cannot go without but in other circumstances they are happy to forego. For example, if water is an absolute necessity we can allow different varieties of water (Dasani, Poland Spring, tap-delivered etc.) to be available or not available as long as the consumer can access some water. In general a specific variety of water (e.g. Dasani) can be made unavailable, but if the only type of water left is Dasani then there is no price where the consumer can be disuaded from purchasing Dasani.

[12]Strictly speaking, this definition imposes limits on the overall constraint set, not just prices. In the homothetic case I focus on, the virtual price domain imposes no limitations on the utility level or the level of income that the function may be evaluated over.

the problems

$$\begin{aligned}
\boldsymbol{q}^*(\boldsymbol{p}^*, Y) &\equiv \arg\max_{\boldsymbol{q}} U(\boldsymbol{q}) \;\; \textit{subject to (1.5)} \\
\boldsymbol{h}^*(\boldsymbol{p}^*, V) &\equiv \arg\min_{\boldsymbol{q}} \boldsymbol{p}^* \cdot \boldsymbol{q} \;\; \textit{subject to (1.6)} \\
V^*(\boldsymbol{p}^*, Y) &\equiv U(\boldsymbol{q}^*(\boldsymbol{p}^*, Y)) = \max_{\boldsymbol{q}} U(\boldsymbol{q}) \;\; \textit{subject to (1.5)} \\
E^*(\boldsymbol{p}^*, Y) &\equiv \boldsymbol{p}^* \cdot \boldsymbol{h}^*(\boldsymbol{p}^*, V) = \min_{\boldsymbol{q}} \boldsymbol{p}^* \cdot \boldsymbol{q} \;\; \textit{subject to (1.6)}
\end{aligned} \tag{1.7}$$

In general, we can construct a hypothetical vector of virtual prices that aligns the choices (and value functions) from the availability-constrained and the virtual price demand systems. That is, we may choose a hypothetical set of prices that rationalizes observed behavior without relying on non-negativity or availability constraints. Formally, we have:

Definition 4 (Behavior-rationalizing virtual prices) *A vector of virtual prices $\boldsymbol{p}^*$ is said to rationalize the behavior of the consumer facing constraint $(\boldsymbol{p}, Y, \Omega_a)$ if it satisfies the relationship*

$$\boldsymbol{p}^*(\boldsymbol{p}, Y, \Omega_a) := \boldsymbol{q}^*(\boldsymbol{p}^*, Y) = \boldsymbol{q}(\boldsymbol{p}, Y, \Omega_a)$$

These virtual prices can be solved for by evaluating the Lagrangian for the availability-constrained optimization. In particular, the first order conditions for the availability-constrained and virtual price demand systems only differ for products with zero purchases given that the budget (utility) constraint binds in either case. This is stated formally in the following proposition.

Proposition 1 (Observed and Reservation Prices) *For a consumer facing constraint $(\boldsymbol{p}, Y, \Omega_a)$ who chooses to purchase products in $\Omega_o \subseteq \Omega_a$, the virtual price vector that rationalizes their behavior is given by*

$$\boldsymbol{p}^*(\boldsymbol{p}, Y, \Omega_a) = \begin{bmatrix} \boldsymbol{p}_o \\ \boldsymbol{r}^*_m(\boldsymbol{p}_o, Y, \Omega_o) \end{bmatrix} \tag{1.8}$$

where $\boldsymbol{p}_o \in \mathbb{R}^{|\Omega_o|}_{++}$ are the elements of the observed price vector $\boldsymbol{p}$ corresponding to goods with

positive purchases, $\Omega_m \equiv \Omega \setminus \Omega_o$ *is the collection of unpurchased goods, and* $\boldsymbol{r}_m^* \in \mathbb{R}_{++}^{|\Omega_m|}$ *are referred to as the reservation prices for unpurchased goods.*

Thus, the only unobserved components of the behavior-rationalizing price vector are exactly the prices for goods with zero consumption.[13]

When the set of actually purchased goods is fixed over time, this can be interpreted in one of two ways. We may suppose that this is because consumers faced a fixed constraint set so that they are simply unable to switch into the unobserved products. Alternatively, we can think of this as consumers who do not face non-negativity and availability constraints but instead always face prices which adjust to ensure they do not consume the non-purchased goods. Written out, there are two functions which can both rationalize the observed behavior:

$$\boldsymbol{q}(\boldsymbol{p}, Y, \Omega_a) = \begin{bmatrix} \boldsymbol{q}_o(\boldsymbol{p}_o, Y, \Omega_o) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{q}_o^* \left(\begin{bmatrix} \boldsymbol{p}_o' & \boldsymbol{r}^{*'}(\boldsymbol{p}_o, Y, \Omega_o) \end{bmatrix}', Y \right) \\ \mathbf{0} \end{bmatrix} \tag{1.9}$$

There are two important distinctions between the functions $\boldsymbol{q}_o(\boldsymbol{p}_o, Y, \Omega_o)$ and $\boldsymbol{q}_o^*(\boldsymbol{p}^*, Y)$. First, only the prices of observed goods are needed to evaluate $\boldsymbol{q}_o(\boldsymbol{p}_o, Y, \Omega_o)$ while $\boldsymbol{q}_o^*(\boldsymbol{p}^*, Y)$ is defined over prices for all goods, including the unobserved reservation prices. Second, the price derivatives of the $\boldsymbol{q}_o(\boldsymbol{p}_o, Y, \Omega_o)$ function enforce that we cannot substitute into the unobserved products while the price derivatives of $\boldsymbol{q}^*(\boldsymbol{p}^*, Y)$ allow for substitution effects between observed and unobserved products, as long as price movements are evaluated in a direction that stays within the virtual price domain for a given Y.

[13] In the rationing context considered in Hicks (1940) and Rothbarth (1941) equality constraints are binding for non-zero quantities, making the relationship between observed and rationalizing virtual prices more complicated.

1.3.2 Cost of Living and the Availability-Constrained Price Index

In this section, I define the cost of living measure and relate it to both the availability-constrained and the virtual price expenditure function. In addition, imposing the standard assumption of homotheticity, I relate the cost of living to observed expenditure shares and the (partly unobserved) changes in virtual prices.

I index the potential constraints faced by consumers by t. That is, t refers to a set of prices ($\boldsymbol{p}_t$), income (Y_t), and product availability (Ω_t). While in principle a consumer may choose not to consume an available good at given prices and income, to economize on notation I will also use Ω_t to refer to goods with non-zero purchases at an observed point.[14] In turn, for every constraint there is a set of associated optimal choices $\boldsymbol{q}_t$, a utility index V_t, and a set of virtual prices $\boldsymbol{p}_t^*$ that rationalizes the observed choices absent the availability and non-negativity constraints. In my empirical exercise, t indexes prices and product availability in different time periods while in a purchasing power parity context t would index different regions.

The cost of living refers to the level of nominal income needed to make a consumer indifferent between two sets of price and availability constraints.[15] This is usually characterized in terms of relative levels of income. Using the value functions from consumer optimization we have the following definition:

Definition 5 (Cost of Living) *The change in the cost of living for a consumer with target utility level $\bar{V}$ is given by*

$$\Delta COL(t_1, t_0, \bar{V}) \equiv \ln E(\boldsymbol{p}_{t_1}, \bar{V}, \Omega_{t_1}) - \ln E(\boldsymbol{p}_{t_0}, \bar{V}, \Omega_{t_0}). \tag{1.10}$$

[14] In the homothetic case that I focus on this is relatively unproblematic since the choice to purchase or not purchase is entirely driven by prices with no dependence on income or the level of utility. With non-homotheticities, an available good may still be unpurchased because the consumer's income is too high or too low to make the good desirable.

[15] The standard notion of cost of living was first given in Konus (1924). The cost of living is also closely related to the proportional (log) versions of the money-metric welfare measures, compensating and equivalent variation, proposed by Hicks.

Given the definition of $\boldsymbol{p}_t^*$ the cost of living can be equivalently written using the virtual price expenditure function:

$$\Delta \text{COL}(t_1, t_0, \bar{V}) = \ln E^*(\boldsymbol{p}_{t_1}^*, \bar{V}) - \ln E^*(\boldsymbol{p}_{t_0}^*, \bar{V}).$$

When the underlying utility function is homothetic, the virtual price expenditure function takes on a particularly convenient form:

$$E^*(\boldsymbol{p}^*, V) = VP(\boldsymbol{p}^*), \tag{1.11}$$

where we refer to $P(\boldsymbol{p}^*)$ as the price index for the consumer. Thus, in the homothetic case[16] the cost of living no longer depends on the benchmark level of utility since we have

$$\Delta COL(t_1, t_0, \bar{V}) = \ln P(\boldsymbol{p}_{t_1}^*) - \ln P(\boldsymbol{p}_{t_0}^*). \tag{1.12}$$

Given that the homotheticity assumption is maintained throughout, from this point on changes in the cost of living are simply referred to as changes in the price index. As with the virtual price expenditure function, the price index $P(\boldsymbol{p}^*)$ is defined only within the virtual price domain.

An additional benefit of the homotheticity assumption is that the functions that characterize the fraction of total spending devoted to each good no longer depend on the level of utility or the level of income, i.e. with homotheticity we have:

$$s_i^*(\boldsymbol{p}^*) = \frac{p_i^* q_i^*(\boldsymbol{p}^*, Y)}{Y} = \frac{p_i^* h_i^*(\boldsymbol{p}^*, V)}{E^*(\boldsymbol{p}^*, V)} \tag{1.13}$$

Optimizing behavior implies that this expenditure share function is exactly the gradient of the log-price index function.[17] This gives rise to the virtual price analog for the

[16] When discussing the cost function for firm input purchases, imposing that the production function exhibits constant returns to scale provides the analogous results to assuming homothetic preferences in the consumer context. See Hulten (1973).

[17] Applying Shepard's lemma, we have $\partial \ln E^* / \partial \ln p_i^* = \partial \ln P / \partial \ln p_i^* = s_i^*(\boldsymbol{p}^*)$ as long as the derivative

standard price index expression.

Proposition 2 (Availability-Constrained Price Index) *The change in the cost of living for a consumer with a homothetic preference may be written as an integral*[18] *of the form:*

$$\Delta \ln P = \int_{t_0}^{t_1} \sum_i s_i^*(\boldsymbol{p}^*(t)) d \ln p_i^*(t) \tag{1.14}$$

There are a wide variety of approaches to evaluating this integral or constructing discrete approximations. A notable example is to use a trapezoid rule approximation, which yields a Tornqvist (1936) index defined over all goods and using changes in virtual prices:

$$\Delta \ln P \approx \sum_{i \in \Omega} \left[\frac{s_{it_0} + s_{it_1}}{2} \right] \Delta \ln p_i^* \tag{1.15}$$

The Tornqvist index is one of the superlative index number formulas identified in Diewert (1976) and selected as a target price index formula by official statistical agencies. In US data, the Bureau of Labor Statistics uses the Tornqvist index to construct the Chained-CPI measure and the quantity-index analog is used in preparing multi-factor productivity measures. Given the homotheticity assumption, superlative indices are recommended because they adjust for the substitution biases of Laspeyres and Paasche indices.[19]

---

is evaluated in a direction that stays within the virtual price domain.

[18]This integral is also known as a divisa price index. In principle if we see data at a higher enough frequency, including allowing prices to rise and fall such that expenditure shares move smoothly up to or down from zero as goods enter and exit, we could evaluate this integral using observed prices. In practice, however, prices and expenditure are only observed at discrete intervals and we do not observe the behavior of the expenditure share functions as goods enter and exit. Without the assumption of stable and homothetic preferences the divisia index defined here (using observed expenditure shares) no longer corresponds to a well-defined welfare object and may exhibit pathological behavior due to path dependence. Silberberg (1972) discusses the path dependence issues in more detail, while Baqaee and Burstein (2021) discusses welfare measures that are consistent with non-homotheticity, taste changes, and endogenous prices due to production activity.

[19]Laspeyres and Paasche indices use expenditure shares from the initial (Laspeyres) or final (Paasche) periods as weights for price changes, which ignores that consumers re-optimize their consumption bundle in the face of price changes. Laspeyres and Paasche indices provide first-order approximations for changes in the cost of living (fixing the level of utility at the initial and final values, respectively) but their justification does not rely on the homotheticity assumptions imposed here.

When the set of goods that consumers purchase does not change, all that is needed to evaluate equation (1.15) is readily observable data on prices and consumer spending. However this happy occurrence is not what we encounter in practice; the problem for the economist then is to impute the missing virtual price changes when (for whatever reason) goods enter and exit the consumer's consumption basket. This can take the form of imputing the whole price change, as done in this paper, or imputing the level of the reservation price in the period where the product is missing.

Even with the homotheticity restriction, there is not enough structure on the demand system to make an inversion of the form in equation (1.2) tractable. There are two main challenges. First, even characterizing the demand system locally for a fixed set of products is infeasible as the number of cross-price effects grows in the square of the number of products, causing a curse of dimensionality. Second, as emphasized in the discussion of equation (1.9) when the set of purchased goods changes over time there are endogenous changes in substitution patterns leading to a parameter instability problem. The observed demand system for a fixed set of goods does not, in general, encode information on the substitution patterns between purchased and unpurchased goods even though this information is needed to know what the reservation price levels are.

Having given a formal statement of the object of inquiry, we can now turn to the issue of how to construct a flexible method for incorporating entry and exit into the price index measure.

1.4 Homothetic with a Group Secondary Aggregate

In this section, I introduce an additional restriction on a homothetic virtual price expenditure function - the group secondardy aggregate (GSA) form - that yields a particularly convenient demand system. This restriction avoids a curse of dimensionality by imposing relatively mild structure on cross-price effects, yielding an entry/exit adjustment formula that

only requires calibrating one parameter per product while still retaining a high degree of flexibility. Finally I compare the general entry/exit adjustment formula for GSA demand systems to the entry/exit adjustment for CES demand, which is a special case of GSA.

While keeping a general treatment, I also discuss the implications of standard demand system restrictions for the GSA demand class. In particular, I focus on how the standard conditions for how the elasticity changes with prices (Marshall's laws of demand) interact with the GSA restriction and some of their implications for entry/exit adjustments through the panel-dimension of the elasticity distribution. I also discuss the difference between the GSA notion of grouping with the restrictions imposed by combining nesting in the expenditure function with the homotheticity restriction.[20]

1.4.1 Group Secondary Aggregate (GSA) Restriction

In this section I introduce the group secondary aggregate (GSA) restriction and discuss some of its implications. In the next section, I derive an entry/exit adjusted price index for this class of demand systems. To begin, define a GSA demand system as follows.

Definition 6 (Homothetic with a Group Secondary Aggregate (GSA)) *A demand system is homothetic with a group secondary aggregate (GSA) if the expenditure function takes the form given in equation (1.11) and the expenditure share functions may be written as*

$$s_i^*(\boldsymbol{p}^*) = f_i(p_i^* / A_{g(i)}^*(\boldsymbol{p}^*)) \tag{1.16}$$

where $f_i : \mathbb{R}_{++} \to [0,1]$ *and* $A_g^* : \mathbb{R}_{++}^{|\Omega|} \to \mathbb{R}_{++}$.

I refer to $p_i^* / A_{g(i)}(\boldsymbol{p}^*)$ as a product's relative price, and to A_g as the base price for group g. With GSA demand, the collection of products Ω is partitioned into a set of mutually exclusive and exhaustive subsets Ω_g. For simplicity I use the expression $i \in g$ to indicate $i \in \Omega_g$, and when the context is clear g may also refer to the number of groups. The GSA

[20]Not all nesting includes the homotheticity restriction. For example, Atkin et al. (2020) explores some implications of nesting while allowing for certains types of non-homotheticity and cross-nest effects.

structure supposes that all expenditure share functions can be interpreted as a good's own price relative to a scalar aggregator of all prices (including, potentially, a product's own price) and that products in the same group compete against the same aggregator. The full vector of expenditure share functions must still satisfy the standard conditions for a homothetic demand system, which limits the admissible combinations of f_i and A_g functions.

A key characteristic of the GSA demand system is the elasticity of the expenditure share with respect to the relative price term, which I refer to as the partial elasticity of good i.

Definition 7 (GSA Partial Elasticity) *The partial elasticity for a product in a GSA demand system is characterized by the function*

$$\eta_i \equiv 1 - \frac{\partial \ln s_i^*}{\partial \ln \left[p_i^* / A_{g(i)}^* \right]} \tag{1.17}$$

Since a good's own price typically enters into the base price for its own group, it is generally the case that the partial elasticity is not equal to the standard Marshallian or Hicksian own-price elasticity.[21] For convenience, I suppress the dependence of η_i on the vector of virtual prices used to evaluate the derivative.

Cross-Sectional Flexibility of Elasticities Without any further restrictions, the GSA demand system allows for full flexibility for the partial elasticity functions, which will be a central focus of the entry/exit adjustment derived in section (1.4.2). By the cross-sectional dimension of elasticities, I mean the distribution of elasticities for observed products at a given level of prices and expenditure shares. Since the f_i function is specific to each product, knowing that product i is more expensive or has a higher market share than some other good j tells us nothing about whether demand for product i is more or less elastic than demand for product j.

[21] i.e. $\eta_i - 1 \neq \zeta_i - 1 = -\partial \ln s_i(\boldsymbol{p}^*) / \partial \ln p_i^*$.

Marshall's Laws, GSA, and Panel Dimension of Elasticities The panel dimension refers to changes in the elasticity of the same good at different levels of a good's relative price. There are two standard restrictions on the panel dimension of the elasticity function. First, the elasticity should be greater than one in the observed range of prices. In general, this condition ensures that increases in a product's relative price leads to a decline in its expenditure share. This is also needed for natural entry/exit effects, i.e. that exiting goods experience price increases while entering goods experience price declines. This is a slightly stronger version of Marshall's first law of demand which says that price increases should lead to declines in quantity demanded; the requirement that the full (rather than partial) elasticity is greater than one is also required for price-setting firms to have a well-behaved optimization problem when setting markups.

A second standard restriction on how the elasticity reacts to changes in prices is the requirement that an increase in price leads to an increase in the elasticity. This condition is known as Marshall's second law of demand and, although it is not strictly required by the optimizing behavior that generates the demand system, it is necessary to generate reasonable behavior by consumers and firms in a variety of contexts.[22] The CES is a notable example that does not impose the Marshall's second law condition.[23] Combining the GSA restriction and Marshall's laws for the partial elasticities implies the following: the partial elasticity when a product is more popular (higher s_i) will be greater than when the product is less popular (lower s_i).

In principle, the panel effects imposed by Marshall's laws need not have any effect on

[22]See for example the discussions in Mrazova and Neary (2017), where the Marshall's second law condition is referred to as "subconvexity", or Melitz (2018). Marshall's second law helps to explain incomplete pass-through of marginal cost changes to prices and is the condition on (symmetric) demand curves that generates love of variety, as discussed in Zhelobodko et al. (2012). The discussion here focuses on applying the Marshall's law conditions to the partial elasticity, which is not the same as imposing these conditions on the full own-price elasticity since this paper works with discrete products rather than a continuum of goods.

[23]This refers to the partial elasticity term in the CES demand system, discussed in more detail in section (1.4.3). Products in a CES demand system exhibit Marshall's laws when there are a finite number of products and the CES partial elasticity is greater than 1. There is still a cross-sectional restriction for the general CES function; more popular (higher s_i) products are relatively inelastically demanded.

the cross-section of elasticities. However, the common assumption of symmetry in the demand system creates a link between the panel and cross-sectional features of the elasticity distribution. In parameteric demand systems, symmetry usually takes the form of a single common price-sensitivity parameter that is common to all products. In the GSA context, Marshall's laws together with a symmetry condition would imply that more expensive products (and low-s_i products) are more elastically demanded than less expensive (and high-s_i) products.[24]

GSA Grouping versus Nesting The GSA notion of grouping is not the same as "nesting" in the sense of Blackorby et al. (1978). This is seen most clearly in the fact that $A_g^*(\boldsymbol{p}^*)$ can depend on all prices, not just the prices of goods in the same group. With nesting and homotheticity, the expenditure function has a two-tier structure. For the lower tier, there are within-nest expenditure functions each with an associated price index defined over products within the nest. In turn, the upper tier aggregates the within-nest price indices into a single overall price index term. The structure implies that a good's expenditure share function can be decomposed into a nest-share and within-nest component, where the within-nest component depends only on the prices of goods in the same nest.

One advantage of using the GSA notion of grouping rather than nesting to consolidate cross-price effects is given in the following proposition:

Proposition 3 (GSA and Cross-Price Effects) *When the demand system is homothetic with a group secondary aggregator (GSA) and all partial elasticity values are greater than one, then:*

---

[24]In the stronger (and more common for theoretical studies) form of symmetry all products face the same demand curve. This implies that more expensive products are always less popular (1st law and strong symmetry) and more elastically demanded (2nd law and strong symmetry). This strong form of symmetry would be imposed in the GSA context if there was a single common f_i function, i.e. if demand was given by $s_i = f(p_i/A)$. A weaker form of symmetry common in empirical applications allows freedom in the level of demand (e.g. a free intercept term) but imposes a common slope term - i.e. that demand curves are (in some sense) parallel. The exact interpretation of the slope term depends on the functional form; e.g. CES (single common elasticity), logit (elasticity is proportional to price), or symmetric translog (elasticity minus one is proportional to inverse of expenditure share). Kroft et al. (2021) discusses some implications of the parallel curves assumption for variety effects in the context of discrete choice models. See the discussion of equation (1.33) for more on the implications of symmetry for the entry/exit adjustment process.

(3a) *if there is only one group all products are substitutes, and*

(3b) *if $i \in g$ is gross substitutes (resp. complements, neutral) with respect to $j \in g'$ then all products in g are gross substitutes (complements, neutral) with respect to all products in g'.*

As noted in the preceding discussion, imposing that the partial elasticity is greater than one is a very natural restriction, particular when working with a large number of products. Thus, Proposition (3a) implies that if we have a nested expenditure function with single-group GSA demand systems in each nest then products within nests must be substitutes. On the other hand, Proposition (3b) allows for products within the same group to be substitutes, complements, or neutral. A second advantage of the GSA notion of grouping is that subsets of groups are also groups in the GSA sense so there is less ex-ante theoretical concern about adding "too many" groups. With nesting, however, breaking products into smaller nests does imply a different demand system.

The standard nesting restriction also has some attractive properties not present with the GSA notion of grouping. First, nesting achieves dimensionality reduction by limiting the amount of cross-price parameters that must be estimated. However, when a nest includes only a small number of products there is no need to impose extra restrictions on the within-nest demand system which provides a way to overcome the issues noted in the discussion of Proposition (3a). Second, nesting allows for intermediate levels of aggregation since standard price index principles may be applied to each nest individually; groups in the GSA sense cannot be treated in this way. While I do not take up the issue of aggregation for GSA demand systems in general, in the appendix I discuss the aggregation properties for the GSA translog demand system (section 1.5).

1.4.2 Entry/Exit Adjustment for GSA Demand Systems

In this section, I show how the GSA demand system yields a tractable entry/exit adjustment formula that only requires observed price and expenditure data and calibration of one

parameter per product, the partial elasticity value.

By definition the partial elasticity allows us to map between changes in expenditure shares and relative prices, so that we have

$$s_i^* d \ln p_i^* = s_i^* d \ln A_{g(i)} - \frac{ds_i^*}{\eta_i - 1} \tag{1.18}$$

which uses the fact that $s_i^* d \ln s_i^* = ds_i^*$.

Using equation (1.18), we can substitute for the unobserved price changes of entering and exiting. Defining the set of goods observed at time t_1 and t_0 as $i \in c$, we have:

$$\Delta \ln P^{GSA} = \int_{t_0}^{t_1} \underbrace{\sum_{i \in c} s_i^* d \ln p_i}_{\substack{\text{Continuing} \\ \text{Goods} \\ \text{Contribution}}} + \underbrace{\sum_g (s_g^* - s_{cg}^*) d \ln A_g^*}_{\substack{\text{Base} \\ \text{Price} \\ \text{Effects}}} \underbrace{- \sum_{i \notin c} \frac{ds_i^*}{\eta_i - 1}}_{\substack{\text{Net Entry/Exit} \\ \text{Partial} \\ \text{Consumer Surplus}}} \tag{1.19}$$

where s_g^* is the the share of expenditure on product in group g and s_{cg}^* is the share of spending on goods in group g that appear in both periods.[25]

I refer to the last term here as the partial consumer surplus by way of analogy with the standard single-product consumer surplus discussed in section (1.2). In the event that (counterfactually) the base-price term were to be held fixed, the partial consumer surplus would correspond to the integral of the expenditure share function with respect to a product's own price. However, since products enter into their own group's base price term the partial consumer surplus does not account for the full effect of entry and exit even in the simple case when only one product enters or exits. Instead, we must also include an adjustment for the gap between the partial (constant-A) effect and the full effect. Thus the full contribution of entering and exiting goods to the price index is the sum of the partial consumer surplus and the base-price effect terms.[26]

[25] i.e. $s_g^* = \sum_{i \in g} s_i^*$ and $s_{cg}^* = \sum_{i \in g \cap c} s_i^*$.

[26] This notion of "contribution" is purely an accounting concept, along the lines of the "contribution to change" values published in national income accounts. This is distinct from the welfare effect of entry

The price index expression in equation (1.19) is still not satisfactory since we do not directly observe the change in the base price term. Indeed, the unobserved virtual price changes will also enter into the change in the base price term so we have simply transferred the problem of missing prices into a new component of the price index. The fact that the base price term is assumed to be common among products in the same group, however, allows us to use the behavior of continuing goods to recover this missing component the virtual price changes. Specifically, rearranging (1.18) and summing over all continuing goods in each group gives us:

$$s_{cg}^{*} d \ln A_g = \sum_{i \in g \cap c} s_i^{*} d \ln p_i + \frac{ds_i}{\eta_i - 1} \tag{1.20}$$

In order to apply this equation, there must be at least one continuing good in each group. The base price effect captures all cross-price effects relevant to a product together with the component of own-price effects neglected in the partial elasticity term. The simple example when no observed prices change while a single good exits highlights how equation (1.20) encodes useful information about own- and cross-price effects. To the extent that the partial elasticity is different from the full own-price elasticity, this is exactly because it ignores a product's contribution to its group aggregator. When products within a group are substitutes, this will mean the partial elasticity is more elastic (smaller exit effect) than the full own-price elasticity; the more substitutable are group members the bigger the mismatch. Due to substitutability, the price increase associated with exit will lead to some extra spending being devoted to fellow group members which will show up in equation (1.20) as an increase in the base-price term and thus boost the total exit effect; the stronger the substitutability within the group the more spending on continuing group members goes up. If products within a group are complements all these effects operate in the opposite direction. Thus, without directly estimating cross-price elasticities equation (1.20) still allows for us to pick up the sign

and exit in an equilibrium counterfactual. For example, if there are strategic complementarities in price setting among monopolistically competitive firms some of the "continuing goods contribution" would reflect changes in the competitive environment due to changes in product availability.

and magnitude of cross-price effects within a group.

Plugging equation (1.20) into equation (1.19) gives the full entry/exit adjusted price index for GSA demand systems, the first main result of this paper.

Proposition 4 (GSA Entry/Exit Adjusted Price Index) *The price index for a demand system that is homothetic with a group secondary aggregate may be written as:*

$$\Delta \ln P^{GSA} = \int_{t_0}^{t_1} \underbrace{\sum_g s_g^* \sum_{i \in g \cap c} \frac{s_i^*}{s_{cg}^*} d \ln p_i}_{\substack{\text{Continuing} \\ \text{Goods} \\ \text{Index}}} + \underbrace{\sum_g \left[\frac{s_g^*}{s_{cg}^*} - 1 \right] \sum_{i \in g \cap c} \frac{ds_i}{\eta_i - 1}}_{\substack{\text{Continuing Goods} \\ \text{Partial Consumer Surplus} \\ \text{Adjustment}}} \underbrace{- \sum_{i \notin c} \frac{ds_i}{\eta_i - 1}}_{\substack{\text{Net Entry/Exit} \\ \text{Partial} \\ \text{Consumer Surplus}}} \tag{1.21}$$

As with the standard price index formula in terms of price changes, it useful to construct discrete approximations for this entry/exit adjusted price index. In particular, if we take trapezoid rule approximations of equations (1.19) and (1.20), then a discrete analog of full entry/exit adjusted price index in the price index from equation (1.21) is given by:

$$\begin{aligned} \Delta \ln P^{GSA} \approx & \sum_g \bar{s}_g \sum_{i \in g \cap c} \frac{\bar{s}_i}{\bar{s}_{cg}} \Delta \ln p_i \\ & + \sum_g \left[\frac{\bar{s}_g}{\bar{s}_{cg}} - 1 \right] \sum_{i \in g \cap c} \left[\frac{1/2}{\eta_{it_1} - 1} + \frac{1/2}{\eta_{it_0} - 1} \right] \Delta s_i \\ & - \sum_{i \notin c} \left[\frac{1/2}{\eta_{it_1} - 1} + \frac{1/2}{\eta_{it_0} - 1} \right] \Delta s_i \end{aligned} \tag{1.22}$$

where overbars indicate the arithmetic averages of the relevant values between t_0 and t_1 and Δ indicates a discrete change.[27] Given the standard condition that we observe expenditure shares and price changes for continuing goods, the only components of this expression that must be calibrated are the partial elasticity values. When there is no entry and exit, this expression simplifies back to being simply the standard Tornqvist index from equation (1.15).

[27] For example, $\Delta s_i = s_{it_1} - s_{it_0}$ and $\bar{s}_i = (1/2)(s_{it_1} + s_{it_0})$.

Consumer Surplus Aggregation When $c = \Omega_{t_0} \cap \Omega_{t_1}$ and all products have finite choke prices, we can simplify the discrete analog of the net entry/exit partial consumer surplus term. Define the set of exiting goods as $i \in x : i \in \Omega_{t_0} \setminus c$ and likewise define the set of entering goods as $i \in n : i \in \Omega_{t_1} \setminus c$. When goods have finite choke prices the partial elasticity at the end point where the good is not purchased is unbounded, so these terms drop out. In addition, by definition $\Delta s_i = -s_{it_0}$ for $i \in x$ and $\Delta s_i = s_{it_1}$ for $i \in n$. This gives us:

$$\begin{aligned} -\sum_{i \notin c} \left[\frac{1/2}{\eta_{it_1} - 1} + \frac{1/2}{\eta_{it_0} - 1} \right] \Delta s_i &= \sum_{i \in x} \frac{s_{it_0}}{2(\eta_{it_0} - 1)} - \sum_{i \in n} \frac{s_{it_1}}{2(\eta_{it_1} - 1)} \\ &= \frac{s_{xt_0}}{2\left(\eta_{xt_0} - 1\right)} - \frac{s_{nt_1}}{2\left(\eta_{nt_1} - 1\right)} \end{aligned} \tag{1.23}$$

where s_{xt_0} and s_{nt_1} are just the total share of spending on exiting goods in the initial period and entering goods in the final period, while $\eta_{xt_0} - 1$ and $\eta_{nt_1} - 1$ are harmonic means of the the product-specific terms. Written out for the exiting goods case, this is:

$$s_{xt_0} \equiv \sum_{i \in x} s_{it_0} \quad \eta_{xt_0} - 1 \equiv \left[\sum_{i \in x} \frac{s_{it_0}}{s_{xt_0}} \left(\frac{1}{\eta_{it_0} - 1} \right) \right]^{-1} \tag{1.24}$$

The fact that the relevant average is a harmonic mean provides some mechanical force leading towards larger entry and exit effects. For a given distribution of elasticities, harmonic means are generally lower than other types of averages. If there is any variation in the values being averaged, the harmonic mean is always less than the arithmetic, logarithmic, and geometric means. In addition, a mean-preserving spread[28] of the distribution of values being averaged leads to a reduction in the harmonic mean.

Remark: Grouping and Entry/Exit Adjustment The influence of the choice of grouping in this sufficient statistic contrasts with standard results for price index formulas without entry and exit. The well-known Laspeyres and Paasche price indices have a convenient numerical property called consistency in aggregation, meaning that aggregating Laspeyres

[28] i.e. a change in the distribution that leaves the arithmetic mean unchanged

(resp. Paasche) price indices yields the same result as applying a Laspeyres (Paasche) index to underlying product-level data. Tornqvist indices, and other similar indices, are affected by the choice of nesting but mechanical numerical properties mean this generally doesn't affect the result substantially (Diewert, 1978).[29]

As exemplified in equation (1.21) and (1.22), with entry/exit adjustment the choice of grouping has a first-order effect even after the price elasticities of entering and exiting goods are taken into account. This is reflected in the difference between $1/s_c$ and s_g/s_{cg}. The first expression, which would be applicable when there is only a single group, increases the weight given to all continuing goods by a single common factor. On the other hand, when there are multiple groups then price changes for goods in groups with more entry and exit are given greater weight in the total index. This also creates the potential for the continuing goods consumer surplus adjustment to be larger (or smaller) for some groups; as the size of the group gets smaller for any given amount of entry and exit s_g/s_{cg} will increase faster.

1.4.3 Constant Elasticity of Substitution (CES) Benchmark

The constant elasticity of substitution (CES) virtual price expenditure function may be written as:

$$E^*(\boldsymbol{p}^*, V) = V \underbrace{\left(\sum_{i \in \Omega} \alpha_i p_i^{*1-\eta} \right)^{\frac{1}{1-\eta}}}_{P(\boldsymbol{p}^*)}$$

It is straightforward to see that CES is a case of a GSA demand systems as the optimal choices for expenditure shares may be written as:

$$s_i^* = \alpha_i \left(p_i^* / P \right)^{1-\eta}$$

---

[29]For a Tornqvist index, the full-sample index may be written as $\sum_i \bar{s}_i \Delta \ln p_i = \sum_g \bar{s}_g \sum_i \bar{s}_i^g \Delta \ln p_i + (1/4) \sum_g \Delta s_g \sum_{i \in g} \Delta s_i^g \Delta \ln p_i$ where the $\sum_g \bar{s}_g \sum_{i \in g} \bar{s}_i^g \Delta \ln p_i$ is a Tornqvist-of-Tornqvists, i.e. a Tornqvist aggregation over groups where each group is given its own Tornqvist index. The statement that a Tornqvist is "approximately consistent in aggregation" simply means that mechanically the remainder term will tend to be small.

where, in this case, the base price term happens to also be the welfare-relevant price index.

In the CES case, the partial elasticity term is constant and common for all products, so that we have $\eta_{it} = \eta$. That is, CES shuts down both the cross-sectional and panel sources of variation in elasticities. In addition, the constant elasticity feature of CES implies unusual behavior for demand as prices increase. When $\eta > 1$, demand for a product may be driven to zero only as the price rises to infinity.[30] However, the fact that the virtual prices for missing goods are unbounded means the "purchased goods" representation of the expenditure function corresponds to just limiting the summation index for the price index to include goods available at each point in time.[31]

A discrete analog of equation (1.21) that is exact for CES demand is given by:

$$\begin{aligned}\Delta \ln P^{CES} &= \underbrace{\left(\sum_{i\in c} w_i\right)^{-1}\left[\sum_{i\in c} w_i \Delta \ln p_i + \frac{\Delta s_c}{\eta - 1}\right]}_{\Delta \ln A} \\ &= \underbrace{\left(\sum_{i\in c} w_i\right)^{-1}\sum_{i\in c} w_i \Delta \ln p_i}_{\substack{\text{Continuing}\\ \text{Goods}\\ \text{Index}}} + \underbrace{\left[\left(\sum_{i\in c} w_i\right)^{-1} - 1\right]\sum_{i\in c}\frac{\Delta s_i}{\eta - 1}}_{\substack{\text{Continuing Goods}\\ \text{Partial Consumer Surplus}\\ \text{Adjustment}}} + \underbrace{\sum_{i\in x}\frac{s_{it_0}}{\eta - 1} - \sum_{i\in n}\frac{s_{it_1}}{\eta - 1}}_{\substack{\text{Net Entry/Exit}\\ \text{Partial}\\ \text{Consumer Surplus}}}\end{aligned} \tag{1.25}$$

where the (un-normalized) weight terms are logarithmic averages[32] of expenditure shares, i.e. $w_i = \Delta s_i / \Delta \ln s_i$. When there is no entry and exit this expression reduces to the Sato-Vartia price index, which offers a calibration-free sufficient statistic for any

[30] When $\eta \leq 1$, although the full demand curve is downward sloping the expenditure share cannot be driven to zero even as the price rises to infinity. Thus, while CES is compatible with products being complements or neutral it cannot rationalize complementarity and entry/exit at the same time.

[31] The multinomial logit case (which is not homothetic) is discussed in more detail in the appendix. Multinomial logit also features unbounded reservation prices and also has this convenient property that the purchased goods representation corresponds to simply truncating the set of products being summed over.

[32] In the limit, as the change in the expenditure share approaches zero, the logarithmic average converges to the constant level of spending, i.e. $s_{it_0} = s_{it_1}$ implies $w_i = s_{it_0}$.

CES preference regardless of the value for η.[33] Relative to equation (1.22), there are two notable features of the sufficient statistic for CES. First, because the distribution of the elasticities is degenerate the average elasticity is just equal to the parameter $\eta - 1$. Second, because the elasticity never declines we no longer get the division by one half discussed in equation (1.23).

This sufficient statistic for the CES price index is very similar, but not identical, to the well-known formula of Feenstra (1994).[34] As I show in the appendix, this can be interpreted as the Feenstra (1994) index providing a discrete analog for the CES price index integral using a particular path for price changes. One advantage of the version of the CES index given in equation (1.25) is that it captures the fact that in the CES case the price index and the GSA base-price term are the same object. Using either Feenstra (1994) or equation (1.25) the net effect of entry and exit for CES has a convenient compact form: the change in the fraction of spending on continuing goods adjusted for the (common) elasticity value.

The CES formula is particularly attractive due to its modest data requirements and ease of implementation. However, CES imposes problematic behaviors for the infra-marginal gains of entry/exit while also imposing strong restrictions on both the cross-sectional and panel dimensions of the elasticity distribution. The cross-sectional restriction of CES may be partly relaxed by using a nested expenditure function with CES demand in each nest, but this retains the strong panel-dimension restrictions (and unbounded reservation prices). Even with many nests, CES cannot accommodate product-specific elasticities because all nests must have more than one good. In addition, with entry/exit adjustment the choice of grouping structure is not as innocuous as it is in the standard price index case, so that the tight link between own- and cross-price effects imposed

[33] See Sato (1976) and Vartia (1976). Although the Sato-Vartia index is not "superlative" by the definition of Diewert (1976) in practice it typically provides a close numerical analog. For example, log-averages are typically close to the arithmetic averages using in the Tornqvist index. For more on the approximation properties of the Sato-Vartia index, see Barnett and Choi (2008).

[34] While the Feenstra (1994) is the standard approach, Hsieh et al. (2020) has recently used a price index formula similar to equation 1.25.

using nests of CES may be undesirable.

Given these limitations of the CES approach, we now turn to finding a convenient parametric demand system that retains the high degree of flexibility present in the GSA entry/exit adjusted price index while providing tractability and imposing reasonable assumptions on the panel- and reservation price characteristics of the demand system.

1.5 Group Secondary Aggregate (GSA) Translog

The homothetic translog, first introduced in Christensen et al. (1975), features a high degree of flexibility as it provides a local second-order approximation to an arbitrary homothetic demand system. This flexibility implies a curse of dimensionality, however, as the number of demand parameters grows with the square of the number of products. In addition, due to translog featuring finite reservation prices the demand parameters that characterize consumer behavior when products are unavailable are generally different from the parameters that encode preferences over all goods regardless of their availability.

The group secondary aggregate (GSA) translog combines translog with the GSA restrictions discussed above. I show that GSA translog yields a price sensitivity parameter that is fixed even as the set of purchased goods is allowed to adjust. GSA translog retains full flexibility for the cross-section of elasticities and also satisfies Marshall's laws of demand. GSA translog features finite reservation prices and for GSA translog the formula from equation (1.22) turns out to be an exact (rather than approximate) sufficient statistic for the consumer's price index.

1.5.1 Unrestricted Homothetic Translog

In this section, I describe the homothetic translog expenditure function. In addition, I show how entry and exit leads to endogenous changes in substitution patterns, yielding a "purchased goods" representation of the translog demand system.

Definition 8 (Homothetic Translog) *The virtual price expenditure function for some set of*

goods Ω is homothetic translog when

$$\ln E^*(\boldsymbol{p}^*, V) = \ln V + \underbrace{\alpha_0 + \boldsymbol{\alpha}' \ln \boldsymbol{p}^* + \frac{1}{2} \ln \boldsymbol{p}^{*'} \boldsymbol{\Gamma} \ln \boldsymbol{p}^*}_{\ln P(\boldsymbol{p}^*)} \tag{1.26}$$

with the following restrictions:

$$\boldsymbol{\alpha}' \mathbb{1} = 1 \qquad \boldsymbol{\Gamma} \mathbb{1} = \mathbf{0} \qquad \boldsymbol{\Gamma} = \boldsymbol{\Gamma}'$$

Rewriting this in scalar notation, we have:

$$\ln E^*(\boldsymbol{p}^*, V) = \ln V + \alpha_0 + \sum_{i \in \Omega} \alpha_i \ln p_i^* + (1/2) \sum_{i \in \Omega} \sum_{j \in \Omega} \gamma_{ij} \ln p_i^* \ln p_j^*$$

The adding up and symmetry restrictions ensure that this function satisfies the restrictions implied by cost minimization. Economic theory requires that an expenditure function be homogeneous of degree 1 in prices, which is ensured by the adding up conditions on $\boldsymbol{\alpha}$ and $\boldsymbol{\Gamma}$.[35] In addition, the symmetry of the $\boldsymbol{\Gamma}$ matrix ensures the expenditure function is twice-differentiable and that the Slutsky matrix is symmetric. Given the adding up and symmetry conditions, there are $n(n-1)/2$ free parameters in the $\boldsymbol{\Gamma}$ matrix. In addition, there are $n-1$ degrees of freedom in the α parameters. The fact that the number of parameters in the $\boldsymbol{\Gamma}$ matrix grows with the square of the number of products reflects the rapid growth in the number of cross-price parameters that must be estimated, leading to the curse of dimensionality.

Given this specification of the expenditure function, Shepard's lemma implies that the

[35] A homothetic translog indirect utility can be written identically to the above, simply replacing $\ln V$ with $\ln Y$ where Y is the level of income. In the indirect utility case, the adding up restrictions on γ_{ij} ensure the preference is homothetic, while the requirement that $\sum_i \alpha_i = 1$ is a convenient, but not strictly required, normalization.

associated demand system[36] is given by

$$s_i^*(\boldsymbol{p}^*) = \alpha_i + \sum_{j \in \Omega} \gamma_{ij} \ln p_j^* \tag{1.27}$$

Here, we can see that the α_i parameter denotes how popular product i would be in the hypothetical setting where all goods have the same price.[37] In turn, the γ_{ij} terms give product-by-product adjustments to the expenditure share for good i as prices move away from a uniform value. The sign of γ_{ij} characterizes whether products are gross substitutes or gross complements.

While not strictly required by consumer theory, there are two additional restrictions that are often placed on the parameters of the translog demand system. First, to ensure elastic demand and the Marshall's second law condition it is necessary and sufficient to impose $\gamma_{ii} < 0$. Second, a sufficient (but not necessary) condition for the Slutsky matrix to be negative-semidefinite is to impose that the $\mathbf{\Gamma}$ matrix of price effects is negative-semidefinite.[38]

Finally, a useful property of the homothetic translog is that we can solve for the "purchased goods" representation for the expenditure function when only a subset of goods have non-zero quantities.

Proposition 5 (Translog - Purchased Goods Representation) *When the virtual price demand system is homothetic translog, the demand system corresponding to purchased goods* $\Omega_o \subset \Omega$

[36] as discussed in equation (1.13) and footnote (17).

[37] Note that α_i may be negative in which case we would need to invoke the purchased good demand system discussed in Proposition (5)

[38] Hurwicz and Uzawa (1971) shows that a symmetric negative-semidefinite Slutsky matrix for all prices and income levels ensures that a demand system can rationalized in terms of an underlying preference ordering. A discussion of concavity conditions for the translog case is given in Diewert and Wales (1987). In addition, a negative-semidefinite $\mathbf{\Gamma}$ ensures that the full own-price elasticities (γ_{ii}) are all negative. In the estimation exercise later in the paper, no concavity conditions are directly imposed.

and unpurchased goods $\Omega_m \equiv \Omega \setminus \Omega_o$ is given by:

$$\boldsymbol{s}_o(\boldsymbol{p}_o) = \boldsymbol{\alpha}_o + \boldsymbol{\Gamma}_{oo} \ln \boldsymbol{p}_o + \boldsymbol{\Gamma}_{om} \ln \boldsymbol{r}_m^*(\boldsymbol{p}_o, \Omega_o) = \tilde{\boldsymbol{\alpha}} + \tilde{\boldsymbol{\Gamma}} \ln \boldsymbol{p}_o$$

which can be rationalized by a homothetic translog of the form

$$\ln E_o(\boldsymbol{p}_o, V) = \ln V + \tilde{\alpha}_0 + \tilde{\boldsymbol{\alpha}}' \ln \boldsymbol{p}_o + \frac{1}{2} \ln \boldsymbol{p}_o' \tilde{\boldsymbol{\Gamma}} \ln \boldsymbol{p}_o$$

where the parameters of the purchased goods demand system are given by:

$$\begin{aligned}
\tilde{\alpha}_0 &= \alpha_0 - (1/2)\boldsymbol{\alpha}_m' \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\alpha}_m \\
\tilde{\boldsymbol{\alpha}} &= \boldsymbol{\alpha}_o - \boldsymbol{\Gamma}_{om} \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\alpha}_m \\
\tilde{\boldsymbol{\Gamma}} &= \boldsymbol{\Gamma}_{oo} - \boldsymbol{\Gamma}_{om} \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\Gamma}_{mo}
\end{aligned}$$

As noted in the discussion of equation (1.9), the purchased goods representation endogenizes the virtual prices of unavailable goods so that consumers always optimally choose to not purchase those products. The transition from $\boldsymbol{\Gamma}$ to $\tilde{\boldsymbol{\Gamma}}$ reflects a change in substitution patterns when product availability changes.[39] For example, when the first cereal with freeze-dried fruit entered the market consumers who chose that product might be relatively price-insensitive because there are few close substitutes. Over time, as more freeze-dried fruit products become available the initial variety may lose market share (as $\tilde{\boldsymbol{\alpha}}$ adjusts) and see changes in its price sensitivity (as $\tilde{\boldsymbol{\Gamma}}$ adjusts).

The presence of the available goods representation highlights the challenge that entry and exit poses for estimation of the demand system. On one hand, we would like to estimate the $\boldsymbol{\Gamma}$ values, but we do not observe all the prices needed to describe this system. On the other hand, conditioning only on the prices of purchased goods corresponds to

[39] This transformation relies on $\boldsymbol{\Gamma}_{mm}$ being invertible, which corresponds to the restriction that the consumer can, in fact, do without all of the products in Ω_m. For example if row i of $\boldsymbol{\Gamma}$ is full of zeros then the product corresponding to that row should always be consumed because its demand is insensitive to price changes (i.e. Cobb-Douglas).

a different, and ever-changing, set of demand parameters since substitution patterns adjust whenever we shift the set of purchased goods.

1.5.2 Group Secondary Aggregate (GSA) Translog

The group secondary aggregate (GSA) translog combines the translog and GSA restrictions. With these restrictions, I identify a price-sensitivity parameter that is invariant to product entry and exit - the partial own-price semi-elasticity. This partial semi-elasticity is central to both the exact entry/exit adjusted price index for GSA translog introduced in Proposition 7 and to the estimation strategy in section (1.6).

Definition 9 (Group Secondary Aggregate (GSA) Translog) *A translog expenditure function has the group secondary aggregate (GSA) form if the matrix of price effects may be written as*

$$\boldsymbol{\Gamma} = -\hat{\boldsymbol{d}}\left[\boldsymbol{I} - \boldsymbol{G}\boldsymbol{B}'\right] \tag{1.28}$$

where $\hat{\boldsymbol{d}}$ is the diagonal matrix with elements d_i, $\boldsymbol{G}$ is a $|\Omega|$-by-g matrix whose columns assign products to groups[40] *and $\boldsymbol{B}$ is a $|\Omega|$-by-g matrix with elements b_{ig}.*

Rewriting this in terms of individual γ_{ij} parameters, we have:

$$\gamma_{ii} = -d_i(1 - b_{ig(i)}) \qquad\qquad \gamma_{ij} = d_i b_{jg(i)} = d_j b_{ig(j)}$$

where $g(i)$ refers to the group to which product i belongs.[41]

This form for the $\boldsymbol{\Gamma}$ matrix allows us to combine the standard translog demand (equation 1.27) with the compact GSA aggregator (equation 1.16), giving us an expenditure share

[40] i.e. the columns of $\boldsymbol{G}$ are a set of exhaustive and mutually exclusive dummy variables. $G_{ig} = 1$ if $i \in g$ and $G_{ig} = 0$ otherwise.

[41] Note that the definition of groups need not be unique. This is because arbitrary subsets of a group are also groups. The condition that groups are fully consolidated is equivalent to requiring the $\boldsymbol{B}$ matrix is of full column rank.

function of the form

$$s_i^* = \alpha_i - d_i \left[\ln p_i^* - \ln A_{g(i)}^* \right] \tag{1.29}$$

where the base price term is given by $\ln A_g^* = \sum_{j \in \Omega} b_{jg} \ln p_j^*$. I refer to d_i as the partial semi-elasticity value, as it relates changes in the log of relative prices to the change in the level of shares.

Within the class of GSA demand systems, the GSA translog is particularly attractive for a few reasons. First, the demand system has a convenient linear form for the demand system. Second imposing the Marshall's law conditions for the partial (constant-A) demand curves is straightforward; both the first and second laws hold when the partial semi-elasticity parameter is positive[42] since with GSA translog we have

$$\eta_i = 1 + d_i / s_i \tag{1.30}$$

This also makes evaluating the elasticity at different points in time straightforward, as the only component that changes is the expenditure share which is consistent across both the virtual price function and the observed data. GSA translog provides a local second order approximation for any GSA demand system with the same grouping structure. While GSA translog imposes no cross-sectional restrictions on the d_i parameters (and thus no cross-sectional restrictions on elasticities) it does impose extra structure on the panel dimension of the elasticity distribution.

A key practical reason to focus on a characterization of the GSA translog that emphasizes the partial semi-elasticity parameter is that this component of the own-price effect is insensitive to product entry and exit as noted in Proposition 6, the second main result of this paper.

Proposition 6 (Stability of Partial Semi-Elasticity) *When the virtual price demand system*

[42] i.e. $d_i > 0$. Unlike with γ_{ii} or $\tilde{\gamma}_{ii}$, imposing that the $\mathbf{\Gamma}$ matrix is negative-semidefinite (in addition to the other GSA restrictions) is not sufficient to ensure this sign restriction for the d_i values. Instead if $\mathbf{\Gamma}$ is negative semi-definite there is, at most, one negative d_i value per group.

is GSA translog, any purchased goods demand system is also GSA translog with the same groups and the same partial semi-elasticity values as the virtual price demand system, i.e.

$$\boldsymbol{\Gamma} = -\begin{bmatrix} \hat{\boldsymbol{d}}_o & 0 \\ 0 & \hat{\boldsymbol{d}}_m \end{bmatrix}\left[\boldsymbol{I} - \begin{pmatrix} \boldsymbol{G}_o \\ \boldsymbol{G}_m \end{pmatrix}\begin{pmatrix} \boldsymbol{B}_o \\ \boldsymbol{B}_m \end{pmatrix}'\right] \implies \tilde{\boldsymbol{\Gamma}} = -\hat{\boldsymbol{d}}_o\left[\boldsymbol{I} - \boldsymbol{G}_o\tilde{\boldsymbol{B}}_o'\right]$$

where $\boldsymbol{I}$ is the conformable identity matrix and generally $\boldsymbol{B}_o \neq \tilde{\boldsymbol{B}}_o$.

Proposition 6 is an application of the Woodbury matrix identity together with the GSA restrictions and the definition of $\tilde{\boldsymbol{\Gamma}}$ from Proposition 5. When adjusting for changes in the set of available products, the full semi-elasticity value (γ_{ij} to $\tilde{\gamma}_{ij}$) adjusts just as it would in the general homothetic translog, but all of this is handled by changes in each product's contribution to the base price aggregators. This property is also useful for estimation, as there is now a stable product-specific parameter that we can hope to estimate even when there is product entry and exit.

The GSA translog yields a convenient entry/exit adjusted price index that only relies on observed spending and price data together with the product-specific partial semi-elasticity values, the third main result of this paper.

Proposition 7 (GSA Translog Entry/Exit Adjusted Price Index) *An exact sufficient statistic for the entry/exit adjusted price index when the demand system is GSA translog is given by:*

$$\Delta \ln P^{GSA\ TL} = \underbrace{\sum_g \bar{s}_g \sum_{i \in g \cap c} \frac{\bar{s}_i}{\bar{s}_{cg}} \Delta \ln p_i}_{\substack{\textit{Continuing} \\ \textit{Goods} \\ \textit{Index}}} + \underbrace{\sum_g \left[\frac{\bar{s}_g}{\bar{s}_{cg}} - 1\right] \sum_{i \in g \cap c} \frac{\bar{s}_i \Delta s_i}{d_i}}_{\substack{\textit{Continuing Goods} \\ \textit{Partial Consumer Surplus} \\ \textit{Adjustment}}} \underbrace{- \sum_{i \notin c} \frac{\bar{s}_i \Delta s_i}{d_i}}_{\substack{\textit{Net Entry/Exit} \\ \textit{Partial} \\ \textit{Consumer Surplus}}} \tag{1.31}$$

Replacing $\bar{s}_i / d_i$ in equation (1.31) with partial elasticity term from equation (1.30) shows that for GSA translog the approximation from equation (1.22) is exact. This mirrors the result from Diewert (1976) that the standard Tornqvist index (equation 1.15) is exact

for any homothetic translog demand system with stable underlying preferences.[43] In addition, since translog features finite reservation prices the simplification for the net entry/exit partial consumer surplus term provided in equation (1.23) applies to equation (1.31). Specifically, in the GSA translog case we have:

$$\frac{\bar{s}_i \Delta s_i}{d_i} = \left[\frac{1/2}{\eta_{it_1} - 1} + \frac{1/2}{\eta_{it_0} - 1} \right] \Delta s_i = \frac{s_{it_1}^2}{2d_i} - \frac{s_{it_0}^2}{2d_i} \tag{1.32}$$

Comparison to symmetric translog The flexibility provided by allowing products to have their own semi-elasticity parameters allows for a wider array of entry/exit effects than permitted with a symmetric translog, which requires that products all have a single common semi-elasticity value. In that case, an alternative version of equation (1.31) can be written which embeds a Herfindahl index[44] into the price index term:

$$\Delta \ln P^{\text{symmetric TL}} = \sum_g \bar{s}_g \underbrace{\sum_{i \in g \cap c} \frac{\bar{s}_i}{\bar{s}_{cg}} \left[\Delta \ln p_i + \frac{\Delta s_i}{d} \right]}_{\Delta \ln A_g} - \frac{1}{2d} \underbrace{\left[\sum_{i \in \Omega} s_{it_1}^2 - \sum_{i \in \Omega} s_{it_0}^2 \right]}_{\Delta \text{Herfindahl Index}} \tag{1.33}$$

With symmetric translog a decline in the Herfindahl index registers as an increase in the cost of living, an effect Feenstra and Weinstein (2017) refers to as a "crowding in product space" effect. This can also be interpreted as a consequence of the cross-sectional restriction that symmetric translog imposes on the elasticity distribution.[45] Consider the case where there are no changes in the prices or expenditure shares for any continuing goods, while one good exits and has its expenditure share replaced by two individually

[43] The fact that equation (1.31) is exact follows from the standard Tornqvist / translog result and the fact that a trapezoid rule approximation is exact for the base price adjustment and the net entry/exit partial consumer surpluses. More generally, Diewert (2002) notes the close connection between trapezoid rule approximations and quadratic forms.

[44] A similar term is present in the model of Fajgelbaum and Khandelwal (2016), which uses an AIDS demand system with symmetry in the price effects. In addition the symmetric QMOR discussed in Feenstra (2018) features a similar Herfindahl index-like term defined over (normalized) values of $s_i / p_i^{r/2}$ rather than s_i.

[45] The Herfindahl index term (scaled by $2d$) reflects the change in the sum of all partial consumer surpluses, rather than just the net entry/exit component.

less-popular products. By definition this lowers the Herfindahl index which, given the symmetric translog assumption, registers as an increase in the price index (hurting consumer welfare).[46] This also follows mechanically from the cross-sectional restriction on elasticities implied by symmetric translog, since the two new goods in this scenario must be elastictically demanded relative to the single popular (and thus inelastically demanded) good they are replacing, so that the gains from entry are outweighed by the losses from exit. For reference, in the CES case shown in equation (1.25) the same pattern of spending and price changes will always imply no change in the price index due to the single common elasticity parameter. By contrast in GSA translog the relationship between expenditure shares, elasticities, and entry/exit adjustment is not so rigid. Given this equal-expenditure-swap scenario the GSA price index could go up, or down, or stay constant as the answer hinges on the (independent) levels of spending and elasticities for entering and exiting goods.

GSA Translog Cross-Price Restrictions In addition to consolidating cross-price effects into an aggregator function, the GSA form for the demand system may be thought of as imposing proportionality on cross-price effects for products in the same group. This is a straightforward consequence of the definition of the GSA translog, since we have

$$\frac{\gamma_{ik}}{\gamma_{jk}} = \frac{d_i b_{gk}}{d_j b_{gk}} = \frac{d_i}{d_j} \quad i, j \in g;\ \forall k \neq i \neq j \tag{1.34}$$

The standard adding up conditions for elasticities require that goods that are elastically demanded are overall more sensitive to price changes of other goods.[47] The GSA assumption imposes that the (relative) increase in cross-price sensitivities needed to satisfy this condition is proportionally distributed across all goods in the demand system.

In the special case that there is only a single group, we recover the separable translog

[46]This exercise ignores how, or if, the underlying preference induces this pattern for the observed expenditure shares. Instead, this discussion just tells us the mechanical effect on the price index formula for the given scenario.

[47]This is necessary to ensure the budget constraint is binding.

demand system of Matsuyama and Ushchev (2017).[48] The restriction to a single group, together with the symmetry and adding up conditions for any translog $\mathbf{\Gamma}$ matrix, imply that in this case the full demand system is characterized by $\boldsymbol{\alpha}$ and $\boldsymbol{d}$ parameters since the aggregator weights must satisfy

$$b_i = \frac{d_i}{\sum_j d_j} \tag{1.35}$$

This shows that in the single-group case there are only n degrees of freedom in the GSA translog $\mathbf{\Gamma}$ matrix. More generally, the combination of the GSA form, symmetry, and adding up conditions implies that there are $n + g(g-1)/2$ degrees of freedom in the GSA translog $\mathbf{\Gamma}$ matrix.[49] Thus, the GSA translog has the same number of degrees of freedom as a nested demand system with a translog upper-tier aggregator and single-group GSA translog within each nest. As noted in Proposition (3b), however, the GSA notion of grouping allows for more flexibility for within-group cross-price effects. In addition, the GSA notion of grouping preserves linearity of the translog demand system defined over all goods while a nested-translog would only allow for linearity of the within-nest demand systems.

Given the convenience and tractability of the GSA translog, the last step in implementing entry/exit adjustment is to choose values for the semi-elasticity parameters. This challenge is taken up in the next section on the generalized random forest (GRF) estimation strategy.

---

[48]The parameterization of the separable translog in Matsuyama and Ushchev (2017) treats the aggregator-weights b_i, together with a free scaling parameter, as the primitives of the model. However, as shown in equation (1.35), the b_i parameters are not preserved when evaluating shifts among different available good representations.

[49]Strictly this count for the degrees of freedom assumes there are no "singleton" groups, i.e. every group has at least two products. If we instead allow for singleton groups, the count of degrees of freedom becomes $n - g^* + g(g-1)/2$ where g^* is the number of groups with only a single product. The loss of one degree of freedom for singleton groups may be thought of as reflecting the fact that without other group members the decomposition of γ_{ii} into d_i and $b_{ig(i)}$ components is not identified.

1.6 Generalized Random Forest (GRF) Estimation Strategy

Given the GSA translog entry/exit adjusted price index in equation (1.31), the only parameters that need to be estimated are the partial semi-elasticity values. As shown in Proposition 6, these values are also unaffected by entry and exit and thus we can use time-series variation in the data to estimate these parameters.

The novel challenge with GSA translog is how to retain flexibility in estimating product-specific parameters without imposing strong additional ex-ante restrictions on the cross-sectional patterns of elasticity in the data. To address this challenge, I use the generalized random forest (GRF) approach of Athey et al. (2019) utilizing my own implementation which adds a high-dimensional fixed effect for use in my panel setting.[50] GRF creates an adaptively-weighted moment condition that allows for non-parametric identification of the product-specific partial semi-elasticity. Standard concerns about endogeneity of the error term and price changes are accounted for using the cross-market instrumental variable proposed in Hausman (1996).

1.6.1 GSA Translog, Cross-market instrument, and Generalized Random Forest

The estimating equation for GSA translog corresponds to differencing the expenditure share function (equation 1.29) and adding an additional error term

$$\Delta s_{it} = -d_i \Delta \ln p_{it} + d_i \Delta \ln A^*_{g(i)t} + \varepsilon_{it} \tag{1.36}$$

where i indexes a product and t indexes the quarter. The error term ε_{it} can be interpreted as a taste shock for individual varieties or measurement error such as from using aggregated data.

The standard estimation challenge for a demand equation is the concern that demand

[50] The publicly available GRF implementation only supports simple linear regression, i.e. the moment condition includes a slope and a single intercept term.

shocks (ε_{it}) and price changes ($\Delta \ln(p_{it})$) may be correlated.[51] To overcome this endogeneity problem, I use the panel nature of the Nielsen data to add a region dimension to the data while instrumenting regional price changes using national leave-out means. This instrumental variable strategy was proposed in Hausman (1996) (where it is applied for AIDS demand systems) and has been used extensively for identification with a variety of demand system and estimation strategies (e.g. Nevo (2001, 2003) for mixed logit, or Faber and Fally (forthcoming) and Handbury (2021) for CES). Indexing regions by r, an observation is then an *i-t-r* combination, and the estimating equation becomes:

$$\Delta s_{itr} = -d_i \Delta \ln p_{itr} + d_i \Delta \ln A^*_{g(i)tr} + \varepsilon_{itr} \tag{1.37}$$

where each region's price change $\Delta \ln p_{irt}$ is instrumented using the national leave-out mean $\frac{1}{N_r - 1} \sum_{r' \neq r} \Delta \ln(p_{itr'})$.[52]

Many different possible distributions of the partial elasticity term are implied by popular demand system specifications. For example, if the CES assumption of a common elasticity is correct then $d_i = \beta s_i$ would be the correct function or if the logit form for the partial elasticity is the correct model then $d_i = \beta(p_i s_i) - s_i$ would be the correct function.

Rather than impose ex-ante which products should have similar semi-elasticity values, we can instead suppose that the d_i and $d_i \Delta \ln A^*_{g(i)tr}$ values are related to observable product characteristics $x_i \in \mathcal{X}$. Denoting the conditional-on-x_i values as $d_i = d(x_i)$ and $d_i \ln A^*_{g(i)tr} = \nu_{tr}(x_i)$ the estimation equation takes the form:

$$\Delta s_{itr} = -d(x_i) \Delta \ln p_{itr} + \nu_{tr}(x_i) + \varepsilon_{itr} \tag{1.38}$$

---

[51] In a standard supply and demand framework, this correlation is expected to lead to an underestimate of the demand elasticity. Positive demand shocks raise both the expenditure share and (by moving up along the supply curve) prices, so that the regression has an upward bias. Since the parameter should be negative an upward bias gives an attenuated estimate of the absolute value or delivers a theoretically proscribed positive estimate.

[52] Keeping the semi-elasticity value as only product-specific assumes that this value is common across all regions. Since the expression is written in differences, we still allow products to have different average levels of market share in different markets insofar as α_i, price levels, or base-prices differ across markets.

As noted in the discussion later on the choice of partitioning variables, in practice I use one value for each product over its life (e.g. its average share or its average deflated price) so that in practice I group products with similar values over their life.

In the formulation in equation (1.38), the problem of estimating $d_i = d(x_i)$ is phrased in terms of the generalized random forest (GRF) approach of Athey et al. (2019). The use of GRF carries with it a set of identifying assumptions. First, the standard exogenity and relevance requirements for an instrument must hold conditional on the auxiliary variables $\mathcal{X}$. Second, the expectation of the conditional-on-x_i IV moment condition must be Lipshitz continuous in the x_i values. In the current context, this second assumption imposes that products which are sufficiently similar in terms of x_i also have similar d_i values and are in the same group.[53]

Beyond the continuous-in-x_i and the identification-conditional-on-x_i conditions, there is no restriction on the particular relationship between x_i and $d(x_i)$. As explained in Athey et al. (2019), if these conditions are satisfied (along with some regularity requirements) then GRF produces estimates $\hat{d}(x)$ which are consistent for the underlying $d(x)$ value. In addition, Athey et al. (2019) also provide a method for constructing asymptotic confidence intervals around the $\hat{d}(x)$ estimates, although these are not estimated in the results below.

There are a few potential pitfalls which may affect estimation with GRF. First, a general concern for GRF, as with other bandwidth-based regression techniques, is that the algorithm may have difficulty matching the true model at the edge of the partitioning space ($\mathcal{X}$) or in other relatively isolated regions. In the results below, this is likely an issue when estimating d_i for the upper tail of high-expenditure share products. Second, the validity of the standard IV restrictions may vary within the partitioning space. Testing for issues with the validity of the IV for subsets of the products is left for future work.

[53] If there is only a single group (i.e. we assume a separable translog), then having a "similar" d_i value is sufficient since $\Delta \ln A_{tr}$ is common to all products in that case. If there are different groups, then we may treat this as saying that the probability of being in one group or another varies continuously (albeit potentially sharply) as we move in the space of partitioning values $\mathcal{X}$.

Threats to Identification The leave-out mean instrument captures components of price movements that are common across regions. A salutary explanation for these co-movements would be cost shocks that are passed through to consumers. The identifying assumption for the use of the leave-out mean is that errors for a given product-region-quarter (ε_{itr}) are uncorrelated with price changes from outside the region. A violation of this assumption would occur if there are geographically broad and simultaneous shocks to preferences and to prices. Notably, this allows for geographically broad taste shocks as long as they are not correlated with prices.[54]

In the GRF context, the IV moment condition must be valid conditional on the partitioning variables x_i. In my estimation procedure, all of the x_i values are time-invariant characteristics for each product, e.g. average deflated price or median market expenditure share. The extra conditioning will affect the validity of the IV exclusion if there are sets that that have similar *average levels* for price or share or some other variable over their life in the sample that are selected together by the tree and for which there are correlations in *changes* in prices and errors.

Comparison to Pooled Regression Estimation for CES and symmetric translog sidestep the problems posed by heterogeneous parameters by assuming that all products share a common price-sensitivity parameter. With these demand systems, it is natural to group all products together in a single regression. In this case the base price effect can be handled using a (within-market) time fixed effect, as in Faber and Fally (2017) (CES). A prominent alternative to linear IV is the double-difference approach as in Feenstra (1994) (CES) and Feenstra and Weinstein (2017) (symemtric translog). To use a fixed effect in the regression, we must have at least two products grouped together; to use a double-difference we must assume that at least four products are appropriately matched.[55]

[54]When demand systems are estimated in levels an additional concern is that price levels may be correlated with unobserved characteristics; since my estimation is done in differences this concern is not present in my context.

[55]In general, the number of products required to be in a group for a double-difference estimator is one plus the number of right-hand side variables. In the standard Feenstra (1994) regression with a constant, there are three RHS variables so we require at least four products (three for the rows of the

In the case of a linear IV regression, as in Faber and Fally (2017) for the CES case, if there is unmodeled underlying heterogeneity the estimated coefficient for the pooled regression returns a weighted average of the underlying heterogeneous terms (Imbens and Angrist, 1994).[56] A similar local average effect interpretation may be given to the parameters generated by the GRF algorithm. However, no such local average effect results are present for the double-difference estimation techniques.[57] In addition, the weighted averages generated from pooled regressions will not typically match the relevant average elasticities for entry/exit adjustment discussed in equation (1.23).

1.6.2 Generalized Random Forest Algorithm

In this section, I review the GRF algorithm for estimating a heterogenous treatment effect. This discussion largely mirrors the description given in Athey et al. (2019). More details on the GRF approach are available in the Athey et al. (2019) paper or in the documentation for the publicly available GRF implementation available on CRAN.[58] Details on my own implementation of GRF with a single (time-region) high-dimensional fixed effect are given in the appendix.

While in my context, an observation in the data is a product-quarter-region (itr), to economize on notation in this section I refer to an individual observation simply using a subscript i. GRF supposes that we observe data $(O_i, X_i) \in \mathcal{O} \times \mathcal{X}$ and that we wish to estimate a quantity $d(x)$ associated with a conditional moment condition of the form

$$\mathbb{E}\left[\psi_{d(x),\nu(x)}(O_i)|X_i = x\right] = 0 \quad \text{for all } X_i \in \mathcal{X} \tag{1.39}$$

design matrix, and one for the reference-product difference). In addition to imposing that products have a common demand parameter, standard double-difference methods require that products included in a regression share a common supply parameter (typically the elasticity of supply). Soderbery (2018) discusses extensions to the Feenstra (1994) method that allow for heterogeneous supply parameters.

[56]Strictly, this result requires that the heterogeneity in the slope term is uncorrelated with the heterogeneity, if any, in the intercept term.

[57]In ongoing work, I show that Feenstra (1994) style double-difference estimators are potentially quite sensitive to improperly pooling products with different parameters.

[58]See documentation at `https://grf-labs.github.io/grf/REFERENCE.html`

where $\psi(O_i)$ is a score function that depends on $d(x)$ and nuisance parameters $\nu(x)$. For example, in my application O_i is vector of values that directly enter the moment condition (share changes, price changes, price instrument, region-quarter dummies) while the X_i value corresponds to the vector of product characteristics (e.g. average share, average price).

GRF estimates $d(x)$, along with the nuisance parameters $\nu(x)$, by inducing a set of weights $\omega(x)$ and solving the minimization problem:

$$\left(\hat{d}(x), \hat{\nu}(x)\right) = \text{argmin}_{d(x),\nu(x)} \left\{ \left\| \sum_{i \in 1}^{n} \omega_i(x) \psi_{d(x),\nu(x)} \left(O_i\right) \right\|_2 \right\} \tag{1.40}$$

The weights $\omega_i(x)$ are based on a forest-growing algorithm described in more detail below. A forest in the GRF case refers to a collection of "trees" where a single tree is characterized by two objects. First, a tree provides a partitioning rule that groups observations based on their $X_i \in \mathcal{X}$ value. Although this partitioning is initially specified based on a specific data sample, the assignment rule can be applied to observations not used in initially growing the tree. Second, a GRF tree contains a set of "leaf" nodes populated with a subset of observations assigned based on the tree's partitioning rule. Importantly, the observations used to populate the leaf node need not be from the same sample used to construct the partitioning rule. Given the full collection of GRF trees, we can construct the $\omega(x)$ weighting vector. Indexing trees by $b = 1, 2, ..., B$, the $\omega_i(x)$ weights for estimating $\hat{d}(x)$ are given by:

$$\omega_{bi}(x) = \frac{1\left(X_i \in L_b(x)\right)}{|L_b(x)|} \qquad \omega_i(x) = \frac{1}{B} \sum_{b=1}^{B} \omega_{bi}(x) \tag{1.41}$$

where $L_b(x)$ refers to the leaf node corresponding to x in tree b, and $|L_b(x)|$ is the number of observations populated in leaf node $L_b(x)$. The single-tree weight $\omega_{bi}(x)$ places equal weight on all observations that are in the leaf node associated with x, while the forest-based weights $\omega_i(x)$ average over all of these single-tree weights.

Growing a tree refers to the process of choosing a partition rule. An exhaustive search of all possible partitions quickly becomes infeasible. Instead the standard approach is to use a greedy step-wise method, at each stage finding the best binary (single two-way) split. A tree-growing algorithm, then, is an iterative procedure for specifying a set of candidate splits, evaluating which split is "best" at each step, and a stopping rule that determines when further splits will no longer be considered.[59]

In the case of GRF, the criteria for evaluating splits is based on the influence function for $\hat{d}$. At each step of the tree-growing procedure, the set of observations from the tree-growing subsample that are currently grouped together is called a node. Given a currently existing node (the parent), GRF tries to find a split that will maximize the difference between $\hat{d}$ values estimated in the two successor (child) nodes.[60] Within the parent node, we can fit the moment condition $\psi_{d,v}(O_i)$ over $i \in P$, yielding estimates $\hat{d}_p$ and $\hat{v}_p$. In turn, we can construct "pseudo-outcomes" ρ_i as:

$$\rho_i = -\xi' A_p^{-1} \psi_{\hat{d}_p, \hat{v}_p}(O_i) \in \mathbb{R} \tag{1.42}$$

where $\psi_{\hat{d}_p, \hat{v}_p}(O_i)$ is observation i's score from fitting the (unweighted) moment condition in the parent node, ξ extracts the element corresponding to the parameter of interest $(\hat{d})$, and A_p is a consistent estimate of the gradient of the score function. The ρ_i value is the influence of observation i on the estimation of $\hat{d}_p$ in the parent node. When the score function ψ is continuously differentiable, A_p corresponds to:

$$A_p = \frac{1}{|i : X_i \in P|} \sum_{i: X_i \in P} \nabla \psi_{\hat{d}_p, \hat{v}_p}(O_i) \tag{1.43}$$

Given pseudo-outcomes ρ_i, GRF evaluates potential splits of the parent node P into two

[59]For a standard regression or classification tree, a final ingredient is to assign a value for some target outcome once an observation is placed in a leaf node of the tree. That step is not present in the GRF context.

[60]This influence function-based criteria is a gradient based approximation to (the computationally intensive task) of an exact evaluation of different parameter estimates in the potential child nodes.

disjoint child nodes $C_1 \cup C_2 = P$ using the variance-reduction criteria of the widely-used CART algorithm of Breiman et al. (1984). This variance-reduction criteria[61] can also be rephrased as maximizing the following quantity:

$$\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{|\{i : X_i \in C_j\}|} \left(\sum_{\{i:X_i \in C_j\}} \rho_i \right)^2 \tag{1.44}$$

The CART algorithm used in GRF limits the set of potential splits to be binary axis-aligned splits based on the set of partitioning variables $\mathcal{X}$. For ordered variables, the restriction to axis-aligned splits leads to the following selection process. Suppose $\mathcal{X} = \mathbb{R}^J$ and that we denote x_{ij} as the value of the j-th partitioning variable for the i-th observation. A possible split based on variable j then takes the form:

$$C_1 = \left\{ i : i \in P \wedge x_{ij} \leq \bar{x}_j \right\} \qquad C_2 = \left\{ i : i \in P \wedge x_{ij} > \bar{x}_j \right\} = P \setminus C_1 \tag{1.45}$$

for some $\bar{x}_j$. For a given partitioning variable j, every value of $\bar{x}_j$ is evaluated based on the variance-reduction criteria given above, giving rise to the best within-j split $(C_1, C_2)_j$ with corresponding split point $\bar{x}_j^*$. In turn, the actual split corresponds to choosing the j^* that optimizes the variance-reduction criteria evaluated over each of the best within-j splits.

In addition to restricting attention to axis-aligned splits, GRF imposes three further limitations on the splits that may be considered at each step. First, at each node only a random subset of variables in $\mathcal{X}$ are considered for possible splits at that step. Second, we require that a child node have a minimum proportion of the observations from the parent node. Third, we also impose a minimum number of observations in a node; if a

[61] Maximizing the criteria given in the main text is equivalent to minimizing the residual sum of squares criteria for a simple linear regression over ρ_i. That is, the same C_1 and C_2 partition that minimizes $\sum_{i \in C_1} (\rho_i - \bar{\rho}_1)^2 + \sum_{i \in C_2} (\rho_i - \bar{\rho}_2)^2$ also maximizes the criteria above, where $\bar{\rho}_1$ and $\bar{\rho}_2$ are the sample mean of ρ_i for $i \in C_1$ and $i \in C_2$, respectively.

node has too few observations then no further splits are attempted.[62]

When growing a forest with GRF, "honest" subsample selection is utilized for each tree.[63] Honest subsampling separates the sample which is used to select a given model structure (in this case, the tree partition) from the sample used for estimation conditional on a model structure. Denoting the full estimation sample as $\mathcal{S}$, we first select a subset $\mathcal{I}_b \subset \mathcal{S}$ for use with tree b by sampling without replacement. In turn, there is a further split of $\mathcal{I}_b$ into the disjoint partition $\mathcal{J}_{1b} \cup \mathcal{J}_{2b} = \mathcal{I}_b$. The $\mathcal{J}_{1b}$ sample is used when applying the tree-growing algorithm described above (i.e. when estimating $\psi_{\hat{d},\hat{\nu}}(O_i)$ at each step or evaluating splits we use observations $i \in \mathcal{J}_{1b}$). The $\mathcal{J}_{2b}$ sample is used to populate the leaf nodes of tree b. That is, sample $\mathcal{J}_{1b}$ is used for choosing the partitioning rule associated with tree b while sample $\mathcal{J}_{2b}$ is used to turn this partitioning rule into weights for the ultimate goal of estimating $\hat{d}(x)$. In particular, the within-tree weights ω_{bi} are equal to 0 for all $i \notin \mathcal{J}_{2b}$ regardless of what leaf node i would be assigned to given its x_i value.

In implementing this algorithm, there are a number of tuning parameters that need to be selected. The main tuning parameters are described in the table below. The standard practice in the machine learning literature is to choose tuning parameters using cross-validation (i.e. testing many possible tuning values on subsets of the data and selecting a "best" set of tuning parameters based on some model-comparison criteria). At present, there is not a clear criteria for cross-validation for instrumental variable GRF.[64] Qualitatively, the results for most products do not seem to be significantly affected by changing the value of the tuning parameters.

---

[62]Details on how these restrictions are applied in practice are available in the GRF documentation and in the appendix.

[63]The "honesty" approach to sampling was introduced in Athey and Imbens (2016) and is discussed in Wager and Athey (2018). While honest subsampling is not necessary to utilize the GRF algorithm, honest subsampling underlies the theoretical results of consistency and asymptotic Gaussianity.

[64]For the closely-related causal forest method, Nie and Wager (forthcoming) propose the R-learner as a cross-validation criteria.

Table 1.2: Generalized Random Forest: Parameter Values

Tuning Parameter	Description
num.trees	Number of trees to grow in a forest
honesty	TRUE/FALSE, whether to use honest subsample or not
sample.fraction	fraction of full sample $(\mathcal{S})$ to use in tree sample $\mathcal{I}_b$
honesty.fraction	if using honest subsampling, fraction of tree sample $\mathcal{I}_b$ placed in tree-growing sample $\mathcal{I}_{1b}$
mtry	at each node, randomly choose k partitioning variables to attempt splits where k is equal to $\min\{\max\{\text{Poisson}(mtry), 1\}, p\}$ where p is the number of potential partitioning variables.
alpha	minimum fraction of parent node P assigned to a child node
min.node.size	target for minimum number of observations in a leaf node

NOTE.—This table reviews the tuning parameters for the generalized random forest algorithm.

1.7 Empirical Setting

As an application of the GSA/GRF procedure, I study the cereal market in the Nielsen Consumer Panel dataset. This setting has been studied in the empirical industrial organization literature (Hausman 1996; Nevo 2001, 2003), and provides an ideal setting for my estimation strategy. The cereal market is one of the largest modules in the Nielsen Consumer Panel data in terms of the number of unique products present in the data. As in most modules, product entry and exit is a pervasive feature of the data.

1.7.1 Nielsen Consumer Panel Cereal Market

Nielsen Consumer Panel tracks the purchasing behavior of about 60,000 panelists every year as they shop in a wide variety of consumer goods categories.[65] A "product" in Nielsen Consumer Panel is defined as a unique Universal Product Code (UPC), corresponding to the barcode printed on product packaging.[66] Due to inventory management

[65] In 2004-2006, the panel included about 40,000 households.

[66] Although not a concern in the cereal data I focus on, non-barcoded or store-barcode purchases such as in-store bakery items are handled separately in the Nielsen data.

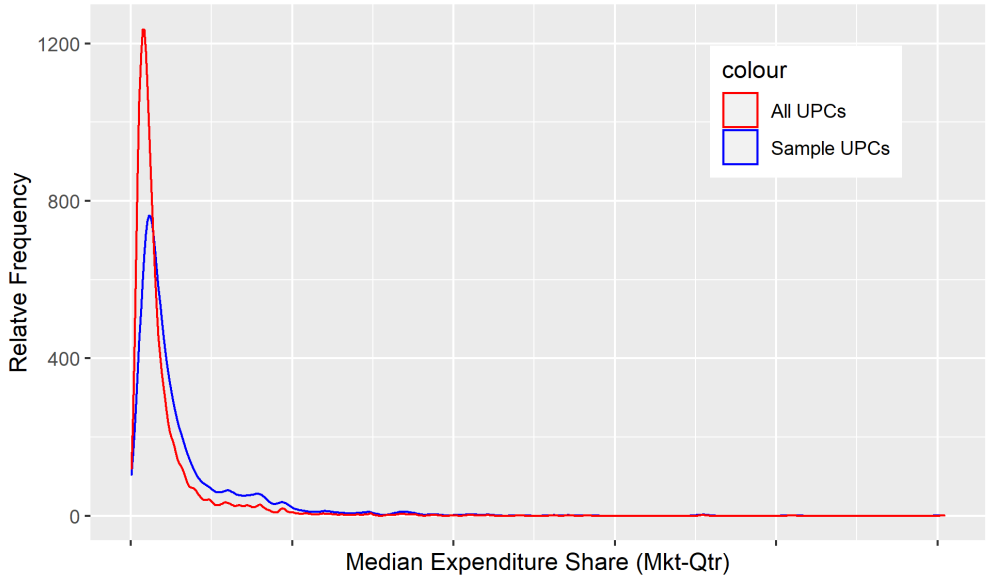
concerns, barcodes are updated even for relatively minor product changes.[67] Panelists record their shopping list, directly scanning the barcodes for items they purchased, and recording the prices they actually paid net of any coupons or discounts. Nielsen also features a variety of product metadata such as organic goods certifications, product size, and flavor notes. Nielsen divides products into three broad nested categories - departments, groups, and modules - based on how goods are generally placed together on store shelves.

The Nielsen Consumer Panel also includes a variety of consumer demographic information for panelists. In particular, Nielsen divides stores and panelists into a set of market areas (roughly major US metropolitan areas) to create a nationally representative panel. Panelists in the data are assigned weights, allowing purchases by the panel to be projected to a representative sample at the market area geography. These market areas are the level of aggregation that I use in my empirical exercise below.

For my empirical exercise, I focus on the ready-to-eat cereal market from 2004 to 2016. One reason to focus on the cereal market is that, even at the lowest level of product grouping specified by Nielsen, there is a large number of unique products available. Between 2004 and 2016, there are 3955 unique cereal UPCs that are purchased by panelists, although only 579 to 1065 are purchased in any given quarter. The median number of quarters in which a UPC is purchased is 8, out of a sample of 52 quarters. In addition, on a quarter-by-quarter basis, between 80 and 95 percent of total spending nationally is on products available in the previous quarter. Within a market-quarter (e.g. Los Angeles metro area in 2005Q1), the vast majority of UPCs have a very low expenditure share. The median expenditure share among purchased UPCs across all market-quarters is only about 0.2 percent. As shown in Figure 1.2, the distribution of region-quarter expenditure shares is concentrated among low-expenditure share products.

[67] Nielsen also provides a UPC version number to track when changes are made to products that are not reflected in a new UPC. This version control ensures that product metadata, such as flavor or size, are constant whenever a given product (UPC plus version number) appears in the Nielsen data. Whenever I refer to UPCs, I am referring to this UPC plus version number pair.

Figure 1.2: Expenditure Share Distribution



NOTE.—This figure shows the distribution of (within-module) expenditure shares for all products and those included in the regression sample. While most products have a very low market share, there is a long tail of more popular UPCs.

The sample that is used for constructing d_i estimates drops products with only a few observations. Specifically, when calibrating the GRF forest I limit the sample to UPCs that have at least 100 region-quarter observations and are present in at least 5 regions in a given quarter. In this limited sample, there are about 1500 UPCs, and the distribution of expenditure shares is somewhat right-shifted towards higher-expenditure share products. However, the overall pattern that most products have relatively low expenditure shares is still present in the GRF sample.

1.7.2 Partitioning Variable Selection

In addition to the tuning parameters, a user must also choose what partitioning variables to include in $\mathcal{X}$. There are a variety of considerations for what variables to include. In this section, we discuss some of the practical issues with choosing variables to include and specify which variables are used in the exercise below. In addition to the concerns given here, the partitioning variables must also be compatible with the identification conditions discussed earlier.

Forest-based methods are able to accommodate many potential partitioning variables. Adding irrelevant variables is relatively harmless as these variables are effectively ignored by rarely being chosen as the basis for splits. For closely correlated relevant variables, at each node a tree will tend to randomly choose one of the correlated variables as the one to actually split on. This has little effect on the actual split chosen[68], although it can complicate the interpretation of which variables distinguish each observation.

The key numerical feature of a partitioning variable is the implicit ordering it imposes on observations. The only way that partitioning variables are used is to determine how to split samples, and the algorithm only uses interior cutoffs. Thus, any transformation of the partitioning variables that preserves the within-variable ordering of observations will have no effect on the splits chosen. Numerically coded ordered categorical variables are treated identically to continuous variables.

Although many tree-growing algorithms can accommodate unordered categorical variables, the current GRF implementation does not accommodate these types of partitioning variables. One way to handle unordered categorical variables is to use dummy variables for each category. However, as discussed below, this approach may be unsatisfactory because the limited variation within a single dummy variable may make it unlikely to be used to split at any given node. A second approach may be to construct some continuous summary variables which capture the heterogeneity embedded in the categorical variable.

The CART splitting criteria incorporated in GRF tends to choose splits based on variables with more unique values. That is, a partitioning variable that more finely groups observations will mechanically be more likely to be used to split the sample rather than a variable that groups observations more coarsely. A partitioning variable that more finely groups observations has more potential split points and thus will tend to pick up more variation in the data, even when that variation is purely random. Alternative tree-growing algorithms, such as the model-based recursive partitioning (MOB) approach of

[68]By definition, correlated variables should lead to splits where the same group of observations go to one or the other child node.

Zeileis et al. (2008), try to adjust the split criteria to ameliorate this variable selection bias.

In the application in the Nielsen data, we are estimating a time-invariant parameter for UPCs (barcodes) within a product module. In addition to individual products, within a module products are grouped into brands (e.g. different variations or packages of a given product type). Thus, we use a mix of UPC-level and brand-level characteristics. To enforce that all observations for a UPC are always assigned to the same leaf node of a tree, the partitioning variables have a uniform value within each UPC. This choice of partitioning variables ensure that all observations of a given UPC across different markets and quarters are always assigned to the same nodes of the tree. This assumption affords significant reductions in computational time and is suitable for the national price index exercise that I focus on below.[69]

As partitioning variables, I use a mix of price, share, and household purchases information. Median values, either over quarters (for national data) or market-quarters, are taken to avoid allowing products to be differentiated by extreme observations. For shares, I use both the median share within markets and quarters that a product is available and the product's median national expenditure share. For price, I use the product's median national unit price level.[70] To avoid mechanically separating products available in different time periods due to overall inflation, prices are deflated using a continuing-goods Laspeyres index before taking the median. I also include the package size as a potential partitioning variable. The median number of households per market-quarter that purchase a product is also included as a potential partitioning variable.[71] In addition,

[69]The sampling process for the GRF exercise below is also done at the UPC level. That is, I draw a sample of 1/2 of all UPCs, and then split this sample of UPCs in half when applying the honest subsampling approach.

[70]Prices are recovered from the Nielsen data as the ratio of dollar purchases (net of coupons) divided by the number of product packages purchased. For example, a particular box of cereal may have $525 of total purchases and see 10 packages sold, so that the product price level is $5.25. In addition, this $5.25 box of cereal may have a size of 12 oz, in which case it's unit price would be $0.4375.

[71]As part of the data preparation process, only UPCs that, in total, have 20 or more unique households purchase that item are included in the sample.

the median national share of a product's brand is included as a potential partitioning variable, to try to keep UPCs within a brand grouped together.[72] Finally, there is a dummy variable for whether a brand is a "control" brand as defined by Nielsen. Including information on expenditure shares, prices, and the number of households purchasing a product can allow GSA translog to (locally) mirror the behavior of a CES or a logit demand system.[73]

1.8 Results

To assist in comparing the results of the GSA translog based estimate to previous work, Table 1.3 reports estimates of elasticities within the cereal market. The first three values correspond to direct estimates of the CES $\hat{\sigma}$ parameter using the same underlying data. The first row gives the Feenstra (1994) double-difference estimate, while the second row uses the double-difference estimation with the weighting suggested in Broda and Weinstein (2010). Both of these are based on national expenditure data rather than market-based data. The third row uses the cross-market Hausman IV together with the single-difference moment condition to estimate $\hat{\sigma}_i$. As noted in Faber and Fally (2021b), the Hausman IV estimate is notably lower than than either of the double-difference methods.

The last two columns report results from the industrial organization literature for the cereal market. The industrial organization papers use a mixed logit ((Nevo, 2001)) and an AIDS demand system (Hausman (1996)), both using a cross-market price instrument. While the "elasticity" in a CES and that I focus on below exclude price index effects, the elasticities reported in Nevo (2001) and Hausman (1996) are not "partial" values and thus are not exactly comparable to the estimates given here.

---

[72] All store-brand products are categorized in a single "control brand" category in Nielsen data. To assign store brands to separate brands, I group products based on the leading digits in their UPCs since these are generally assigned based on companies.

[73] As noted earlier, at a given level of consumption the full elasticity for a product in a CES demand system is $(\sigma - 1)(1 - s)$.

Table 1.3: Cereal Market Elasticities: CES Point Estimates and Literature Examples

Method/Paper	**UPC-level CES $\hat{\sigma}$**	**Brand-Level Elasticity (median)**	**Brand-Level Elasticity (median)**
National Double-Difference (Feenstra 1994 Weights)	7.27		
National Double-Difference (Broda Weinstein 2006 Weights)	5.69		
Cross-Market IV	2.77		
Nevo 2001		3.06	3.04
Hausman (1996)		2.17	2.31

NOTE.—This table reviews alternative measures of elasticity. The first column features pooled CES estimates using the sample of the current paper under three pre-existing methods; the heteroskedasticity-identified method using weights from Feenstra (1994) and Broda and Weinstein (2006) as well as a pooled cross-market regression using the same instrument as used for the translog case in this paper. The second two columns include median and mean for two prominent studies on the cereal market in the industrial organization literature. The center of the elasticity distribution in the current study most closely matches that from Nevo (2001).

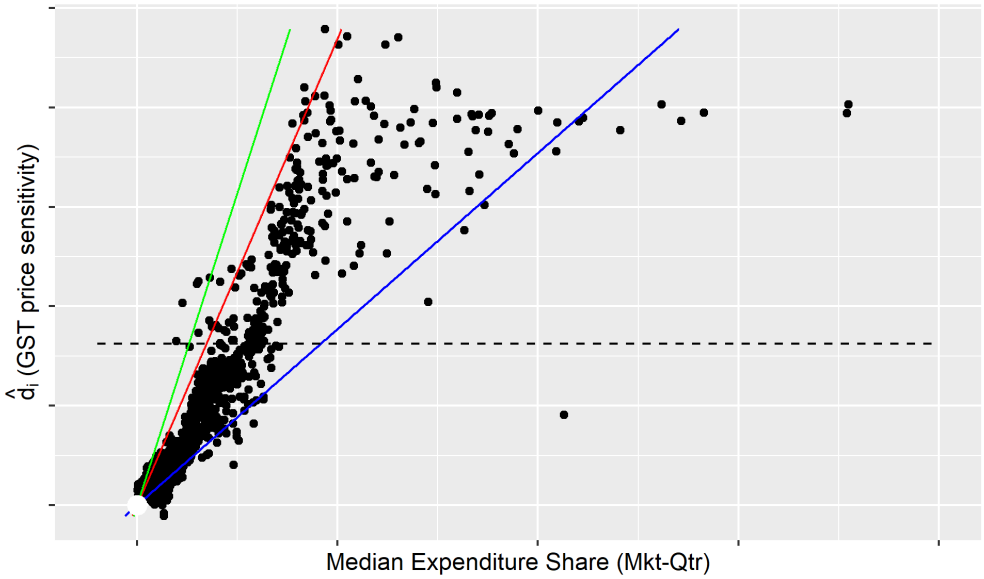
To compare the GSA translog elasticity to those from the table below, Figure 1.3 plots the estimated $\hat{d}_i$ values against each product's median expenditure share.[74] In such a graph, we estimate a good's elasticity based on the slope of the line connecting a given point to the origin, since we can rearrange the definition of the constant base-price elasticity from equation 1.30 as:

$$s_i(\eta_i - 1) = d_i \tag{1.46}$$

In Figure 1.3, the rays corresponding to the three pooled CES elasticity estimates are shown by the green (unweighted double difference), red (BW-weighted), and blue (pooled cross-market IV) lines. A prominent feature of the results is the strong relationship between d_i values and each product's (median) expenditure share, which allows the semi-elasticity to capture the scaling effect necessary to allow the elasticity to be similar across products of differing levels of popularity. For comparison, the estimated value for a regression which assumes all products have a single common semi-elasticity value is plotted in the dotted horizontal line.

[74]Median is taken over all quarters and markets that the product is available.

Figure 1.3: Semi-Elasticity vs. Expenditure Share



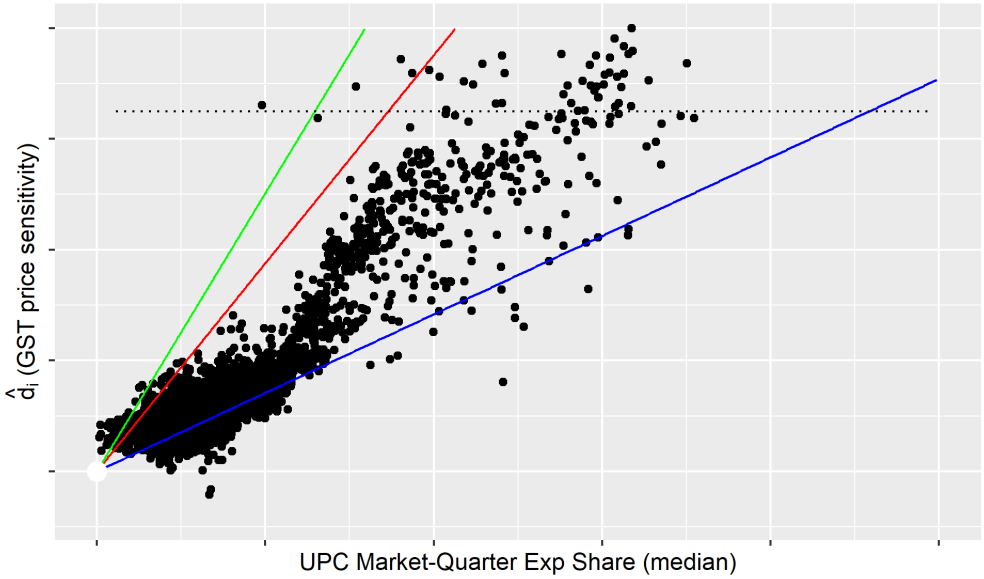
NOTE.—This figure shows the estimated product-specific semi-elasticities plotted against each good's median expenditure share (over all markets and quarters the good is available). Lines on the same ray have equivalent partial elasticities; rays corresponding to point estimates for different CES estimation techniques are plotted.

Looking at the median expenditure share-elasticity, the unweighted double-difference estimate is at the edge of the range seen in the GSA/GRF estimation, while the Broda-Weinstein weighting brings the elasticity closer to a range similar to those of the GSA estimates. Notably, despite the general trend of an increase in semi-elasticity value as product the share increases, the most popular products are estimated to have the same semi-elasticity values as lower-demanded products. This is likely an artifact of the sparsity of products in this region, so that GRF is unable to meaningfully separate these products and, roughly speaking, products are estimated to have a common semi-elasticity value.

While the upward trend in semi-elasticity values with increasing expenditure share is a prominent feature of the data, the more zoomed-out view makes it difficult to assess the extent of variation in d_i, and in turn elasticity, among the majority of low-expenditure share products. Zooming in on the low-expenditure share and low-d_i products, Figure 1.4 shows that there is substantial heterogeneity in the point estimates of the semi-

elasticities.

Figure 1.4: GSA Price Sensitivity vs. Expenditure Share (Low expenditure share)



NOTE.—This figure includes the same data as Figure 1.3, but with the axes adjusted to zoom in on the mass of low-expenditure share products. For this mass of products, it is clear to see that there is substantial variation in the elasticity, as indicated by products with a similar semi-elasticity (d_i point estimate) having very different expenditure shares.

To gain some intuition for what is driving the patterns GRF is finding in the data, we can observe how products with similar median expenditure shares differ along other partitioning variable dimensions. This can be done either by looking at the cross-tabulations of the point estimates or by tracing out the $\hat{d}(x_i)$ function as we vary the value the partitioning variables. Using shading to indicate different dimensions of the data, we can see that product price and the brand (rather than UPC-level) appear to be related to shifts in the price sensitivity. Figure 1.5 shows that UPCs within a popular national brand tend to have more elastic demand than other products at the same level of expenditure. This has some intuitive appeal, as this suggests that the gains from adding an individual product into a large brand portolio are smaller than an equally popular but unique product. Figure 1.6 shows that the lowest-elasticity products (below the blue line) tend to be expensive relative to products with a similar expenditure share (while from Figure 1.5 we know these products are not varieties of a large brand). This suggests that these products may be examples of an expensive but niche market segment

which has higher appeal for its customers.

Figure 1.5: GSA Price Sensitivity and National Brand Share

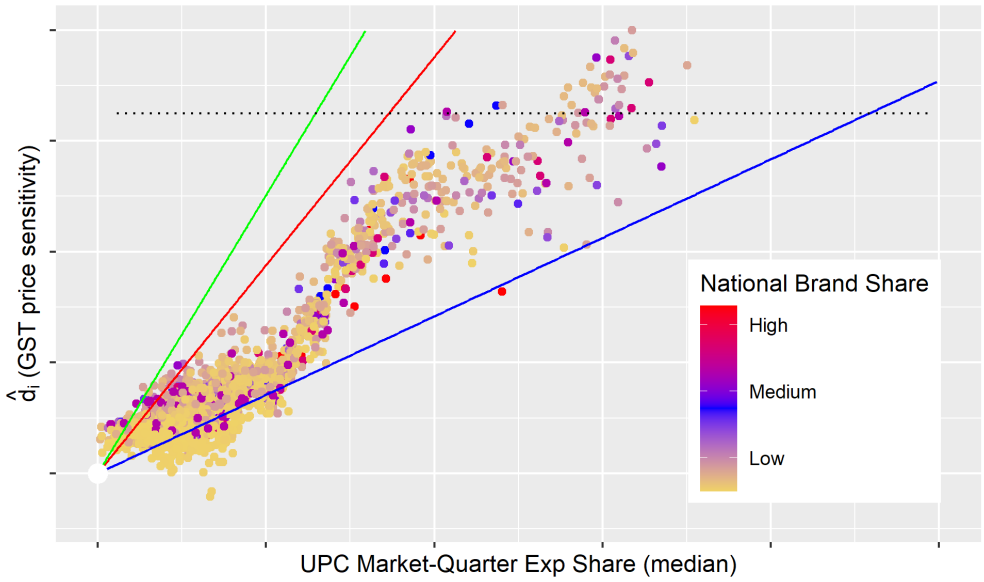
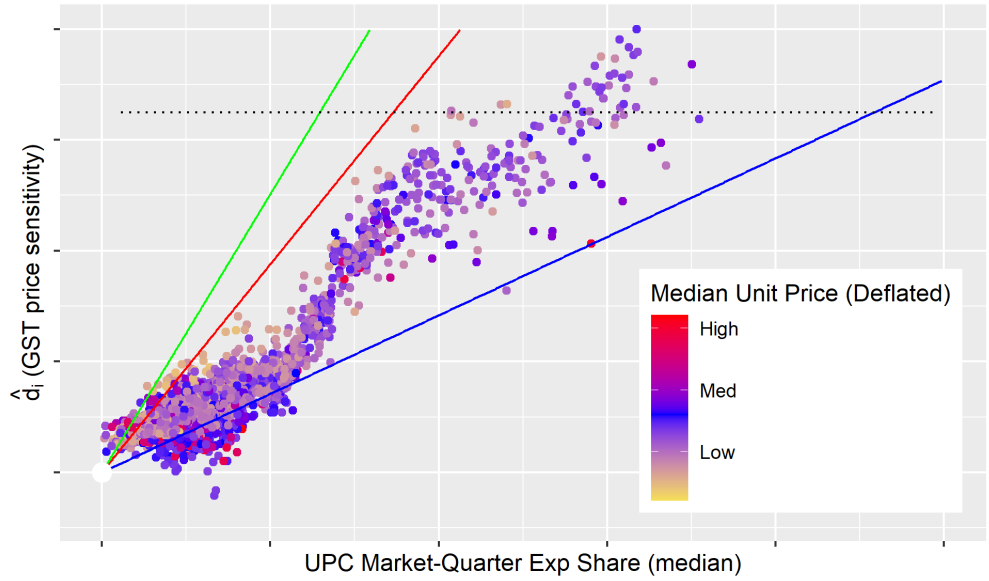


Figure 1.6: GSA Price Sensitivity and Unit Price

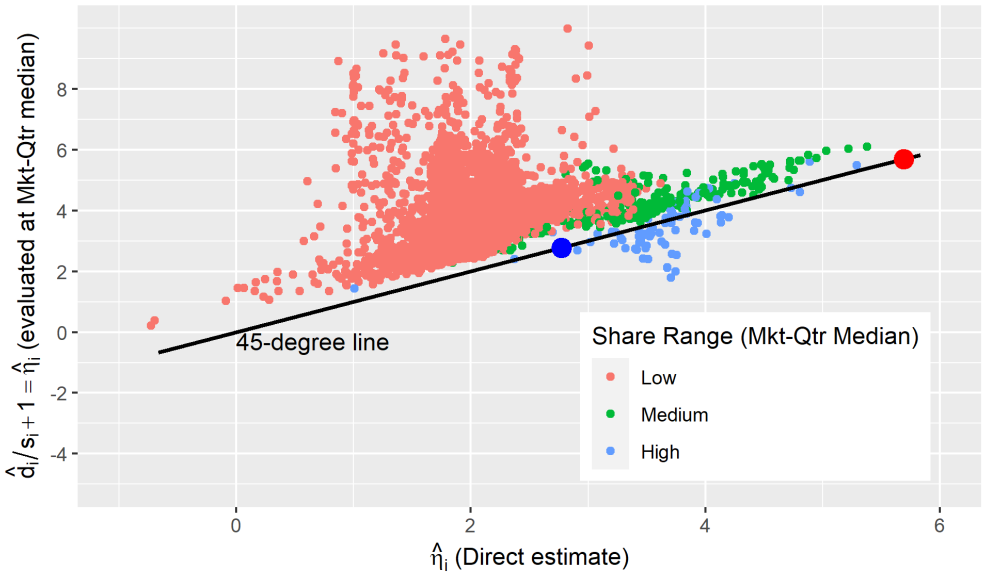


NOTE.—This figure plots semi-elasticities against expenditure shares, using shading to indicate unit prices. Unlike the implication of combining symmetric demand curves and Marshall's second law, the pattern that emerges is that price is only weakly related to elasticity, and it is very common to see expensive products with relatively low elasticity (i.e. they sit on a lower ray).

As noted earlier, the GRF algorithm roughly fits a common d_i value for the most popular

products in turn implying that the most popular products have a low elasticity. This is likely due to the GRF algorithm's difficulty extending the upward trend in d_i values in this sparsely popular region of products. Support for this interpretation of the result comes from estimating product elasticities directly using GRF, rather than estimating semi-elasticities.[75] To the extent that elasticities are similar across different groups of products, the direct estimate does not have to group products of similar expenditure shares to uncover this result. Figure 8 compares the median-share GSA translog elasticity to this direct-estimation elasticity. The GSA elasticity is typically above the direct estimate, except for the group of high expenditure share products which dip below the 45 degree line.

Figure 1.7: GSA Elasticity vs. Directly-Estimated Elasticity (GRF)



NOTE.—This figure compares the elasticity using the GSA translog method (elasticities evaluated at median expenditure share) to the point estimates for elasticity estimated using the Generalized Random Forest procedure (i.e. using $\Delta \ln s_i$ as the dependent variable instead of Δs_i). Except for the most popular products, directly estimating the elasticity finds relatively inelastic demand. One possible explanation for this is greater attenuation bias due to higher error-term variance when taking logs for low-expenditure share products.

Given that bandwidth issues likely explain the semi-elasticity values for the upper tail of

[75]That is, the GRF procedure is run with a left-hand side variable of $\Delta \ln(s_{itr})$ instead of Δs_{itr}. The right-hand side variables of $\Delta \ln(p_{itr})$ and the time-region dummy are used in both cases. All the partitioning variables are kept the same.

the expenditure share distribution, it seems improper to use the estimated d_i values for the most highly-demanded products. In the results presented below, I instead use the direct-estimation elasticity (i.e. from a GRF estimation where $\Delta \ln s_{itr}$ is the left-hand side variable) to impute a value for d_i among the highly-demanded products. Specifically, I use $\tilde{d_i} = s_i(\hat{\eta}_i - 1)$ where s_i is the median region-quarter expenditure share and $\hat{\eta}_i$ is the direct-estimation of the elasticity using GRF. Graphically, this corresponds to adjusting the d_i values so that the light-blue dots in Figure 8 are shifted back towards the 45-degree line.

The large blue and red dots along the 45-degree line in Figure 1.7 denote the pooled cross-market IV and the Broda Weinstein-weighted double difference estimates, respectively, of a single common CES elasticity. The GRF direct estimation elasticities are, not surprisingly, clustered around the cross-market IV (the blue dot), albeit with a wide range. The lowest point estimate is -0.73 while the highest is 5.4, with the middle 50% of products having point estimates for their elasticity between 1.8 and 2.3.[76] Notably, the Broda Weinstein-weighted double difference (the red dot) is outside the range of elasticities calculated using the direct GRF approach.

While the median-share elasticity is a useful benchmark, in practice with a GSA demand system each product's elasticity will change over time as its expenditure share changes. Evaluating the elasticity using national expenditure shares and the estimated $\hat{d}_i$ values[77], figure 9 shows a broad distribution of elasticity estimates taken over all product-quarters. Within the sample, about 15% of product-quarters are below the pooled cross-market IV estimate (the blue line), 35% of product-quarters are above the Broda Weinstein-weighted double difference, while the remaining 50% of product-quarters are between these two levels. Overall, the median product-quarter elasticity is 4.8 (shown by the black dashed line). Due to a long tail of low expenditure share items at the national level (and, in turn, high elasticity) the average is pulled higher to 7.3.

Given the estimates for the semi-elasticity, we can now turn to the issue of the gains from entry and exit. Figure 1.8 shows the cumulative change in overall price index measurements using different entry/exit adjustment formulas and calibrations. If we only used the standard Sato-Vartia continuing goods index, as shown in the black line, we would estimate a modest increase in the cost of living of 13.1% between 2004 and 2016 (about 1.03% per year).[78]

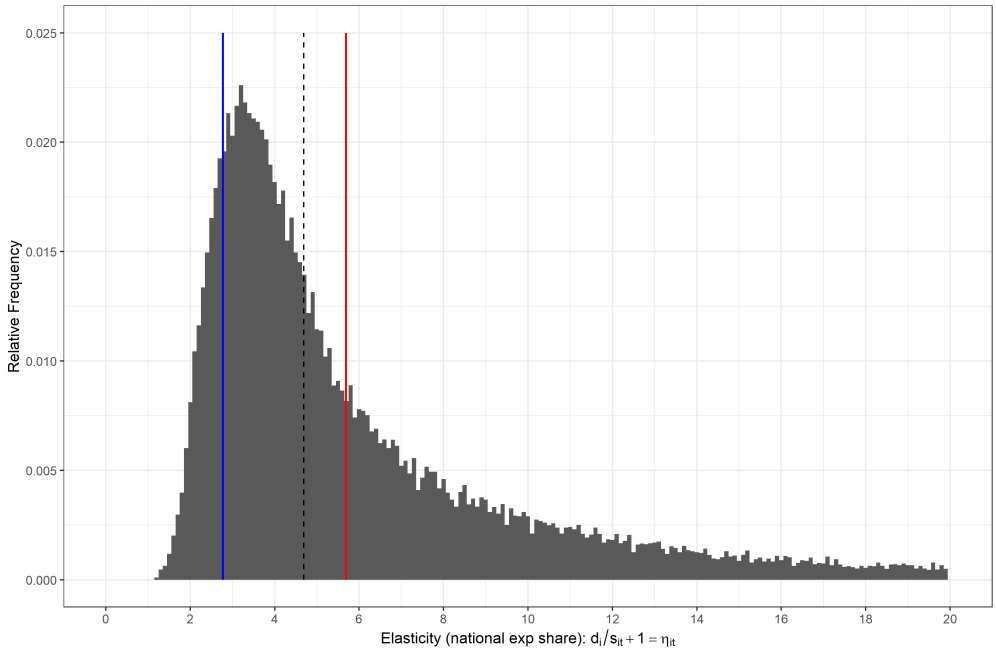
---

[76] A broad literature in treatment effect estimation notes that the point estimate of a pooled (common slope) regression in the presence of heterogeneity is akin to a weighted average of the heterogenous slope terms.

[77] Including the adjustment in d_i values for the most highly-demanded products.

[78] All price index calculations are constructed as chained quarterly indices. The full set of products is

Figure 1.8: Elasticity Distribution: Product-Quarters (National Expenditure Share)



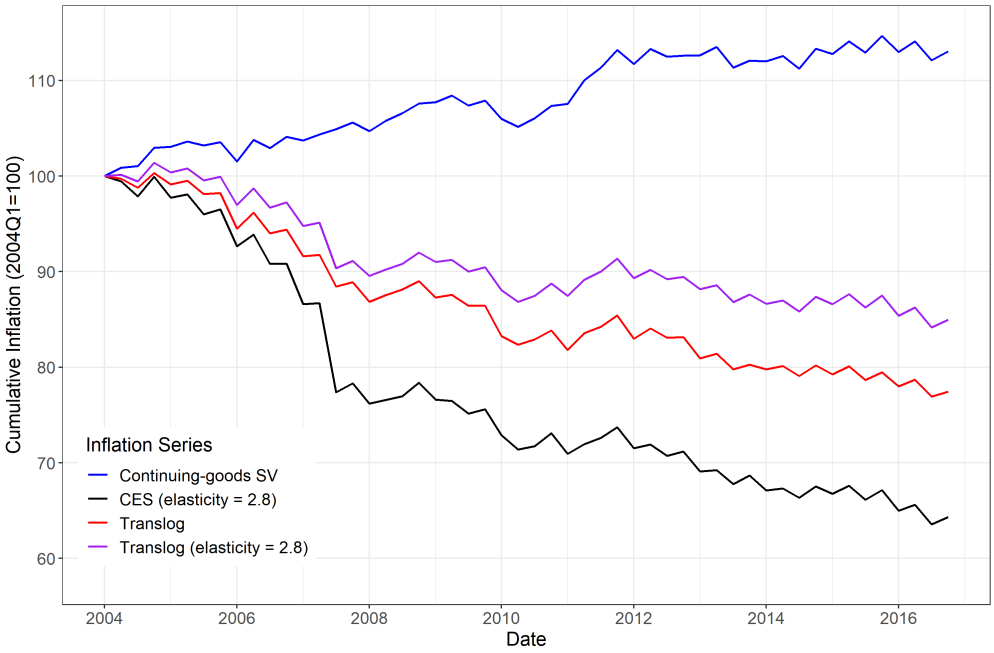
NOTE.—This figure plots the full distribution of product elasticities over all quarters, using national expenditure shares. The mass of products lies between the standard CES estimates, more elastic than the IV-based CES estimate but inelastic compared to the heteroskedasticity-based point estimates.

All of the entry/exit adjustments considered in Figure 1.9 reflect an estimated decline in the expenditure function. The magnitude of this decline is substantially affected by the demand system estimates, however. Using the weighted double-difference estimate for the CES elasticity leads to an estimated decline in the cost of living of -8.6% while using the pooled cross-market IV estimate corresponds to a -35.7% decline in the cost of living. The translog-based adjustment lies between these two extremes, with an estimated decline of -22.3%. For reference, Figure 1.9 also includes an alternative GSA price index where the elasticities are pinned at 2.8 for all products and time periods, as considered in the comparison of equations (1.31) and equation (1.25).

The translog price index calibrated with the product-specific semi-elasticity values from

considered a single group for both the CES and the GSA translog price indices. The behaviour of the continuing goods component of the GSA translog index is similar to that of the continuing goods Sato-Vartia index, with an approximately 12.1% increase in this index over the twelve years of the sample (or 0.95% per year).

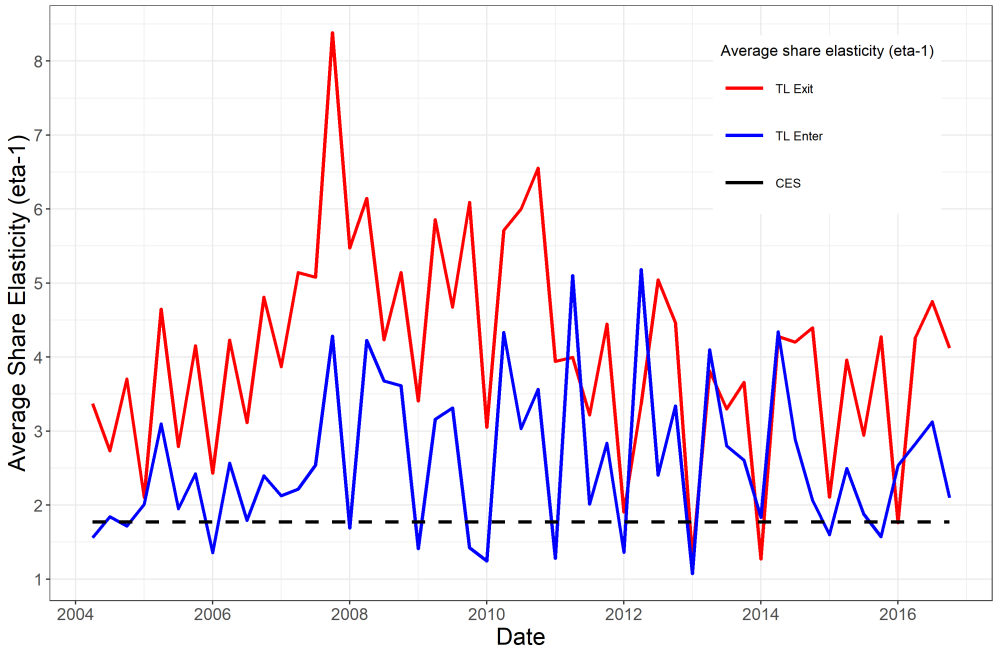
Figure 1.9: Cumulative Inflation with Entry/Exit Adjustment



NOTE.—This figure plot cumulative inflation using a standard continuing goods index, along with different entry/exit adjustments. The path of inflation using the new method, with product-specific and time-varying elasticities lies between the CES with the IV-style point estimate for the elasticity and a translog-style formula that (counterfactually) assumes a constant-and-common elasticity at the same level.

the random forest estimations values yields an even larger adjustment than this benchmark, despite the fact that the translog / random forest calibration finds most product-periods are more elastically demanded than the CES point estimate suggests. Overall, this reflects an imbalance of entry and exit between low- and high-elasticity products especially in the early part of the sample period. Figure 1.10 calculates the average share-elasticity for each quarter using the random forest calibration. In almost all quarters, the average elasticity is above the CES point estimate of 2.8. However, the elasticity for entering goods is systematically lower than the elasticity for exiting goods, so that the gains from entry fall less than the losses from exit. This asymmetry leads to a boost in the net gain over time from product entry and exit.

Figure 1.10: Entry and Exit Average Elasticity



NOTE.—This figure plots the welfare-relevant average partial elasticities for entering and exiting goods using the product-specific and time-varying elasticities, compared to the CES point estimate from the IV regression. Although translog-based elasticities are systematically higher (more elastic), there is also a novel asymmetry: entering goods (on entry) are relatively inelastic compared to exiting goods (on exit).

1.9 Conclusion

This paper has introduced the group secondary aggregate (GSA) translog functional form, which allows for arbitrary own-price effects and embeds a finite reservation price, together with an associated exact entry/exit adjusted price index and an estimation strategy which imposes little ex-ante restriction on the relevant demand parameters. While in general substitution effects depend on the set of available products, I show that the relevant demand parameter in GSA translog is invariant to product entry and exit. The GSA structure allows goods to be substitutes or complements both within- and between groups, without requiring additional demand parameters to be estimated or necessitating a non-linear transformation of the demand system. The estimation strategy builds on the generalized random forest (GRF) method of Athey et al. (2019) by incorporating a time-region fixed effect suited to the panel data.

In the empirical application to the ready-to-eat cereal market, I document a wide dispersion in product elasticities under the GSA calibration and also the spread in elasticity estimates based on different approaches to estimating the CES demand. While translog demand yields about half the entry/exit effects suggested by CES *for a given level of elasticity*, the overall estimated gains also depend crucially on the product-by-product elasticity estimates. In the empirical application, incorporating heterogenous price effects tended to raise the raise the net gains from entry and exit. This was due, primarily, to an important asymmetry between entering and exiting goods that is inadmissible with CES: entering goods are inelastically demanded (on entry) relative to exiting goods (on exit). This asymmetry partially offsets the effects of the higher average elasticity level and the shift from unbounded (CES) to finite (translog) reservation prices. In subsequent work, I intend to investigate the extent to which this pattern is present in other consumption categories in Nielsen as well as in other data sets.

In principle, the methods proposed in this paper can supplement hedonic adjustment methods used by statistical agencies. First, GRF estimation could be used to estimate product prices each period, providing a non-parametric approach to standard hedonic adjustment.[79] Second, product characteristics can be used in combination with a demand system specification when estimating each good's elasticity in the context of the GSA translog price index. In this case, an important advantage of GRF relative to other estimation techniques is that it provides a basis for calibrating own-price elasticities for products outside the initial training sample. While choosing a model specification and constructing a forest may be time-intensive, once a model is specified estimating treatment parameters for new test cases can be done quickly. This is a potential important consideration given the time constraints involved in preparing official statistics.[80]

While this paper focuses on the measurement of the entry/exit adjusted price index, it also highlights new mechanisms that are excluded by the CES benchmark and case

[79]Standard regression tree or random forest algorithms could also serve this purpose.

[80]See, for example, the discussion in Pakes (2003).

studies that focus on fixed samples of goods. In particular, the asymmetry between entering and exiting goods that I find calls for a more systematic explanation. One explanation is a product-cycle interpretation, where new goods have a high relative price when they enter but by the time they exit they have been superseded by other varieties and have been pushed to more elastic segments of their demand curve. An alternative explanation for this pattern could be the different ways the CES and translog sufficient statistics interpret expenditure share fluctuations due to taste shocks.[81] In principle consumers could have a "novelty aversion" where goods have lower taste-values on entry relative to their long-run pattern or a "love of novelty" where there is a spike in demand on entry but then goods settle into a more modest level of popularity. Finally, given the flexibility in own-price effects that my specification allows for, it may be that goods are entering and exiting from product segments with different elasticities. Accounting for the relative importance of these forces, and embedding them into a model for which products enter and exit, is left for future work.

[81]The standard theory for price indices used in this paper must be adapted to accommodate taste shocks. Redding and Weinstein (2020) introduced a method that treats taste shocks as analogous to changes in quality or price mismeasurement. Baqaee and Burstein (2021) addresses taste shocks by evaluating the cost of living for a stable set of preferences, which requires an adjustment to expenditure shares away from the point where the preference specification is fixed; analytically the treatment of taste shocks appears similar to the treatment of non-homotheticities.

APPENDIX

1.A Appendix: Proofs

1.A.1 Proposition 1 (observed and reservation prices)

Partition the set of goods into quantities of observed goods $\boldsymbol{q}_o$ and missing goods $\boldsymbol{q}_m$ so that the utility function may be written as:

$$U(\boldsymbol{q}) = U(\boldsymbol{q}_o, \boldsymbol{q}_m)$$

By definition, the availability constrained demand system satisfies the first order conditions:

$$\nabla_{\boldsymbol{q}_o} U(\boldsymbol{q}_o, 0) = \lambda_c \boldsymbol{p}_o \implies \left(\frac{1}{\lambda_c}\right) \nabla_{\boldsymbol{q}_o} U(\boldsymbol{q}_o, 0) = \boldsymbol{p}_o$$

where λ_c is the Lagrange multiplier for the availability constrained problem.

In turn, the virtual prices $\boldsymbol{p}^*$ that rationalizes the consumption bundle $(\boldsymbol{q}_o, 0)$ at nominal income Y must satisfy:

$$\nabla_{\boldsymbol{q}} U(\boldsymbol{q}_o, 0) = \begin{bmatrix} \nabla_{\boldsymbol{q}_0} U(\boldsymbol{q}_o, 0) \\ \nabla_{\boldsymbol{q}_m} U(\boldsymbol{q}_o, 0) \end{bmatrix} = \lambda_a \boldsymbol{p}^*$$

where $\boldsymbol{p}_o \cdot \boldsymbol{q}_o = Y$. To satisfy the first order conditions, it must be that $\lambda_a = \lambda_c$ and $\boldsymbol{p}_o^* = \boldsymbol{p}_o$ while the remaining virtual prices are given by:

$$p_m^* = \left(\frac{1}{\lambda_a}\right) \nabla_{\boldsymbol{q}_m} U(\boldsymbol{q}_o, 0)$$

Given a fixed preference and the choice of the set Ω_o, λ_a and q_o are pinned down by the observed prices and the level of nominal income. Thus, we may write p_m^* as a function of p_o, Y, and Ω_o.

1.A.2 Proposition 2 (availability-constrained price index)

As already noted, the change in the cost of living is given by:

$$\Delta COL(t_o, t_1, \bar{V}) = \ln E^*(p^*_{t_1}, \bar{V}) - \ln E^*(p^*_{t_0}, \bar{V})$$

Since quantity choices are optimized in the expenditure minimization problem, Shepard's lemma implies

$$\nabla_{\ln \boldsymbol{p}^*} \ln E^* = \left(\frac{\hat{\boldsymbol{p}}^*}{E^*}\right) \nabla_{\boldsymbol{p}^*} E^* = \boldsymbol{s}^*(\boldsymbol{p}^*)$$

In turn, using the line integral, we have:

$$\Delta COL(t_o, t_1) = \Delta \ln E^*(t_o, t_1) = \int_{t_0}^{t_1} \nabla_{\ln \boldsymbol{p}^*} \ln E^* \cdot \ln \boldsymbol{p}^* = \int_{t_0}^{t_1} \boldsymbol{s}^* \cdot \ln \boldsymbol{p}^*$$

1.A.3 Proposition 3 (GSA and Cross-Price Effects)

A GSA demand system for $i \in g$ is defined as

$$s_i = f_i(p_i / A_{g(i)}(\boldsymbol{p}))$$

or, rearranging to get quantities we have:

$$q_i = \left(\frac{Y}{p_i}\right) f_i(p_i / A_g(\boldsymbol{p}))$$

The sign of the gross (Marshallian) cross-price elasticities matches the sign of the off-diagonal partial derivatives which can be evaluated as

$$\frac{\partial \ln q_i}{\partial \ln p_j} = \frac{\partial \ln s_i}{\partial \ln p_j} = (\eta_i - 1)\frac{\partial \ln A_g}{\partial \ln p_j}$$

When all elasticities are greater than 1, then the sign of the cross-price elasticity is equal to the sign of the elasticity of group aggregator with respect to the price of product j.

Although the elasticity matrix is not itself symmetric, it differs from a symmetric matrix only by multiplication with a diagonal matrix with all positive values, specifically:

$$J\ln \boldsymbol{p} \ln \boldsymbol{s} = \hat{s}^{-1} J\ln \boldsymbol{p}\boldsymbol{s} = H_{\ln \boldsymbol{p}} \ln E$$

Thus, for $j \in g'$ the sign of $\partial \ln A_{g'} / \partial \ln p_i$ must match the sign of $\partial \ln A_g / \partial \ln p_j$. In turn, the sign of $\partial \ln A_{g'} \partial \ln p_i$ is common to all $j' \in g'$ by definition of group membership so that, again invoking symmetry, the sign of $\partial \ln A_g / \partial \ln p_{j'}$ must match the sign of $\partial A_g / \partial \ln j$. This completes the proof of the second part of Proposition 3.

The first part of proposition 3 comes from the adding up conditions on elasticities. Specifically, when there is only a single group with aggregator A the following relationship holds:

$$d \ln s_i = -(\eta_i - 1) \left[d \ln p_i - d \ln A \right]$$

In turn, multiplying both sides by s_i and adding up over all goods, we have:

$$0 = d \ln A \left[\sum_i s_i (\eta_i - 1) \right] - \sum_i s_i (\eta_i - 1) d \ln p_i$$

and, rearranging we have:

$$d \ln A = \left(\frac{1}{\sum_i s_i (\eta_i - 1)} \right) \sum_i s_i (\eta_i - 1) d \ln p_i$$

When all products have an elasticity greater than 1, then $\partial \ln A / \partial \ln p_i$ is positive for all i meaning all products are gross substitutes.

1.A.4 Proposition 4 (GSA Entry/Exit Adjusted Price Index)

Follows directly from discussion in text.

1.A.5 Proposition 5 (Translog - Purchased Goods Representation)

This result is a restatement of an intermediate result given in Feenstra (2003).

The expenditure share functions corresponding to missing goods are given by

$$0 = \boldsymbol{s}_m = \alpha_m + \boldsymbol{\Gamma}_{mo} \ln \boldsymbol{p}_o + \boldsymbol{\Gamma}_{mm} \ln \boldsymbol{p}_m^*$$

Rearranging to solve for $\ln \boldsymbol{p}_m^*$, we have:

$$\ln \boldsymbol{p}_m^* = -\boldsymbol{\Gamma}_{mm}^{-1} \alpha_m - \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\Gamma}_{mo} \ln \boldsymbol{p}_o$$

Expanding out the interaction term in the definition of the homothetic translog expenditure function, and noting that $\boldsymbol{\Gamma}$ is symmetric, we have:

$$\ln \boldsymbol{p}' \boldsymbol{\Gamma} \ln \boldsymbol{p} = \ln \boldsymbol{p}' \boldsymbol{\Gamma}_{oo} \ln \boldsymbol{p} + 2 \ln \boldsymbol{p}_m' \boldsymbol{\Gamma}_{mo} \ln \boldsymbol{p}_o + \ln \boldsymbol{p}_m' \boldsymbol{\Gamma}_{mm} \ln \boldsymbol{p}_m$$

Plugging in the solution for $\ln \boldsymbol{p}_m^*$, and collapsing terms we have

$$\ln \boldsymbol{p}' \boldsymbol{\Gamma} \ln \boldsymbol{p} = \ln \boldsymbol{p}_o' \boldsymbol{\Gamma}_{oo} \ln \boldsymbol{p}_o - \ln \boldsymbol{p}_o \boldsymbol{\Gamma}_{om} \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\Gamma}_{mo} \ln \boldsymbol{p}_o + \boldsymbol{\alpha}_m' \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\alpha}_m$$

In addition, for the $\boldsymbol{\alpha}' \ln \boldsymbol{p}$ term making the substitution gives us:

$$\boldsymbol{\alpha}' \ln \boldsymbol{p} = \boldsymbol{\alpha}_o' \ln \boldsymbol{p}_o - \alpha_m \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\alpha}_m - \alpha_m' \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\Gamma}_{om} \ln \boldsymbol{p}_o$$

Putting this all together, we have:

$$\begin{aligned} \ln E(\boldsymbol{p}_o, \boldsymbol{p}^*) &= \alpha_0 + \boldsymbol{\alpha}' \ln \boldsymbol{p} + \frac{1}{2} \ln \boldsymbol{p}' \boldsymbol{\Gamma} \ln \boldsymbol{p} \\ &= \alpha_0 - \left(\frac{1}{2}\right) \boldsymbol{\alpha}_m' \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\alpha}_m + \left[\boldsymbol{\alpha}_o - \boldsymbol{\Gamma}_{om} \boldsymbol{\Gamma}_{mm}^{-1} \boldsymbol{\alpha}_m\right]' \ln \boldsymbol{p}_o + \frac{1}{2} \ln \boldsymbol{p}' \left[\boldsymbol{\Gamma}_{oo} - \boldsymbol{\Gamma}_{om} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_{mo}\right] \ln \boldsymbol{p} \end{aligned}$$

which matches the definitions of the $\tilde{\alpha}_0$, $\tilde{\boldsymbol{\alpha}}$, and $\tilde{\boldsymbol{\Gamma}}$ terms in the Proposition.

1.A.6 Proposition 6 (Stability of Partial Semi-Elasticity)

The decomposition of the $\mathbf{\Gamma}$ matrix afforded by the GSA structure implies that $\mathbf{\Gamma}_{oo} = -\hat{\boldsymbol{d}}_o + \hat{d}_o G_o B_o$ and $\mathbf{\Gamma}_{om} = \hat{d}_o G_o B_m$.

In turn, applying this decomposition to the definitions in Proposition 5, we have:

$$\begin{aligned}
\tilde{\mathbf{\Gamma}} &= \mathbf{\Gamma}_{oo} - \mathbf{\Gamma}_{om}\mathbf{\Gamma}_{mm}^{-1}\mathbf{\Gamma}_{mo} \\
&= -\hat{\boldsymbol{d}}_o + \hat{\boldsymbol{d}}_o \boldsymbol{G}_o \boldsymbol{B}_o - \hat{\boldsymbol{d}}_o \boldsymbol{G}_o \boldsymbol{B}_m \mathbf{\Gamma}_{mm}^{-1}\mathbf{\Gamma}_{mo} \\
&= -\hat{\boldsymbol{d}}_o \left[I + \boldsymbol{G}_o \underbrace{\left(\boldsymbol{B}_o \boldsymbol{B}_m \mathbf{\Gamma}_{mm}^{-1}\mathbf{\Gamma}_{mo} \right)}_{\tilde{\boldsymbol{B}}_o} \right]
\end{aligned}$$

1.A.7 Proposition 7 (GSA Translog Entry/Exit Adjusted Price Index)

It is will know that the Tornqvist index is exact for a homothetic translog demand system, so we have:

$$\Delta \ln E = \sum_{i \in \Omega} \left[\frac{s_{it_o} + s_{it_1}}{2} \right] \Delta \ln p_i^* = \sum_{i \in \Omega} \bar{s}_i \Delta \ln p_i^*$$

To recover the missing price changes, we can invert the demand system, which given a stable preference (i.e. constant α_i and d_i in the all-goods representation) gives us:

$$\Delta \ln p_i^* = \Delta \ln A_{g(i)} - \frac{\Delta s_i}{d_i}$$

Using this substitution only for entering and exiting goods, we have:

$$\Delta \ln E = \sum_{i \in c} \bar{s}_i \Delta \ln p_i + \sum_g (\bar{s}_g - \bar{s}_{cg}) \Delta \ln A_g - \sum_{i \notin c} \frac{\Delta s_i \bar{s}_i}{d_i}$$

In turn, we can recover ΔA_g terms by using the behavior of continuing goods within

each group. One way to do this to note that for each product separately we have:

$$\Delta \ln A_{g(i)} = \Delta \ln p_i + \frac{\Delta s_i}{d_i}$$

Using expenditure shares to average over all continuing goods within a group, we have:

$$\Delta \ln A_g = \bar{s}_{cg}^{-1} \left[\sum_{i \in g \cap c} \bar{s}_i \Delta \ln p_i - \frac{\bar{s}_i \Delta s_i}{d_i} \right]$$

Plugging this into the earlier expression, we have:

$$\Delta \ln E = \sum_{i \in c} \bar{s}_i \Delta \ln p_i + \sum_g \left(\frac{\bar{s}_g}{\bar{s}_{cg}} - 1 \right) \left[\sum_{i \in g \cap c} \bar{s}_i \Delta \ln p_i + \frac{\bar{s}_i \Delta s_i}{d_i} \right] - \sum_{i \notin c} \frac{\Delta s_i \bar{s}_i}{d_i}$$

Finally, rearranging we get the result from the proposition:

$$\Delta \ln E = \sum_g \bar{s}_g \sum_{i \in g \cap c} \frac{\bar{s}_i}{\bar{s}_{cg}} \Delta \ln p_i + \sum_g \left(\frac{\bar{s}_g}{\bar{s}_{cg}} - 1 \right) \sum_{i \in g \cap c} \frac{\bar{s}_i \Delta s_i}{d_i} - \sum_{i \notin c} \frac{\Delta s_i \bar{s}_i}{d_i}$$

1.B Appendix: Discussion of CES and Logit

Constant Elasticity of Substitution

Constant elasticity of substitution demand may be written as:

$$\ln s_i = \alpha_i - (\sigma - 1) \left[\ln p_i - \ln P \right]$$

where P is the CES price index. In turn, taking differences to eliminate α_i we have:

$$\Delta \ln s_i = -(\sigma - 1) \left[\Delta \ln p_i - \Delta \ln P \right]$$

First, we can make a quick proof that the Sato-Vartia index is exact for the CES. Note that the Sato-Vartia (un-normalized) weights $w_i = \Delta s_i / \Delta \ln s_i$. Thus, applying this weight

and adding up over all goods, we have:

$$\begin{aligned}
\sum_i w_i \Delta \ln s_i &= -(\sigma - 1) w_i \left[\Delta \ln p_i - \Delta \ln P\right] \\
0 &= -(\sigma - 1) \sum_i w_i \Delta \ln p_i + (\sigma - 1) \Delta \ln P \left[\sum_i w_i\right] \\
\Delta \ln P &= \left(\frac{1}{\sum_i w_i}\right)^{-1} \left[\sum_i w_i \Delta \ln p_i\right]
\end{aligned}$$

where the second line notes that $\sum_i \Delta s_i = 0$ when the sum is taken over all goods.

This formula doesn't work, however, when products are entering and exiting. First $\Delta \ln s_i$ is not defined when one of the end-points is zero. Second, we do not observe a price change and the model-consistent price when the product is unavailable is infinite. Thus it doesn't make sense to use Sato-Vartia weights or the expression in changes for entering and exiting goods. In this case, we have:

$$\begin{aligned}
\sum_{i \in c} w_i \Delta \ln s_i &= -(\sigma - 1) \sum_{i \in c} w_i \left[\Delta \ln p_i - \Delta \ln P\right] \\
\frac{-\Delta s_c}{\sigma - 1} &= \sum_{i \in c} w_i \Delta \ln p_i - \Delta \ln P \left(\sum_{i \in c} w_i\right) \\
\Delta \ln P &= \left(\sum_{i \in c}\right)^{-1} \left[\sum_{iinc} w_i \Delta \ln p_i + \frac{\Delta s_c}{\sigma - 1}\right]
\end{aligned}$$

This matches the first line of equation (1.25). To rearrange to the second line, may note that $\Delta s_c = (1 - s_{nt_1}) - (1 - s_{xt_0}) = s_{xt_0} - s_{nt_1}$.

The Feenstra (1994) price index is instead:

$$\Delta \ln P = \sum_{i \in c} w_i^c \Delta \ln p_i + \frac{\Delta \ln s_c}{\sigma - 1}$$

where $w_i^c = \Delta s_i^c / \Delta \ln s_i^c$ where $s_{it}^c = s_{it} / s_{ct}$ is the share of product i among continuing goods in each period.

The Feenstra (1994) version of the CES price index corresponds to evaluating the integral for the exact price index in three steps. First, letting all exiting products experience a price increase to infinity while holding all other prices fixed. Second, letting all continuing goods experience their observed price changes. Then, third, letting all entering goods experience a decline in price from infinity while holding all other prices fixed. Given this path of price changes, the CES price index takes the form:

$$\Delta \ln P = \int_{t_0}^{t_x} \boldsymbol{s}_x \cdot d \ln \boldsymbol{p}_x + \int_{t_x}^{t_n} \boldsymbol{s}_c \cdot d \ln \boldsymbol{p}_c + \int_{t_n}^{t_1} \boldsymbol{s}_n \cdot d \ln \boldsymbol{p}_n$$

For exiting and entering goods these are indefinite integrals since the limit point is unbounded. However, as long as $\sigma > 1$ the indefinite integral is still well defined even as price rise to infinity.

Taking the exiting good leg of the integral, we may use a change of variables for the integrand by noting that

$$d \ln p_i = \frac{-d \ln s_i}{\sigma - 1} + d \ln P$$

In turn, plugging this into the integral for the t_0 to t_x interval, we have:

$$d \ln P = \sum_{i \in x} s_i d \ln p_i = \sum_{i \in x} \frac{-ds_i}{\sigma - 1} + d \ln P \sum_{i \in x} s_i$$

In turn, this implies that in the t_0 to t_x interval we have:

$$d \ln P = \frac{1}{\sigma - 1} \sum_{i \in x} \frac{-1}{1 - \sum_{i \in x} s_i} ds_i$$

The indefinite integral for the right-hand side is given by:

$$\frac{1}{\sigma - 1} \int \sum_{i \in x} \frac{-1}{1 - \sum_{i \in x} s_i} ds_i = \left(\frac{1}{\sigma - 1} \right) \ln \left(1 - \sum_{i \in x} s_i \right)$$

Given the definitions of the end-points evaluating this integral yields:

$$\int_{t_0}^{t_1} \boldsymbol{s}_x \cdot d \ln \boldsymbol{p}_x = \left(\frac{1}{\sigma - 1}\right) [\ln(1) - \ln(1 - s_{xt_0})] = \frac{-\ln s_{ct_0}}{\sigma - 1}$$

Since $s_{ct_0} < 1$ and $\sigma > 1$ this term is positive, corresponding to the intuition that prices rise to generate exit. The analogous result holds for product entry, except without the minus sign.

Finally, to calculate the continuing goods component we can note that by definition with a CES demand the effect of price increases for one good leads to an equi-proportional increase in expenditure shares for all other goods (this is the IIA property). Thus, the expenditure share vectors at the end points for t_x and t_n are exactly the shares of each product *among continuing goods* at the initial and final time periods. Thus, we can evaluate the t_x to t_n component using a Sato-Vartia index (which is excct for CES) using the appropriately updated expenditure shares.

Logit Demand System (Representative Agent)

A logit demand system corresponds to quantity shares given by the expression

$$\pi_i = \frac{q_i}{\sum_j q_j} = \frac{e^{\alpha_i - \beta p_i}}{\sum_j e^{\alpha_j - \beta p_j}}$$

for $i \in 1, ..., N$. Using M to denote the number of units sold for all products $1, ..., N$, and defining an aggregator $A = \frac{-1}{\beta} \ln \left(\sum_j e^{\alpha_j - \beta p_j}\right)$ we may rewrite the demand equation as:

$$q_i = M\pi_i = Me^{\alpha_i - \beta[p_i - A]}$$

Thus, in the logit case we can define an analog to the partial elasticity for the homothetic GSA class where, again, we define the partial elasticity as the effect of raising a product's own price while holding all aggregators constant. This gives the logit partial elasticity

as:

$$\eta_i = -\frac{q_i}{p_i}\frac{\partial q_i(p_i, A)}{\partial p_i} = \frac{\beta}{p_i}$$

The standard practice in the discrete choice literature is to calculate an expected consumer surplus as proposed in Small and Rosen (1981). Given the logit case, this evaluates to:

$$\Delta\mathbb{E}\left[CS\right] = -\int \sum_i \pi_i dp_i = -\frac{1}{M}\int \sum_i q_i dp_i = \Delta\frac{1}{\beta}\ln\left(\sum_i e^{\alpha_i - \beta p_i}\right) = -\Delta A$$

Alternatively, we can rationalize the logit demand system using a representative agent with the following expenditure function (which has a quasi-linear form with numeraire good q_0) of the form

$$E(\boldsymbol{p}, U) = p_0\left[U - \frac{M}{\beta}\ln\left(\sum_i e^{\alpha_i - \beta(p_i/p_0)}\right)\right] = p_0\left[U + MA(\boldsymbol{p})\right]$$

As is standard, we normalize the price of the numeraire good to 1.[82] In this case the cost of living index (at initial utility level) is given by:

$$\Delta\ln E(\boldsymbol{p}, U_{t_0}) = \ln\left(1 + \frac{1}{Y_{t_0}}\int_{t_0}^{t_1}\sum_i q_i dp_i\right) = \ln\left(1 + \frac{M}{Y_{t_0}}\Delta A\right)$$

Here we can see the close relationship between the expected consumer surplus and the welfare index for the representative agent.[83]

It turns out that we can solve for ΔA using an index number formula similar to that used in the CES case. Taking the demand system for a single good, we have:

$$\ln q_i = \ln M + \alpha_i - \beta p_i + \beta A$$

[82]Note that the numeraire good included in the representative agent is in addition to the "outside good" normally included in the discrete choice setting. The inclusion of an outside good means that observed purchases do not have to add up to the full population M.

[83]Note the sign switch; a higher value of A is associated with higher prices which raises the cost of living and reduces welfare.

In turn, taking differences to eliminate the α_i term, we have

$$\Delta \ln q_i = \Delta \ln \pi_i = -\beta p_i + \beta A$$

Using Sato-Vartia weights, but with the logarithmic average of quantity shares[84] instead of expenditure shares, and summing over all goods, we have:

$$0 = -\beta \sum_i w_i \Delta p_i + \beta \Delta A \sum_i w_i$$

In turn, solving for ΔA we have:

$$\Delta A = \left(\sum_i w_i \right)^{-1} \sum_i w_i \Delta p_i$$

In addition, we may repeat this process using a partial sum rather than a full sum, in which case we have:

$$\Delta \pi_c = -\beta \sum_{i \in c} w_i \Delta p_i + \beta \Delta A \sum_{i \in c} w_i$$

Rearranging to solve for ΔA we have:

$$\Delta A = \left(\sum_{i \in c} w_i \right)^{-1} \left[\sum_{i \in c} w_i \Delta p_i + \frac{\Delta \pi_c}{\beta} \right]$$

Plugging this into the cost of living expression, we have:

$$\begin{aligned} \Delta \ln E &= \ln \left(1 + \frac{M}{Y_{t_0}} \Delta A \right) \\ &= \ln \left(1 + \frac{M}{Y_{t_0}} \left(\sum_{i \in c} w_i \right)^{-1} \sum_{i \in c} w_i \Delta p_i + \frac{M}{Y_{t_0}} \left[\left(\sum_{i \in c} w_i \right)^{-1} - 1 \right] \frac{\Delta \pi_c}{\beta} + \frac{M}{Y_{t_0}} \frac{\Delta \pi_c}{\beta} \right) \end{aligned}$$

Finally, we can rewrite the last term in this sum using the expenditure share and partial

[84] Note that it would also be valid to use the logarithmic average of the quantity itself rather than the quantity share. Since we normalize the shares by dividing the $\sum_i w_i$ these two approaches are equivalent.

elasticity form we have seen earlier. Specifically, we have:

$$\begin{aligned}\frac{M}{Y_{t_0}}\frac{\Delta\pi_c}{\beta} &= \sum_{i\in x}\frac{M\pi_{it_0}}{Y_{t_0}\beta} - \sum_{i\in n}\frac{M\pi_{it_1}}{Y_{t_0}\beta} \\ &= \sum_{i\in x}\frac{s_{it_0}}{p_{it_0}\beta} - \left(\frac{Y_{t_1}}{Y_{t_0}}\right)\sum_{i\in n}\frac{s_{it_1}}{p_{it_1}\beta} \\ &= \sum_{i\in x}\frac{s_{it_0}}{\eta_{it_0}} - \left(\frac{Y_{t_1}}{Y_{t_0}}\right)\sum_{i\in n}\frac{s_{it_1}}{\eta_{it_1}}\end{aligned}$$

CHAPTER 2

Is Heteroskedasticity-Based Identification Robust to Parameter Heterogeneity?

2.1 Introduction

This paper considers the sensitivity of heteroskedasticity-based regression strategies to variation in the structural parameters to be estimated. I review the assumptions of the heteroskedasticity-based identification methods and emphasize that these are at odds with the conditions needed to treat OLS estimates as weighted averages of underlying heterogeneous parameter values. Using simulated data, I show that the heteroskedasticity-based regression is very sensitive to the presence of heterogeneity in the structural parameters.

To evaluate the practical relevance of these concerns, I conduct three empirical exercises using U.S. trade data. First, using strictly weaker conditions than those required by standard double-difference heteroskedasticity-based regressions, I can construct bounds on one structural parameter (either supply or demand) per product. I show that for all industries (HS4 categories) there is at least one product per industry where the product-specific bound does not include the industry-level point estimate. Second, using the HBI-based point estimate for the industry supply elasticity, I can estimate a product-specific demand elasticity. This method finds noticeable dispersion in the demand elasticity, and a relatively weak correlation between industry-level point estimates and the median product-specific demand elasticity. Third, I evaluate how sensitive the HBI regression is to the choice of which product to include in the regression. There are economically

important swings in these parameters as we adjust the set of products from the top 10 to the top 15 products in the industry.

In the context of standard economic models, where observable values like price and expenditure are equilibrium outcomes, it is often difficult to directly recover estimates of the underlying structural parameters. One prominent approach is to search for exogenous variation that is useful for estimating the relevant parameter. This could take the form of a natural experiment, where circumstances conspire to create effectively random variation in the data, or an instrumental variable (IV) that generates exogenous variation in a relevant explanatory variable. A notable feature of the natural experiment and IV methods is that they generally identify an "average" effect; under relatively mild conditions even if there is variation in the underlying parameters (e.g. a treatment effect) the pooled regression is bounded by the underlying observation-specific values.

One prominent alternative to natural experiment and instrumental variable approaches is heteroskedasticity-based identification (HBI). In the standard supply and demand framework, the parameters for the supply and demand curves appear in the variance-covariance matrix for quantities (expenditures) and prices, but their effects are conflated with the relative importance of supply and demand shocks. The insight of the heteroskedasticity-based identification is that if there is variation in the relative importance of supply and demand shocks, we can use the induced variation in observed variance-covariance matrices to recover the underlying structural parameters.

This paper investigates an over-looked assumption in the heteroskedasticity-based identification strategy. A key condition is that the structural patterns are common across the observational units used in evaluating the moment condition. Unlike with natural experiments or standard IV frameworks, allowing for heterogeneity in the structural parameters leads to large swings in the estimation equation; there is no guarantee that the estimated parameters are even bounded within the set of underlying structural parameters.

Related Literature Feenstra (1994) originated the estimation framework studied here.

Feenstra (1994) showed how, within a CES framework and using the standard heteroskedasticity identification assumptions, we can recover both the CES import demand elasticity parameter and the supply elasticity for import sources. The Feenstra (1994) method was based on a two-stage least squares implementation for the imposed moment condition, and also made some allowance for measurement error in prices.[1]

This paper is most closely related to Imbs and Mejean (2015). That paper also notes the possibility for bias in the methodology of Feenstra (1994) when heterogeneous parameter values are improperly pooled. Imbs and Mejean (2015) argues for a particular direction of bias given likely correlations in the data. I supplement Imbs and Mejean (2015) by adding Monte Carlo simulations of the effects of improperly grouping heterogeneous goods together. Relative to Imbs and Mejean (2015) I focus on the erratic behavior of the elasticity point estimate rather than taking a stand on the sign of the bias. This also extends the critique of Imbs and Mejean (2015) beyond the levels-of-aggregation scenario they focus on to the issue of within-industry trade as well. Related to the concerns noted in this paper, Mohler (2009) documents the sensitivity of the Feenstra (1994) estimation procedure to arbitrary choices in the data preparation process.

A number of papers have proposed variations on the specific implementation of the moment condition and identification strategy developed in Feenstra (1994). The Monte Carlo evidence that I focus on applies to all of these refinements since I abstract from mismeasurement and finite sample biases. Broda and Weinstein (2006) proposes an alternative model for measurement error which is captured in a different weighting scheme for the two-stage least squares estimate, and also proposed a grid-search procedure to handle cases where unrestricted estimation yields results outside a valid range for the parameters. Soderbery (2010) noted that both Feenstra (1994) and Broda and Weinstein (2006) implementations may suffer from small-sample biases, while Soderbery (2015)

[1]Formally, the Feenstra (1994) regression may be stated in the form of a two-stage least squares where the "instrument" in the first stage is a country dummy. In terms of the point estimates, this is equivalent to an OLS where the observational units are country-industries and the regressor values are within-country time averages.

proposes a limited information maximum likelihood (LIML) estimation procedure to mitigate these small-sample bias issues. Further refinements to handle estimation at the edge of the eligible parameter space are taken up in von Brasch and Raknerud (2021).

Other notable examples of the use of the Feenstra (1994) estimation strategy include: Redding and Weinstein (2020), Argente and Lee (2020), Jaravel (2019), Arkolakis et al. (2019), Feenstra et al. (2018), Feenstra and Weinstein (2017), Hottman et al. (2016), Broda and Weinstein (2010).

Heteroskedasticity-based estimation is also used outside the trade and panel data cases. For example, Rigobon (2003) uses structural breaks in the variances of supply and demand shocks in high-frequency financial data (together with the same constant-parameter noted here). Rigobon (2003) documents that heteroskedasticity-based estimation is robust to modest misspecification of the form of changes in the relative importance of supply and demand shocks.

2.2 Model Environment

The elasticity of substitution between two goods, i and some base product b, may be denoted as σ_i or:

$$\frac{d\ln(s_{it}/s_{bt})}{d\ln(p_{it}/p_{bt})} = -(\sigma_{it}-1) \tag{2.1}$$

That is, the elasticity of substitution relates the change in relative quantities to the change in relative prices. In principle, the value of the elasticity of substitution depends on both the choice of the primary product (i) and the base-good (b) but to economize on notation I will neglect this issue. In the standard trade setting, where a symmetric demand system is assumed, the dependence on b is assumed away.

Generally we require that $\sigma > 1$ so that we can define the offset-elasticity as $\beta_d = (\sigma - 1) > 0$. Evaluating this expression using discrete changes, and adding an error term to capture deviations from some average level of the elasticity of substitution, we

have:

$$\Delta_2 \ln s_{it} = -\beta_{di} \Delta_2 \ln p_{it} + \varepsilon_{it} \tag{2.2}$$

where Δ_2 indicates a double difference; first with respect to time and second with respect to the base product. Written out for the share change, the double difference is defined as:

$$\Delta_2 \ln s_{it} = (\ln s_{it} - \ln s_{it-1}) - (\ln s_{bt} - \ln s_{bt-1}) \tag{2.3}$$

with the analagous definition holding for price double differences.

The standard concern about using OLS to estimate an equation of the form (2.2) is that changes in prices may be correlated with changes in the error term. In a standard supply and demand framework, the basic idea is that positive demand shocks for one or the other products will move us along that good's supply curve and potentially change the price. When supply curves are upward sloping, this creates a positive correlation between the error term (ε_{it}) and the price changes, leading to an upward bias in the estimate of β_d. Since β_d is supposed to be negative, for low levels of correlation this creates an attenuation bias and for high level of correlation the regression will give positive point estimates, contrary to standard restrictions. The analogous problem occurs for the supply curve relationships; shifts in supply move us along demand curve creating a correlation between supply errors and prices.

As shown in Leamer 1981, this attenuation intuition can be formalized with an assumption about the correlation between supply and demand shocks. First, consider a cost function of the form

$$\ln c_{it} = \alpha_i + \omega_i \ln q_{it} + \nu_{it} \tag{2.4}$$

Rearranging this equation and using a change of variables to write this in terms of shares and sales prices, we have:

$$\ln s_{it} = \frac{\alpha_i}{\omega_i} + \left(\frac{1+\omega_i}{\omega_i}\right) \ln(p_{it}) + \left(\frac{1}{\omega_i}\right) \ln \mu_{it} - \ln E_t + \nu_{it}/\omega_i \tag{2.5}$$

where E_t refers to total (consumer) expenditures and μ_{it} is the markup (if any) above prices.

If we suppose that product i and product b have the same supply elasticity ($\omega_i = \omega_b$) then taking the double difference of this this expression yields

$$\Delta_2 \ln s_{it} = \beta_{si} \Delta_2 \ln p_{it} + \delta_{it} \tag{2.6}$$

where $\beta_{si} = (1 + \omega_i)/\omega_i$ and $\delta_{it} = (\Delta_2 \nu_{it} + \Delta_2 \ln \mu_{it}) / \omega_i$.

So far, we have already imposed the key structural assumptions that constrict the signs of demand and supply parameters, specifically:

Assumption 1 [Supply and Demand Slopes]: $\sigma_i > 1$ and $\omega_i > 0 \implies \beta_{di} > 0$ and $\beta_{si} > 1$

The key statistical assumption made in Leamer (1981) is that the errors terms in the reduce form supply and demand expressions are uncorrelated, i.e.

Assumption 2 [Uncorrelated Supply and Demand Shocks]: $\mathbb{E}\left[\varepsilon_{it}\delta_{it}\right] = 0$

Motivated by this moment condition, we may combine the reduced form supply (equation 2.6) and demand (equation 2.2) expressions to match observed double differences with the stated moment condition:

$$\begin{aligned}
\varepsilon_{it}\delta_{it} &= \left[\Delta_2 \ln s_{it} + \beta_{di} \Delta_2 \ln p_{it}\right] \left[\Delta_2 \ln(s_{it}) - \beta_{si} \Delta_2 \ln(p_{it})\right] \\
&= \left[\Delta_2 \ln s_{it}\right]^2 - \beta_{di}\beta_{si} \left[\Delta_2 \ln p_{it}\right]^2 + \left[\beta_{di} - \beta_{si}\right] \Delta_2 \ln s_{it} \Delta_2 \ln p_{it}
\end{aligned}$$

In the limit, as the total number of time periods observed (T) increases, the left-hand side of this expression (normalized by T) converges to zero given the assumption that the errors terms are uncorrelated. Denoting the limiting values of the right-hand term

as follows:

$$\begin{aligned} \frac{1}{T}\sum_t \left[\Delta_2 \ln s_{it}\right]^2 &\to v_{si}^2 \\ \frac{1}{T}\sum_t \left[\Delta_2 \ln p_{it}\right]^2 &\to v_{pi}^2 \\ \frac{1}{T}\sum_t \left[\Delta_2 \ln s_{it}\right]\left[\Delta_2 \ln p_{it}\right] &\to v_{spi} \end{aligned} \tag{2.7}$$

we have the asymptotic expression

$$0 = v_{si}^2 - \beta_{di}\beta_{si}v_{pi}^2 + \left[\beta_{di} - \beta_{si}\right] v_{spi} \tag{2.8}$$

2.3 Leamer Bounds

Given the covariances for shares and prices, this expression induces a tight link between the demand and supply parameters. Rearranging this expression, we have:

$$(\beta_{di} + b)(\beta_{si} - b) = \frac{v_s^2}{v_p^2}\left[1 - r^2\right] \tag{2.9}$$

where $b \equiv v_{sp}/v_p^2$ and $r^2 \equiv \frac{v_{sp}^2}{v_s^2 v_p^2}$. Note that r^2 is guaranteed to be less than 1 due to the Cauchy-Schwartz inequality. Further rearranging to isolate the demand parameter, we have:

$$\beta_{di} = -b_i + \left(\frac{v_{si}^2}{v_{pi}^2}\right)\left[\frac{1 - r_i^2}{\beta_{si} - b_i}\right] \tag{2.10}$$

There are two notable features of this expression. First, the b_i term is exactly the value that is calculated if we were to run OLS on equation (2.2). Second, when b_i is negative given the structural assumption on β_{si} the second term is guaranteed to be positive. Thus when $b_i < 0$, we know that $|b_i|$ is an under-estimate of the true demand parameter.

We may also rearrange this expression to yield an upper bound for the demand parameter.

Specifically, we have:

$$\beta_{di} = \frac{(-v_{spi}/v_{pi}^2)\beta_{si} + v_{si}^2/v_{pi}^2}{\beta_{si} - v_{spi}/v_{pi}^2} = (-b_{ri}) \left[\frac{\beta_{si} r_i^2 - b_i}{\beta_{si} - b_i} \right] \tag{2.11}$$

Note that the value $-b_{ri} \equiv v_{si}^2/v_{spi}$ is the inverse of the OLS estimate from the reverse regression, i.e. a regression of the form

$$\Delta_2 \ln p_{it} = \beta_{ri} \Delta_2 \ln s_{it} + \gamma_{it}$$

for some error term γ_{it} and $b_{ri} = 1/\hat{\beta}_{ri}$.

When b_i is negative, then so is b_{ri}, since the sign of both matches the sign of the share-price covariance term (v_{spi}). In addition, the factor multiplying b_{ri} is guaranteed to be less than 1 in this case because r_i^2 is less than 1. Since we have to scale down $|b_{ri}|$ to get back to β_{di} we know that $|b_{ri}|$ is an overestimate of β_{di}. Summarizing, this gives us the following result

$$b_i < 0 \text{ and Assumptions } 1+2 \implies |b_i| < \beta_{di} < |b_{ri}|$$

An analogous set of results holds for the supply parameter when the sign of the OLS-estimated slopes (b_i and b_{ri}) are positive. That is, we have:

$$b_i > 0 \text{ and Assumptions } 1+2 \implies |b_i| < \beta_{si} < |b_{ri}|$$

Remark: Measurement Error A common concern in the international trade data is that there may be measurement error in prices. This is likely to occur since prices are constructed using average unit values rather than product-specific prices. If a source country supplies many varieties within a given HS category then unit values may not capture the effective average price relevant for consumer maximization. In addition, when data are aggregated to lower frequencies (e.g. annual) then the average price over a year may be an improper measure of the prices consumers are actually reacting to.

In an extension, I show that in the presence of measurement error the Leamer bound results continue to hold. That is, the direct OLS value (b_i) still under-estimates the true parameter while the inverse-OLS estimate (b_{ri}) are over-estimates the true parameter, where the parameter being bounded depends on the sign of v_{sp}. The intuition that classical measurement error does not affect the bound result is straightforward. Measurement error in the RHS variable creates attenuation bias leading the direct-OLS estimate to move towards zero. Thus, when there is measurement error in prices the extent to which $|b_i|$ is too low relative to one or the other parameter is magnified. On the other hand, when there is only measurement error in a LHS variable there is no effect on the asymptotic behavior of the OLS estimates so measurement error in prices has no effect on the bound result for $|b_{ri}|$. The same logic applies if there is measurement error in shares; there is no effect on the asymptotic value for $|b_i|$ while there is attenuation bias for $\hat{\beta}_r$ which, in turn, leads $|b_{ri}|$ to be too large relative to the relevant parameter.

2.4 Heteroskedasticity-based Identification

Rearranging equation (2.8), we can construct an asymptotic linear equation of the form

$$v_{pi}^2 = \underbrace{\left(\frac{1}{\beta_{di}\beta_{si}}\right)}_{\theta_{1i}} v_{si}^2 + \underbrace{\left[\frac{1}{\beta_{si}} - \frac{1}{\beta_{di}}\right]}_{\theta_{2i}} v_{spi} \tag{2.12}$$

This expression cannot be used as a regression without further assumptions, however. In its current form, this expression has two right-hand side variables but only one product[2] so that there are not enough observations to identify both β_{di} and β_{si}.

To achieve point identification, we first have to assume that there are multiple observations for which the underlying parameters are common. That is, we have:

Assumption 3 [Uniform Supply/Demand Slopes]: $\beta_{di} = \beta_d$ and $\beta_{si} = \beta_s$

[2]We may also add a constant in the regression, which would increase the number of right-hand side variables to three.

In the Feenstra (1994) setting, this panel dimension is introduced by adding additional products (countries) and assuming a CES demand system (for β_{di}) and a common slope for supply curves (for β_{si}). There is also a finance literature that adds a panel dimension by breaking the time-series for a given product into two components. In that context, Assumption 3 is requiring that the parameters for a product are fixed across this break in the sample.

In order for this panel dimension to deliver identification, we also have to ensure that the right-hand side of the regression is invertible. Rearranging equations (2.2) and (2.6) we can solve for the share and price double differences in terms of the β parameters and the shock terms as

$$\begin{aligned}(\beta_{si} + \beta_{di})\Delta_2 \ln s_{it} &= \beta_{si}\varepsilon_{it} + \beta_{di}\delta_{it} \\ (\beta_{si} + \beta_{di})\Delta_2 \ln p_{it} &= \varepsilon_{it} - \delta_{it}\end{aligned}$$

In turn, the asymptotic limits (in T) for the variance-covariance matrix of the share and price interactions is given by:

$$\begin{aligned}(\beta_{si} + \beta_{di})^2 v_{si}^2 &= \beta_{si}^2\sigma_{\varepsilon i}^2 + \beta_{di}^2\sigma_{\delta i}^2 \\ (\beta_{si} + \beta_{di})^2 v_{pi}^2 &= \sigma_{\varepsilon i}^2 + \sigma_{\delta i}^2 \\ (\beta_{si} + \beta_{di})^2 v_{spi} &= \beta_{si}\sigma_{\epsilon i}^2 - \beta_{di}\sigma_{\delta i}^2\end{aligned} \tag{2.13}$$

To ensure that the panel dimension indeed allows for an invertible design matrix when β_{di} and β_{si} are common, we must rely on there being differences in the relative importance of supply and demand shocks for each product. Specifically, we require that there are at least two products $i \neq j$ such that

Assumption 4 [Heteroskedasticity]: There is at least one pair of goods $i \neq j$ such that $\dfrac{\sigma_{\epsilon i}^2}{\sigma_{\epsilon j}^2} \neq \dfrac{\sigma_{\delta i}^2}{\sigma_{\delta j}^2}$

With these two additional assumptions, we now have the sufficient conditions to write the Feenstra (1994) regression equation in reduced form:

$$v_{pi}^2 = \theta_1 v_{si}^2 + \theta_2 v_{spi} \tag{2.14}$$

The conditions of Assumption 1 induce the following restrictions on the θ parameters:

$$\theta_1 \geq 0 \quad \theta_1 + \theta_2 \leq 1 \tag{2.15}$$

For the $\theta_1 > 0$ constraint, this follows from the fact that both β_d and β_s are positive values. The θ_1 value can approach zero either as demand moves towards perfect substitutes or as supply become inelastic (ω_i goes to zero, β_s goes to infinity). For the adding up condition, this is equivalent to requiring that $\beta_s > 1$ and $\theta_1 > 0$.

Inverting the θ_1 and θ_2 system to recover the underlying β_d and β_s parameters, we have:

$$\beta_d = \frac{\theta_2 + \left(\theta_2^2 + 4\theta_1\right)^{1/2}}{2\theta_1} = \frac{2}{(\theta_2^2 + 4\theta_1)^{1/2} - \theta_2}$$
$$\beta_s = \frac{-\theta_2 + \left(\theta_2^2 + 4\theta_1\right)^{1/2}}{2\theta_1} = \frac{2}{(\theta_2^2 + 4\theta_1)^{1/2} + \theta_2}$$

The expression in equation (2.14) may be estimated using OLS and using the sample analogs for the variance/co-variance terms previously defined in equation (2.7). When the time-dimension differs for each country we may weight a country-observation by the number of periods it appears in the data. While OLS suffices for the point estimates, this does not properly capture the fact that this expression is consistent in the time dimension. To incorporate this feature into the estimation of the confidence intervals, Feenstra (1994) shows how this expression may be rewritten as a country-time two-stage least squares estimation where the first stage simply uses country dummies to collapse values down to their mean values. Finally, Feenstra (1994) also incorporates a constant in the regression to control for measurement error in prices.

2.5 Misspecification Results

In this section I first describe the intuition for why least-squares estimates of equation (2.14) perform poorly when there is unmodeled underlying heterogeneity. The key issue is that, given the structural model that underlies the regression equation, incorrectly imposing uniform values for the supply and demand parameters is likely to create an omitted variable bias. In addition, given the non-linear mapping from the reduced form parameters θ_1 and θ_2 back to the underlying structural parameters this bias can lead to point estimates well outside the range in the underlying data. This point is reinforced with Monte Carlo estimates; even with only one incorrectly included observation and a small offset for the elasticity parameter the point estimates can be well outside the elasticity values in the true distribution.

The intuition for a bias for estimating equation (2.14) with unmodeled underlying heterogeneity is straightforward. The slope coefficients for an OLS of the form (2.14) will vary with the structural structural supply and demand parameters. In addition, by definition, the right-hand side variables from the sample variance-covariance matrix for each product vary systematically with these structural parameters. For example, a product with a high elasticity of substitution will see larger variances for share changes and a large (negative) covariance between share and price changes. At the same time, such a product will have a smaller value for θ_{1i} and a larger (more positive) value for θ_{2i}. In addition, the two RHS variables will themselves be correlated; greater variance in shares will (all else equal) be associated with a higher-magnitude covariance between prices and shares. This pattern of close correlation between the heterogeneous parameters and the variation in the observed regression values violates the conditions that allow us to consider the OLS parameter as a weighted average of the underlying parameters.

2.5.1 OLS as Weighted Average

Consider a generic regression equation of the form

$$y_i = \theta_{i1} x_{i1} + \theta_{i2} x_{i2}$$

If we impose that there are uniform θ values then we may rewrite this equation as:

$$y_i = \theta_1 x_{i1} + \theta_2 x_{i2} + \underbrace{[(\theta_{i1} - \theta_1) x_{i1} + (\theta_{i2} - \theta_2) x_{i2}]}_{\text{error}}$$

In this case, the OLS estimator of the single common parameter vector (θ_1, θ_2) takes the form:

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \frac{1}{\left(\sum_i x_{i1}^2\right)\left(\sum_i x_{i2}^2\right) - \left(\sum_i x_{i1} x_{i2}\right)^2} \begin{bmatrix} \sum_i x_{i2}^2 & -\sum_i x_{i1} x_{i2} \\ -\sum_i x_{i1} x_{i2} & \sum_i x_{i1}^2 \end{bmatrix} \begin{bmatrix} \sum_i \theta_{i1} x_{i1}^2 + \sum_i \theta_{i2} x_{i1} x_{i2} \\ \sum_i \theta_{i1} x_{i1} x_{i2} + \sum_i \theta_{i2} x_{i2}^2 \end{bmatrix}$$

In the typical case where x_{i2} is a constant, $\hat{\theta}_2$ corresponds to the intercept term in the regression and the OLS estimator $\hat{\theta}_1$ yields:

$$\hat{\theta}_1 = \frac{\sum_i \theta_{i1}(x_{i1} - \bar{x}_1)^2 + \bar{x} \sum_i (\theta_{i1} - \bar{\theta}_1)(x_{i1} - \bar{x}_1) + \sum_i (\theta_{i2} - \bar{\theta}_2)(x_{i1} - \bar{x}_1)}{\sum_i (x_{i1} - \bar{x}_1)^2}$$

where bars over variables indicate mean values. The standard assumption is that the regressor x_{i1} is uncorrelated with the unobserved i-specific parameter values. In this case, the additional right-hand side terms drop out and the remaining value is simply:

$$\hat{\theta}_1 = \frac{\sum_i \theta_{i1}(x_{i1} - \bar{x}_1)^2}{\sum_i (x_{i1} - \bar{x}_1)^2}$$

which is a weighted average of θ_{i1} values where the weights corresponds to the normalized deviations for each observation.

The structural model developed in section 2.4 does not imply the lack-of-correlation properties that would be needed to ensure this average value result. Instead, we can

see that the structural parameters β_{di} and β_{si} are embedded in both the reduced-form parameters (θ_{i1} and θ_{i2}) and the regressor values as shown in equation (2.13). The only thing that can be said ex-ante is that the function is continuous so that vanishingly small differences in structural parameters will not have an effect. However, as I show in the next section with Monte Carlo results the patterns in the data are potentially very volatile even for minor changes in the structural parameters.

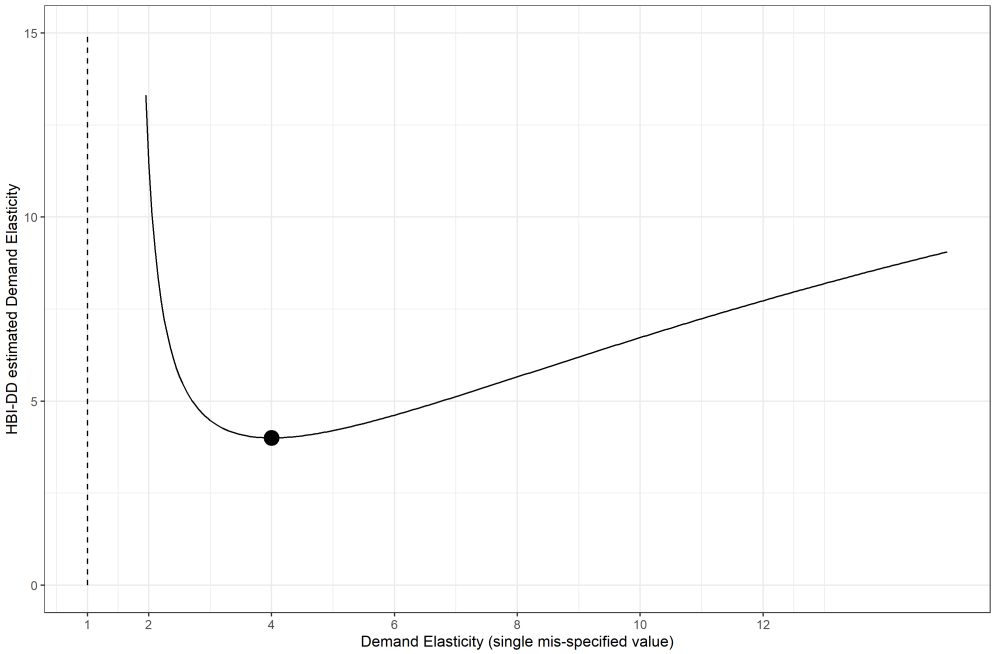
2.5.2 Monte Carlo Exercise

I consider two Monte-Carlo exercises to document the sensitivity of the HBI double-difference estimator to variation in the underlying parameters. In this section, I focus solely on the point estimate for the demand elasticity while maintaining the assumption that there is a single common supply elasticity. In addition, all exercises are done using the asymptotic form of the expression - i.e. all values are constructed without error using the formulas from (2.13). Thus, these while my implementation for the Monte Carlo uses OLS the issues I identify are likely to affect other estimation techniques as well as there are no finite-sample issues in my simulated setting.

First, I consider a sample of 50 observations where 49 observations all have a common elasticity while the final observation has a demand elasticity that is allowed to vary. In general, the asymptotic result will also depend on the choice of the supply elasticity and the distribution of variances for the supply and demand shocks. As a benchmark, I impose that the supply elasticity is equal to one (so that $\beta_s = 1/2$), and that the standard demand elasticity is equal to 4. In addition, I normalize the variance of the supply and demand shocks for the single misspecificed product to be equal to the mean supply and demand variances among the 49 correctly specified products. Thus, the patterns in this section are entirely driven by fact that the demand elasticity creates a correlation between the reduced form parameters and the regressor values.

The solid dot in Figure 2.5.1 notes the point where the regression is correctly specified because the varying-elasticity product matches the demand elasticity for the other 49

Figure 2.5.1: Point Estimate with Single Incorrectly Included Product

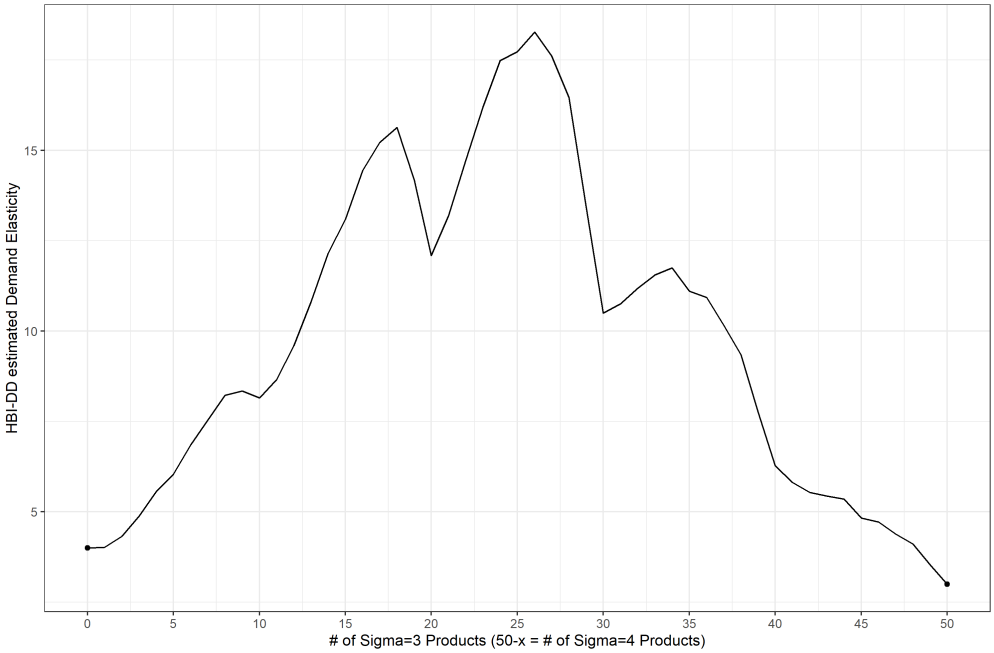


NOTE.—This figure plots how, in a sample of 50 goods, the overall point estimate changes as we vary the true elasticity of a single product. While the point estimate is continuous, it diverges quickly from the elasticity of the other 50 goods and is clearly outside the range of the underlying elasticities in the data-generating process.

goods. As we can see in the diagram, the point estimate quickly diverges when the single mis-specified product has a relatively low demand elasticity. Mechanically, the low demand elasticity estimate reduces the share variance and raises the share covariance but by non-proportional amounts. OLS tries to fit this pattern by reducing θ_1 and increasing θ_2; given the inversion formula for the demand elasticity this shift in the θ values leads to a sharp spike in the σ point estimate. As this process proceeds, θ_1 moves towards zero but this induces a point estimate for the demand parameter that is extremely large. As the single misspecified product moves towards a region of high-elasticity, a similar process unfolds. However, in this case (given the imposed pattern for the supply and demand shocks) the effect of increasing the odd-man-out demand elasticity runs into a diminishing marginal effect and there is slow-down in the effect of an increasing. The pattern may become unstable again, however, as the underlying θ_1 once again approaches zero.

Second, I consider how changing the mixture of products affects the estimate of the demand elasticity. I consider two groups of products, one with a demand elasticity equal to 4 and one with a demand elasticity equal to 3. As in the first exercise, I assume there is a single common supply elasticity equal to 1. I allow the relative frequency of $\sigma = 4$ and $\sigma = 3$ products to vary, moving from all $\sigma = 4$ to all $\sigma = 3$ and all intermediate cases from zero to 50. Again, in general, the specific results may depend on the values of the supply and demand shocks variances. In this exercise, I impose a fixed set of 50 supply and demand variance values ($\sigma^2_{\varepsilon i}$ and $\sigma^2_{\delta i}$) and then adjust the associated demand elasticities one at a time so that the variation in the diagram only reflects changes to the demand parameter.

Figure 2.5.2: Point Estimate with Varying Mixture ($\sigma = 3$ and $\sigma = 4$)



NOTE.—This figure evaluates how changing the mixture of elasticities within a narrow range affects the point estimate for a pooled regression. The point estimate diverges substantially from the small range of the underlying data.

Even with the relatively modest gap of an elasticity of 3 versus an elasticity of 4, the path for the point elasticities as we adjust the mixture quickly moves outside the region of the underlying product-elasticities. The specific zig-zag pattern, with notable whiplash, reflects the large changes that may be induced as we move over a more extreme shock-

variance observation. The types of paths that may develop is difficult to characterize in a simple way. However, the key finding here is simply the observation that the HBI double-difference point estimate can be radically different from the underlying distribution even under mild cases of parameter variation.

The mechanics of the second case are similar; when variation is driven by the differences in the demand parameters the asymmetric effect on the moment condition for θ_1 and θ_2 leads to a too-large estimate of the demand parameter. These results seem to be at odds with the expectation from Imbs and Mejean (2015) that the Feenstra (1994) point estimates are downward-biased. This, in part, reflects the particular setting for the Imbs and Mejean (2015) "elasticity optimism" result. Imbs and Mejean (2015) is concerned with estimation of excessively-aggregated demand systems, which motivates their assumption for the underlying correlations that give rise to omitted variable bias. The form of bias that I identify is driven entirely by unmodeled parameter heterogeneity and does not take a stand on the distribution of supply and demand shock variances per se. The finding in my Monte Carlo results that it is easy to generate too-high estimates of the demand elasticity in this setting matches the observation in the literature that Feenstra (1994)-style regressions tend to generate higher values for the demand elasticity than is typically seen using other regression techniques.

2.6 Empirical Evidence: Trade Data

The Monte Carlo exercise in section 2.5.2 shows that double-difference estimators may be sensitive even to relatively small misspecification error. While this may be a theoretical concern, it does not tell us whether this is likely to be an issue in actual data. In this section, I provide model-consistent ways to evaluate whether the conditions for the HBI double-difference estimates are present in the data.

I consider a standard trade environment where the HBI double-difference estimation strategy has been widely applied. I use data on US imports at the HS4 level of aggregation.

A "product" in this data is a country source within the HS4 category. For example, HS4 code 8703 covers passenger vehicles and a product in this category could be vehicles from Canada or vehicles from Japan.

I consider two exercises regarding the performance of the Feenstra (1994) estimation procedure.

First, I evaluate whether the HBI double-difference point estimates for an HS4 category lies within the Leamer (1981) bounds for individual products. In all but one HS4 category there is at least one product where the Leamer (1981) bounds do not include the corresponding HBI double-difference point estimate.[3]

The car market is also one of the industries where there are products for which the industry-level point estimate is not included in the product-specific Leamer bounds. For example France, Germany, and Spain all have 30 years of data in the sample and for each the Leamer upper bound is below the industry-level point estimate. Allowing for uncertainty in the (inverse) OLS coefficient that defines the upper bound, the upper-tail 95% confidence level for simple OLS regression does not include the industry-level point estimate in the case of France or Spain.[4] At the other end of the distribution, for Turkey and Vietnam (with 12 and 16 years in the data, respectively) the Leamer lower bound is above the industry-level point estimate. In addition, allowing for uncertainty in the Leamer bounds themselves the lower-tail 95% confidence level for the Vietnam OLS regressor still does not reach the industry-level point estimate.

Second, using the industry-level point estimate for the supply estimate, I calculate product-specific estimates of the demand elasticities by evaluating equation (2.10). In

[3]Strictly, there are 824 HS4 categories in the data. Of these, using the method of Feenstra (1994) to evaluate equation (2.14) there are only only 552 industries with point estimates in the valid range given by equation (2.15). In addition, using the weighting scheme proposed in Broda and Weinstein (2006) there are 691 valid point estimates. Among either the 552 or the 691 industries with valid point estimates, at least one product has a Leamer bound that does not include the industry-level point estimate.

[4]Given that there are three countries, the collection of Germany, France, and Spain is sufficient to run a regression. However, including only these three countries yields a negative point estimate for θ_1 which is outside the valid range.

this case, the correlation between the median product-specific elasticities and the HBI double-difference point estimate is only 0.3.

Figure 2.6.1: Car Market (HS4 code 8703) Demand Elasticity Distribution

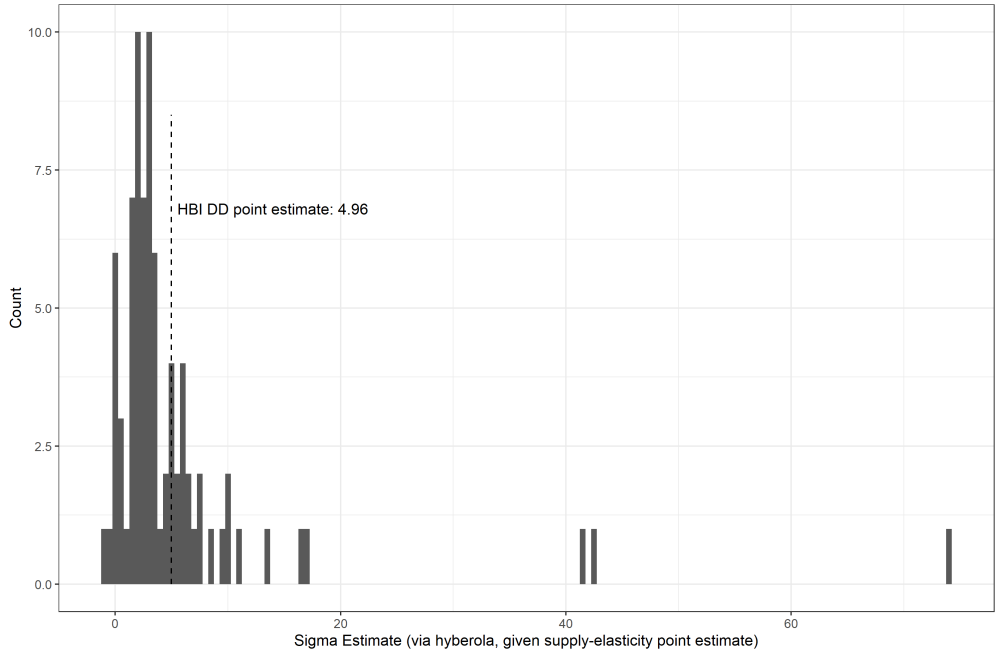
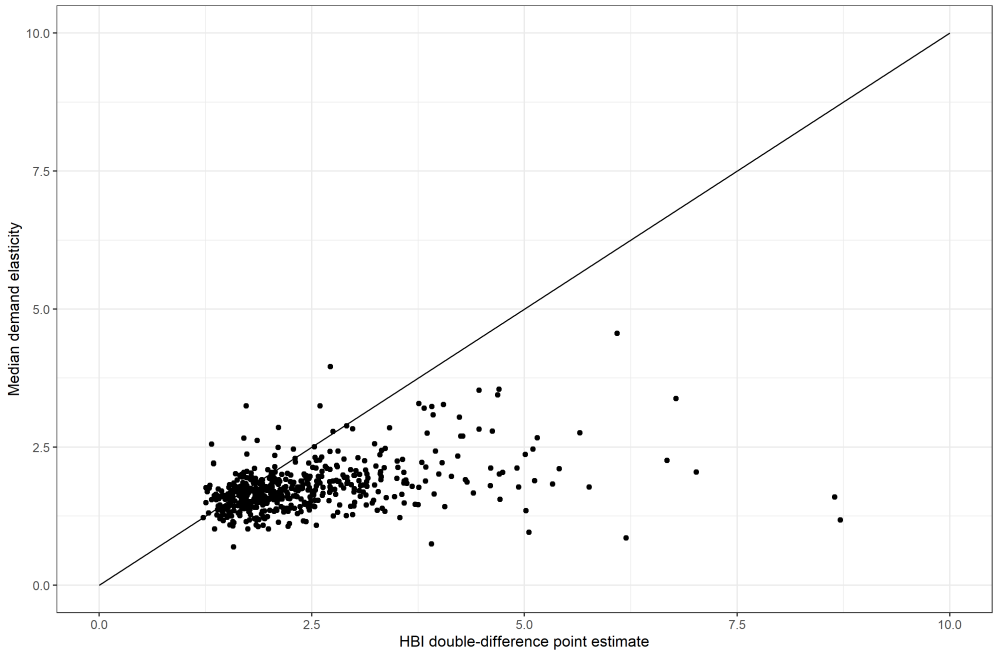


Figure 2.6.1 gives an example of the distribution of demand elasticities constructed by using this hyperbola result and a single common supply elasticity. In the car market, there are 81 import source countries in the sample period. Using the procedure outlined above, there are same products with negative demand parameters ($\sigma \leq 1$) and some extreme upper-tail outliers as well. In this case, the median demand elasticity is 3.0 while the simple average of product-specific demand elasticities is 5.6, relative to the HBI double-difference based point estimate of 4.96.

While the car market provides a useful example, I consider this exercise for every industry in the HS4 data. As shown in Figure 2.6.2 the median elasticity is systematically quite low while there is a wide range of HBI double-difference point estimates. The 45-degree line is plotted in 2.6.2; here, we can see that most of the HBI double-difference point estimates are well above the corresponding median using the hyperbola inversion.

Figure 2.6.3 repeats the same exercise, except in this case I look at the the simple mean

Figure 2.6.2: Median Demand Elasticity vs. Point Estimate

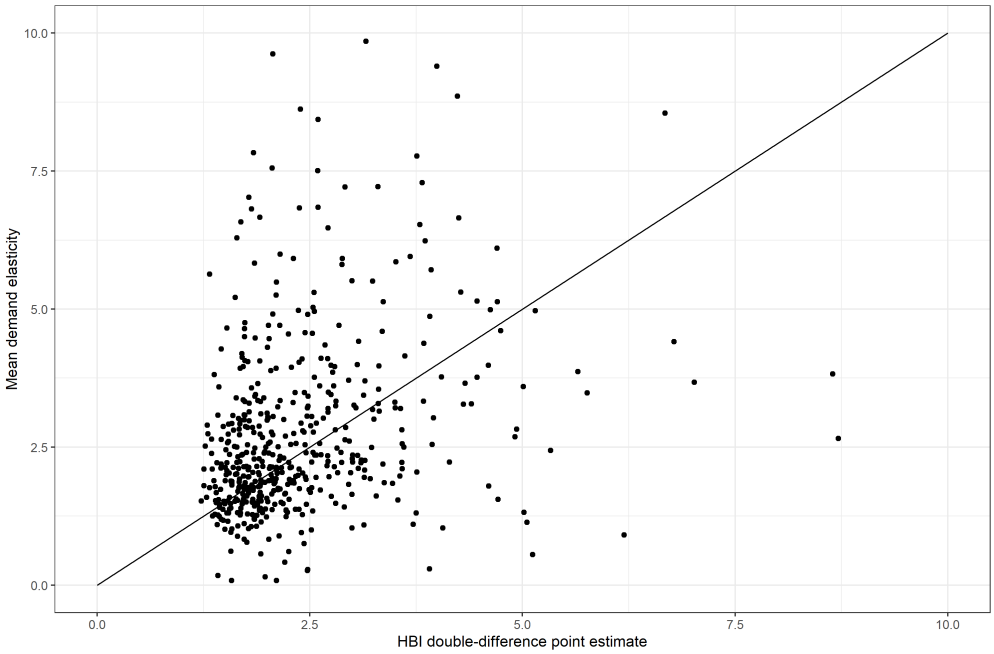


value rather than the median. The mean is not systematically lower than the HBI double-difference point estimate. However, there is still substantial heterogeneity between these two measures of the elasticity. In this case, the correlation coefficient falls to only 0.03, relative to the 0.3 correlation coefficient between the distribution-median and the HBI double-difference point estimates for each industry.

2.7 Conclusion

This paper has studied the properties of heteroskedasticity-based identification estimation procedures when there is unmodeled heterogeneity in the parameters of interest. Using Monte Carlo simulations, I document that the point estimates are sensitive to variation in the true underlying demand elasticity. This result follows from the structure of the model, as the demand elasticity affects both the observed regressors and the structural parameters that we seek to estimate. The HBI regression is sensitive to even modest variation in the underlying parameters and can also exhibit invalid point estimates even when standard conditions are satisfied on a product-by-product basis.

Figure 2.6.3: Mean Demand Elasticity vs. Point Estimate



I also use model-consistent methods to evaluate the potential prevalence of parameter heterogeneity in international trade data, a common setting for the HBI estimation procedure. I document that in almost all cases the industry-level point estimates do not lie within product-specific bounds for the elasticity parameters that are assumed to be uniform across all goods. I also show that, conditional on the point estimate for the supply elasticity there is substantial heterogeneity in the demand elasticity and a substantial discrepancy between these product-specific estimates and the industry-level point estimates drawn from the HBI procedure.

Both the Monte Carlo and empirical exercises seem to suggest that there is a potential for the HBI estimation procedure to yield an over-estimate for the underlying demand parameter. This result is at odds with the predictions in Imbs and Mejean (2015), but does match the observation that HBI estimation procedures tend to generate higher values for the elasticity than seen using other methods.

The results in this paper suggest two notes of caution for the literature using HBI estimates of demand elasticities. While substantial attention has been paid to the statistical

assumption of uncorrelated supply and demand shocks as well as methods for addressing finite sample biases the centrality of the uniform parameter assumption has seen a dirth of emphasis. This paper suggests that satisfying the uniform parameter assumption is of first-order importance in applying the HBI method. In addition, I document that parameter heterogeneity interacts with the non-linear estimation procedure to create substantial instability when one of the reduced form coefficients moves towards zero. Given the inherent uncertainty present in point estimates, the bounding results that I review may also be useful in deciding on calibrations on the occasion when the bounds are reasonably tight. The search for the true elasticity, or at least the true distribution of elasticities, thus continues.

Bibliography

Aghion, P., Bergeaud, A., Boppart, T., Klenow, P. J., and Li, H. (2019). Missing growth from creative destruction. *American Economic Review*, 109(8):2795–2822.

Argente, D. and Lee, M. (2020). Cost of living inequality during the great recession. *Journal of the European Economic Association*, 19(2):913–952.

Arkolakis, C., Costinot, A., and Claire, A. R. (2012). New trade models, same old gains? *American Economic Review*, 102(1):94–130.

Arkolakis, C., Costinot, A., Donaldson, D., and Rodriguez-Clare, A. (2019). The elusive pro-competitive effects of trade. *The Review of Economic Studies*, 86(1):46–80.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogenous causal effects. *Proceedings of the National Academy of Science*, 113(27):7353–7360.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11(1):685–725.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.

Atkin, D., Faber, B., Fally, T., and Gonzalez-Navarro, M. (2020). Measuring wealth and inequality with incomplete price information. NBER Working Paper 26890.

Atkin, D., Faber, B., and Gonzalez-Navarro, M. (2018). Retail globalization and household welfare: Evidence from mexico. *Journal of Political Economy*, 126(1):1–73.

Bajari, P., Cen, Z., Chernozhukov, V., Manukonda, M., Wang, J., Huerta, R., Li, J., and Leng, L. (2021). Hedonic prices and quality adjusted price indices powered by ai. Working Paper CWP04/21, Centre for microdata methods and practice.

Baqaee, D. and Burstein, A. (2021). Welfare and output with income effects and taste shocks. NBER Working Paper 28754.

Barnett, W. A. and Choi, K.-H. (2008). Operational identification of the complete class of superlative index numbers: An application of galois theory. *Journal of Mathematical Economics*, 44(708):603–612.

Behrens, K., Kanemoto, Y., and Murata, Y. (2017). On measuring welfare change when varieties are endogenous. CEPR Discussion Paper No. DP11774.

Bergin, P. R. and Feenstra, R. C. (2009). Pass-through of exchange rates and competition between floaters and fixers. *Journal of Money, Credit, and Banking*, 41(1):35–70.

Blackorby, C., Primont, D., and Russel, R. R. (1978). *Duality, separability, and functional structure: Theory and economic applictions*. North-Holland.

Braun, R. and Lein, S. M. (2021). Sources of bias in inflation rates and implications for inflation dynamics. *Journal of Money, Credit, and Banking*, 53(6):1553–1572.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Broda, C. and Weinstein, D. E. (2006). Globalization and the gains from variety. *The Quarterly Journal of Economics*, 121(2):541–585.

Broda, C. and Weinstein, D. E. (2010). Product creation and destruction: Evidence and price implications. *American Economic Review*, 100(3):671–723.

Chernozhukov, V., Hausman, J. A., and Newey, W. K. (2019). Demand analysis with many prices. Working Paper 26424, National Bureau of Economic Research.

Christensen, L. R., Jorgenson, D. W., and Lau, L. J. (1975). Transcendental logarithmic utility functions. *American Economic Review*, 65(3):367–383.

Crawford, I. and Neary, J. P. (forthcoming). New characteristics and hedonic price index numbers. *The Review of Economics and Statistics*.

Deaton, A. and Muellbauer, J. (1980). An almost ideal demand system. *American Economic Review*, 70(3):312–326.

Diewart, W. E. and Wales, T. J. (1988). A normalized quadratic semiflexible functional form. *Journal of Econometrics*, 37(3):327–342.

Diewert, W. E. (1976). Exact and superlative index numbers. *Journal of Econometrics*, 4(2):115–145.

Diewert, W. E. (1978). Superlative index numbers and consistency in aggregation. *Econometrica*, 46(4):883–900.

Diewert, W. E. (2002). The quadratic approximation lemma and decompositions of superlative indexes. *Journal of Economic and Social Measurement*, 28(1-2):63–88.

Diewert, W. E. and Feenstra, R. C. (2021). Estimating the benefits of new products. In Abraham, K. G., Jarmin, R. S., Moyer, B., and Shapiro, M. D., editors, *Big Data for Twenty-First Century Economic Statistics*. University of Chicago Press.

Diewert, W. E. and Wales, T. J. (1987). Flexible functional forms and global curvature conditions. *Econometrica*, 55(1):43–68.

Dixit, A. K. and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *American Economic Review*, 67(3):297–308.

Ehrlich, G., Haltiwanger, J., Jarmin, R., Johnson, D., Olivares, E., Pardue, L., Shapiro, M. D., and Zhao, L. Y. (2021). Quality adjustment at scale: Hedonic vs. exact demand-based price indices. Working Paper.

Faber, B. and Fally, T. (2017). Firm heterogeneity in consumption baskets: Evidence from

home and store scanner data. Working Paper 23101, National Bureau of Economic Research.

Faber, B. and Fally, T. (2021a). Firm heterogeneity in consumption baskets: Evidence from home and store scanner data. The Review of Economic Studies (forthcoming).

Faber, B. and Fally, T. (2021b). Firm heterogeneity in consumption baskets: Evidence from home and store scanner data. *Review of Economic Studies*.

Fajgelbaum, P. D. and Khandelwal, A. K. (2016). Measuring the unequal gains from trade. *The Quarterly Journal of Economics*, 131(3):1113–1180.

Fally, T. (2020). Integrability, generalized separability, and a new class of demand systems. Working Paper.

Feenstra, R. C. (1994). New product varieties and the measurement of international prices. *American Economic Review*, 84(1):157–177.

Feenstra, R. C. (2003). A homothetic utility function for monopolistic competition models without ces. *Economics Letters*, 78:79–86.

Feenstra, R. C. (2010). New products with a symmetric aids expenditure function. *Economics Letters*, 106(2):108–111.

Feenstra, R. C. (2018). Restoring the product variety and pro-competitive gains from trade with heterogenous firms and bounded productivity. *Journal of International Economics*, 110:16–27.

Feenstra, R. C., Luck, P., Obstfeld, M., and Russ, K. N. (2018). In search of the armington elasticity. *The Review of Economics and Statistics*, 100(1):135–150.

Feenstra, R. C. and Shiells, C. R. (1996). Bias in u.s. import prices and demand. In Bresnahan, T. F. and Gordon, R. J., editors, *The Economics of New Goods*, pages 249–276. National Bureau of Economic Research.

Feenstra, R. C. and Weinstein, D. E. (2017). Globalization, markups, and us welfare. *Journal of Political Economy*, 125(4):1040–1074.

Groshen, E. L., Moyer, B. C., Aizcorbe, A. M., Bradley, R., and Friedman, D. M. (2017). How government statistics adjust for potential biases from quality change and new goods in an age of digital technologies: A view from the trenches. *Journal of Economic Perspectives*, 31(2):187–210.

Handbury, J. (2021). Are poor cities cheap for everyone? non-homotheticity and the cost of living across u.s. cities. *Econometrica*, 89(6):2679–2715.

Handbury, J. and Weinstein, D. E. (2015). Goods prices and availability in cities. *The Review of Economic Studies*, 82(1):258–296.

Hausman, J. (1996). Valuation of new goods under perfect and imperfect competition. In Gordon, T. F. B. . R. J., editor, *The Economics of New Goods*, chapter 5. University of Chicago Press.

Hausman, J. (2003). Sources of bias and solutions to bias in the consumer price index. *Journal of Economic Perspectives*, 17(1):23–44.

Hausman, J. A. and Newey, W. K. (2017). Nonparametric welfare analysis. *Annual Review of Economics*, 9(1):521–546.

Hicks, J. R. (1940). The valuation of the social income. *Economica*, 7(26):105–124.

Hottman, C., Redding, S., and Weinstein, D. (2016). Quantifying the sources of firm heterogeneity. *Quarterly Journal of Economics*, 131:1291–1364.

Hsieh, C.-T., Li, N., Ossa, R., and Yang, M.-J. (2020). Accounting for the new gains from trade liberalization. *Journal of International Economics*, 127:1–29.

Hulten, C. R. (1973). Divisa index numbers. *Econometrica*, 41(6):1017–1025.

Hurwicz, L. and Uzawa, H. (1971). On the integrability of demand functions. In

Chipman, J. S., editor, *Preferences, Utility and Demand*, pages 114–148. Harcourt Brace, Jovanovich.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Imbs, J. and Mejean, I. (2015). Elasticity optimism. *American Economic Journal: Macroeconomics*.

Jaravel, X. (2019). The unequal gains from product innovations: Evidence from the u.s. retail sector. *The Quarterly Journal of Economics*, 134(2):715–783.

Kee, H. L., Nicita, A., and Olarreaga, M. (2008). Import demand elasticities and trade distortions. *The Review of Economics and Statistics*, 90(4):666–682.

Klenow, P. J. and Li, H. (2020). Innovative growth accounting. NBER Working Paper 27015.

Konny, C. G., William, B. K., and Friedman, D. M. (2019). Big data in the u.s. consumer price index: Experiences and plans. In Abraham, K. G., Jarmin, R. S., Moyer, B., and Shapiro, M. D., editors, *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Konus, A. A. (1924). The problem of the true index of the cost of living. translated in Econometrica 7 (1939) 10-29.

Kroft, K., Laliberté, J.-W. P., Leal-Vizcaíno, R., and Notowidigdo, M. J. (2021). Parallel inverse aggregate demand curves in discrete choice models. *Economic Theory*.

Leamer, E. E. (1981). Is it a demand curve, or is it a supply curve? partial identification through inequality constraints. *The Review of Economics and Statistics*, 63(3):319–327.

Matsuyama, K. and Ushchev, P. (2017). Beyond ces: Three alternative classes of flexible

homothetic demand systems. CEPR Discussion Papers 12210, C.E.P.R. Discussion Papers.

Melitz, M. J. (2018). Competitive effects of trade: theory and measurement. *Review of World Economics*, 154(1):1–13.

Mohler, L. (2009). On the sensitivity of estimated elasticities of substitution. FREIT Worker Paper No. 38.

Mrazova, M. and Neary, J. P. (2017). Not so demanding: Demand structure and firm behavior. *American Economic Review*, 107(12):3835–3874.

Neary, P. (2004). Rationalizing the penn world tables: True multilateral indices for international comparisons of real income. *American Economic Review*, 94(5):1411–1428.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342.

Nevo, A. (2003). New products, quality changes and welfare measures computed from estimated demand systems. *The Review of Economics and Statistics*, 82(2):266–275.

Nie, X. and Wager, S. (forthcoming). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*.

Novy, D. (2013). International trade without ces: Estimating translog gravity. *Journal of International Economics*, 89(2):271–282.

Pakes, A. (2003). A reconsideration of hedonic price indexes with an application to pc's. *American Economic Review*, 93(5):1578–1596.

Redding, S. J. and Weinstein, D. E. (2020). Measuring aggregate price indices with taste shocks: Theory and evidence fro ces preferences. *The Quarterly Journal of Economics*, 135(1):503–560.

Rigobon, R. (2003). Identification through heteroskedasticity. *The Review of Economics and Statistics*, 85(4):777–792.

Romer, P. M. (1990). Endogenous technological change. *The Journal of Political Economy,*, 98(5):S71–S102.

Rothbarth, E. (1941). The measurement of changes in real income under conditions of rationing. *The Review of Economic Studies*, 8(2):100–107.

Sato, K. (1976). The ideal log-change index number. *The Review of Economics and Statistics*, 58(2):223–228.

Silberberg, E. (1972). Duality and the many consumer's surpluses. *American Economic Review*, 62(5):942–952.

Small, K. A. and Rosen, H. S. (1981). Applied welfare economics with discrete choice models. *Econometrica*, 49(1):105–130.

Soderbery, A. (2010). Investigating the asymptotic properties of import elasticity estimates. *Economics Letters*, 109(2):57–62.

Soderbery, A. (2015). Estimating import supply and demand elasticities: Analysis and implications. *Journal of International Economics*, 96(1):1–17.

Soderbery, A. (2018). Trade elasticities, heterogeneity, and optimal tariffs. *Journal of International Economics*, 114:44–62.

Tornqvist, L. (1936). The bank of finland's consumption price index. *Bank of Finland Monthly Bulletin*, 16(10):27–32.

Ueda, K., Watanabe, K., and Watanabe, T. (2019). Product turnover and the cost-of-living index: Quality versus fashion effects. *American Economic Journal: Macroeconomics*, 11(2):310–347.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economics Perspectives*, 28(2):3–28.

Vartia, Y. (1976). Ideal log-change index numbers. *Scandinavian Journal of Statistics*, 3:121–126.

von Brasch, T. and Raknerud, A. (2021). A two-stage pooled panel data estimator of demand elasticities. Statistics Norway, Research Department Discussion Paper No. 951.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogenous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

Zhelobodko, E., Kokovin, S., Parenti, M., and Thisse, J.-F. (2012). Monopolistic competition: beyond the constant elasticity of substitution. *Econometrica*, 80(6):2765–2784.