WORKING PAPER · NO. 2026-101

Supply and Demand with Market Heterogeneity

Ingvil Gaarder, Lancelot Henry de Frahan, Magne Mogstad, Alexander Torgovitsky, and Oscar Volpe

JULY 2026

THE UNIVERSITY OF CHICAGO
BECKER FRIEDMAN INSTITUTE FOR ECONOMICS

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

Supply and Demand with Market Heterogeneity
Ingvil Gaarder, Lancelot Henry de Frahan, Magne Mogstad, Alexander Torgovitsky, and Oscar Volpe
July 2026
JEL No. C14, C18, C36, H21, H22

ABSTRACT

We revisit the classic identification problem of separating supply and demand for a homogeneous good using data from multiple markets. We allow markets to be heterogeneous according to unobservables, a feature that arises if there are unobservable differences in consumer preferences or firm technology. We develop a new identification analysis based on hypothetical market types. We use this analysis to show how nonparametric, economically motivated assumptions carry empirical restrictions for a wide range of target parameters, including elasticities, but also welfare parameters, such as consumer surplus. Then, we develop computationally tractable methods for implementing partially identified linear random coefficients models in which the slopes of supply and demand are heterogeneous. We apply these methods to estimate the welfare impact and incidence of sales taxes in the United States.

Ingvil Gaarder
The University of Chicago
ingvil.gaarder@gmail.com

Lancelot Henry de Frahan
The University of Chicago
lancelot@uchicago.edu

Magne Mogstad
The University of Chicago
Department of Economics
and NBER
magne.mogstad@gmail.com

Alexander Torgovitsky
The University of Chicago
Department of Economics
atorgovitsky@gmail.com

Oscar Volpe
Harvard University
ovolpe@fas.harvard.edu

Supply and Demand with Market Heterogeneity*

Ingvil Gaarder Lancelot Henry de Frahan Magne Mogstad
Alexander Torgovitsky Oscar Volpe

May 20, 2026

Abstract

We revisit the classic identification problem of separating supply and demand for a homogeneous good using data from multiple markets. We allow markets to be heterogeneous according to unobservables, a feature that arises if there are unobservable differences in consumer preferences or firm technology. We develop a new identification analysis based on hypothetical market types. We use this analysis to show how nonparametric, economically motivated assumptions carry empirical restrictions for a wide range of target parameters, including elasticities, but also welfare parameters, such as consumer surplus. Then, we develop computationally tractable methods for implementing partially identified linear random coefficients models in which the slopes of supply and demand are heterogeneous. We apply these methods to estimate the welfare impact and incidence of sales taxes in the United States.

1 Introduction

The original identification problem in econometrics is to separate the supply and demand for a single homogeneous good using data from multiple markets (Wright, 1915, 1928; Tinbergen, 1930/1995; Schultz, 1938). This problem has traditionally been analyzed using linear models with constant coefficients (for example, Haavelmo, 1943; Koopmans and Hood, 1953; Fisher, 1966). Identification in that classical model was already well-understood when Goldberger (1972) surveyed it in his Schultz Lecture.

*Gaarder: Harris School of Public Policy, University of Chicago. Henry de Frahan and Torgovitsky: Kenneth C. Griffin Department of Economics, University of Chicago. Mogstad: Kenneth C. Griffin Department of Economics, University of Chicago; NBER; Statistics Norway. Volpe: Department of Economics, Harvard University. Margaret Chen, Utkarsh Dandanayak, and Arnstein Vestre provided outstanding research assistance. An early draft of this paper was circulated under the title "Linear Supply and Demand in Heterogeneous Markets," dated November 15, 2024. We thank Josh Angrist for a helpful comment.

Disclaimer: *Researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.*

In the time since Goldberger's lecture, there has been a shift in econometrics towards methods that incorporate unobserved heterogeneity. This shift has been particularly important in models with instrumental variables (Heckman, 1976, 2001; Imbens and Angrist, 1994; Mogstad and Torgovitsky, 2024). The vast majority of this research has focused on models without simultaneity.

Yet simultaneity continues to be an important concern in many strands of empirical research. Examples include estimating the demand for oil (Kilian, 2009) or gasoline (Hughes et al., 2008; Coglianese et al., 2017), estimating regional Phillips curves (McLeay and Tenreyro, 2020; Hazell et al., 2022), estimating the supply of housing (Saiz, 2010), health insurance (Hackmann et al., 2015), the incidence of employer-provided non-pecuniary benefits (Saez et al., 2012) or corporate taxes (Suárez Serrato and Zidar, 2016), and the demand for agricultural goods in developing countries (Bergquist and Dinerstein, 2020). The industrial organization literature makes heavy use of simultaneous equations, often using discrete choice models in a more complicated context with multiple differentiated products (for example, Berry, 1994; Berry et al., 1995; Gandhi and Nevo, 2021; Berry and Haile, 2021).

In this paper, we take a fresh look at incorporating unobserved market-level heterogeneity into the classical model with a single homogeneous good. Our starting point is Manski's (1994, 1997) observation that even in a model with unrestricted heterogeneity, unrestricted nonlinearity, and no instruments, the natural economic assumption of downward-sloping demand has empirical content. The bounds produced by only this assumption tend to be wide (unsurprisingly) and are usually too wide to be particularly useful. However, they point to the optimistic possibility that combining multiple assumptions with instruments could lead to more informative conclusions. Exploring this possibility requires taking an approach to identification analysis that is more flexible than the one used by Manski.

We develop a new approach based on hypothetical market types. Each hypothetical market type is a combination of equilibrium price and quantity pairs taken across different possible realizations of an instrument. In the simplest case of a binary instrument, a market type is defined as two pairs of price and quantity. We develop a graphical analysis that shows how assumptions motivated by economic reasoning, such as downward-sloping demand or monotonicity in a shifter, restrict the number of a priori possible market types. Assumptions that eliminate market types leave fewer types to explain the data, leading to tighter identified sets (bounds) on interesting target parameters.

Our approach uses similar logic as arguments about choice types (Robins and Greenland, 1992; Heckman and Pinto, 2018), also called principal strata (Frangakis and Rubin, 2002). These arguments have been used to good effect in models without simultaneity, most famously through the monotonicity condition introduced by Imbens and Angrist (1994). We show how the version of the monotonicity condition suggested by Angrist

et al. (2000) can be combined with downward-sloping demand, and with the "Ramsey Exclusion Restriction" considered by Zoutman et al. (2018). We develop a computational approach for computing sharp nonparametric bounds under any combination of these assumptions, as well as others that a researcher may wish to maintain.

As a first step, we apply our approach to the Fulton fish market data used by Graddy (1995) and Angrist et al. (2000). We show that fully nonparametric bounds that allow for arbitrary market heterogeneity can rule out the possibility of inelastic demand. The nonparametric bounds also do not contain the linear instrumental variables (IV) estimand, which is attenuated towards zero. We show that this finding has a clear theoretical explanation that is likely at play in many empirical settings: linear IV estimands overweight markets where the instrument has a larger impact on moving equilibrium prices. This will typically mean that linear IV estimands will overweight markets that are less elastic, leading to attenuation.

Next, we adapt our methodology to incorporate the assumption that each market has a linear (or log-linear) demand and/or supply function, but with heterogeneous slopes. The result is a random coefficients model. We show that the random coefficients model is still partially identified, but that conservative prior assumptions on the magnitude of the demand and/or supply functions can produce remarkably informative bounds on interesting target parameters. A notable consequence of our partial identification approach is that we can also produce informative bounds on target parameters that involve both demand and supply together, such as deadweight loss, even with only a supply shifter as an instrument. A classical analysis with constant coefficients would conclude that such a parameter is unidentified without the availability of a demand shifter to identify the supply coefficient.

Our findings are interesting in the context of Masten's (2018) study of the same random coefficients model under simultaneity. Masten showed that continuous instruments are necessary and—under additional strong assumptions—also sufficient for point identification. The bounds we obtain use only a binary instrument, but in many cases are remarkably narrow. The implication is that even when the random coefficients model is not point identified, it can still contain considerable empirical content.

We then make our analysis more scalable by smoothing the heterogeneity across market types in the random coefficients model. We assume that the distribution of the random coefficients follows a linear basis expansion, as in sieve methods (for example, Chen, 2007). This provides the researcher the flexibility to consider anything from a tightly specified parametric model to an extremely expressive model that becomes nonparametric as the number of terms in the basis expansion increases.

We show how to compute and estimate sharp bounds on standard target parameters, such as features of the distribution of demand, surplus measures, or deadweight loss. We entertain two methods: a full information method that matches the full distribution of

the observables and a limited information method that matches only the coefficients produced by reduced form linear regressions. Computation for either method involves solving convex, linearly constrained quadratic programming problems, so is feasible in high dimensions and with typical sorts of covariate specifications. Because estimation acknowledges the possibility of partial identification, we do not need to make any assumptions about the available instrument variation, making the methods applicable with the types of discrete instruments and/or covariates commonly encountered in practice.

We apply our methodology to estimate the welfare impact and incidence of sales taxes in the United States using changes in sales taxes as the instrument. A constant coefficients model employing the Ramsey exclusion restriction (Zoutman et al., 2018) yields a point estimate of the (constant) demand elasticity of 0.62 and a point estimate of the (constant) supply elasticity of 5.81. The random coefficients model produces bounds on the cross-market average demand elasticity of $[0.71, 0.96]$. Bounds on the average supply elasticity are $[3.91, 11.60]$, considerably wider than for demand, which suggests more ambiguity in supply than demand when allowing for unobserved heterogeneity across markets. Despite this ambiguity, the random coefficients model produces a near-point estimate of 0.91 for consumer incidence, which is almost identical to the constant coefficients estimate. The model also produces near-point estimates of \$0.05 per dollar of tax revenue for the deadweight loss of a marginal tax increase, again similar to the constant coefficients estimate. The results show that the estimated welfare impacts of sales taxes from a constant coefficients model are quite robust to allowing for rich unobserved heterogeneity across markets, even while this heterogeneity produces considerable ambiguity in the underlying elasticities of supply and demand.

The paper is organized as follows. In the next section, we lay out the general model and identification problem, then provide intuition for our approach based on market types. In Section 3, we consider the random coefficients model, including the attenuation under heterogeneity that occurs with linear IV estimators. In Section 4, we develop our approach for smoothing heterogeneity in the linear random coefficients model. Section 5 contains our application to sales taxes. We give some brief concluding remarks in Section 6.

2 Nonparametric Models with Unrestricted Heterogeneity

In this section, we consider a fully nonparametric model that allows for unrestricted unobserved heterogeneity. We show that common IV assumptions still have identifying power for this model, despite its lack of restrictions on functional form or heterogeneity.

2.1 Model and notation

Suppose that demand and supply are given by the structural functions

$$q^{\mathrm{D}}(p, Z, B) \quad \text{and} \quad q^{\mathrm{S}}(p, Z, B), \tag{1}$$

where p is a hypothetical price, Z is a vector of observable covariates and/or instruments, and B is a latent variable of arbitrary dimension that captures unobserved heterogeneity across markets. Quantity and/or price could be measured in levels or logs.

We assume throughout that a unique equilibrium exists for any realization of (Z, B). The researcher observes price-quantity pairs (P, Q) that are assumed to satisfy the equilibrium condition

$$Q = q^{\mathrm{D}}(P, Z, B) = q^{\mathrm{S}}(P, Z, B). \tag{2}$$

The equilibrium price and quantity pairs can also be represented as reduced form functions of (Z, B) that we denote as

$$Q = q^{\mathrm{E}}(Z, B) \quad \text{and} \quad P = p^{\mathrm{E}}(Z, B). \tag{3}$$

The population consists of distinct, separated markets.

Both the structural equations (1) and the reduced form (3) can be expressed with potential outcomes notation (Manski, 1997; Angrist et al., 2000). For any fixed values p and z, let

$$Q^{\mathrm{D}}(p, z) \equiv q^{\mathrm{D}}(p, z, B) \quad \text{and} \quad Q^{\mathrm{S}}(p, z) \equiv q^{\mathrm{S}}(p, z, B) \tag{4}$$

denote the quantities that would be demanded or supplied if the price and instrument were externally set to p and z. Let

$$Q^{\mathrm{E}}(z) \equiv q^{\mathrm{E}}(z, B) \quad \text{and} \quad P^{\mathrm{E}}(z) \equiv p^{\mathrm{E}}(z, B) \tag{5}$$

denote the equilibrium quantity and price that would satisfy equation (2) if the instrument alone were externally set to z. The randomness in both the "structural" potential outcomes $Q^{\mathrm{D}}(p, z), Q^{\mathrm{S}}(p, z)$ and the reduced form potential outcomes $Q^{\mathrm{E}}(z), P^{\mathrm{E}}(z)$ is due to the latent variable B, which is left implicit in potential outcomes notation. In this section, we use the potential outcomes notation without explicit reference to B, which turns out to be more straightforward for a nonparametric analysis. In Section 3, we incorporate B explicitly into a linear random coefficients specification.

Matzkin (2008, 2015) and Berry and Haile (2018) considered a model like (1) under the assumption that B is two-dimensional, with one component entering q^{D} and one entering q^{S}, together with the assumption that both of these functions are strictly increasing in

their respective component of B. This model imposes rank invariance in the structural potential outcomes (see Chernozhukov and Hansen, 2005; Torgovitsky, 2015), which does not allow for the type of flexible unobserved heterogeneity we want to consider. For example, it does not allow for a simple linear random coefficients model, which has four unobservables (two intercepts and two slopes). The benefit of assuming that B is two-dimensional is that it makes it possible to construct an invertible mapping between B and the two-dimensional observable pair (P, Q). Matzkin (2008) and Berry and Haile (2018) used this observation to provide sufficient conditions for point identification. A model with multi-dimensional B does not admit such an invertible mapping, so we begin by confronting the identification problem from first principles.

2.2 The identification problem

The data consists of market-level observations on (P, Q, Z) sampled from a population of markets. Suppose that we knew the population distribution of (P, Q, Z). What could we learn about the supply and demand curves from this knowledge? This is the essential question of identification, the answer to which must form the foundation of any coherent empirical analysis.

The starting point for the identification analysis is the relationship between the observed distribution of price and quantity and the unobserved potential equilibrium outcomes:

$$\mathbb{P}[P \leq p, Q \leq q | Z = z] = \mathbb{P}[P^{\mathrm{E}}(z) \leq p, Q^{\mathrm{E}}(z) \leq q | Z = z]. \tag{6}$$

The left-hand side of (6) is assumed to be known for any p, q, and z. Because $Q = Q^{\mathrm{E}}(Z)$ and $P = P^{\mathrm{E}}(Z)$, the distribution of the potential reduced form outcomes, $(P^{\mathrm{E}}(z), Q^{\mathrm{E}}(z))$, is directly identified conditional on $Z = z$, but not conditional on $Z = z' \neq z$, nor unconditionally.

The distributions of the reduced form potential outcomes, $P^{\mathrm{E}}(z)$ and $Q^{\mathrm{E}}(z)$, will often be insufficient to answer questions of interest. Many questions of interest involve the structural functions, $Q^{\mathrm{S}}(p, z)$ and $Q^{\mathrm{D}}(p, z)$. Simultaneity means that these structural potential outcomes are only indirectly related to the reduced form potential outcomes through the equilibrium condition (2).

Manski (1994, 1995, 1997) showed that if the structural demand function is assumed to slope in the expected downward direction, then non-trivial bounds can be placed on the distribution of $Q^{\mathrm{D}}(p, z)$ at any given (p, z). In particular, Manski (1997, Proposition M1, Corollary M1.1) showed that if $Q^{\mathrm{D}}(p, z)$ is decreasing in p, then (6) implies that

$$\mathbb{P}[Q^{\mathrm{D}}(p, z) \leq q | Z = z] \in \Big[\mathbb{P}[P \leq p, Q \leq q | Z = z],\ \mathbb{P}[P < p \text{ or } Q \leq q | Z = z] \Big]. \tag{7}$$

Manski observed that these bounds are sharp—best possible given the assumptions—and

that they translate immediately into sharp bounds on any feature of the distribution of structural demand that is monotone with respect to first-order stochastic dominance, such as a mean or quantile. Sharp as they are, we will see ahead that they tend to be quite wide in practice, motivating the adoption of additional assumptions.

2.3 Instruments

The classic additional assumption is that Z contains an instrumental variable (IV) that shifts supply but not demand. For notational purposes, we assume for now that Z doesn't contain any other non-instrument covariates, although in principle they could be conditioned on. The baseline IV assumptions can then be stated as follows.

Baseline IV Assumptions.

Exclusion: $Q^{\text{D}}(p, z) = Q^{\text{D}}(p, z') \equiv Q^{\text{D}}(p)$ for all z and z'.

Exogeneity: Z is jointly independent of $\{(Q^{\text{D}}(p), Q^{\text{S}}(p, z))\}_{p,z}$.

The exclusion and exogeneity assumptions are standard for IV models: exclusion says that Z is a supply shifter that has no impact on demand, while exogeneity says that Z is independent of all other latent factors that determine supply and demand. The exogeneity assumption can be stated equivalently using the latent random variable B.

Exogeneity (stated with latent variables): Z is independent of B.

Manski and Pepper (2000, Proposition 2, Corollary 2) showed that the baseline IV assumptions strengthen the bounds in (7) to the intersection bounds

$$\mathbb{P}[Q^{\text{D}}(p) \leq q] \in \Big[\sup_z \mathbb{P}[P \leq p, Q \leq q | Z = z],\ \inf_z \mathbb{P}[P < p \text{ or } Q \leq q | Z = z]\Big]. \tag{8}$$

These bounds are also sharp, although they still tend to be quite wide in practice. In addition, as Manski and Pepper (2000, pg. 1005) acknowledge, because (8) are *pointwise* sharp bounds, they will generally produce non-sharp bounds on target parameters that compare $Q^{\text{D}}(p_0)$ and $Q^{\text{D}}(p_1)$ at two different prices, such as the change in demand between two hypothetical prices. With this pessimistic backdrop as our starting point, we now consider alternative strategies for obtaining more informative conclusions. This involves taking a different approach to the identification analysis.

2.4 Market types

Let $Y \equiv (P, Q)$ denote the equilibrium price and quantity pairs. We begin by focusing on simple cases in which Y has discrete support and $Z \in \{0, 1\}$ is a binary supply shifter. Suppose that the population distribution of $Y|Z = z$ has only two points of support for each value of z. We write these points as $\{y_{0\mathtt{a}}, y_{0\mathtt{b}}\}$ for $z = 0$ and as $\{y_{1\mathtt{a}}, y_{1\mathtt{b}}\}$ for $z = 1$. These are also the support points of the potential equilibrium outcomes,

$Y(z) \equiv (P^{\text{E}}(z), Q^{\text{E}}(z))$. Exogeneity implies that the marginal distribution of $Y(z)$ is identified by the marginal distribution of $Y|Z = z$ for $z = 0, 1$. These two marginal distributions tell us the proportion of markets in equilibria $y_{z\mathsf{a}}$ or $y_{z\mathsf{b}}$, so they tell us something about the distribution of heterogeneity across markets.

Most target parameters will depend on the joint distribution of $Y(0)$ and $Y(1)$. For example, the distribution of the average slope of demand between the two equilibrium prices,

$$\frac{Q^{\text{E}}(1) - Q^{\text{E}}(0)}{P^{\text{E}}(1) - P^{\text{E}}(0)} = \frac{Q^{\text{D}}(P^{\text{E}}(1)) - Q^{\text{D}}(P^{\text{E}}(0))}{P^{\text{E}}(1) - P^{\text{E}}(0)} \equiv \overline{\varepsilon}^{\text{D}}, \tag{9}$$

depends on the joint distribution of $Y(0)$ and $Y(1)$. We think of a joint realization of $M \equiv (Y(0), Y(1))$ as a market type. The support of possible market types is some subset of the set of four points $\{(y_{0\mathsf{a}}, y_{1\mathsf{a}}), (y_{0\mathsf{a}}, y_{1\mathsf{b}}), (y_{0\mathsf{b}}, y_{1\mathsf{b}}), (y_{0\mathsf{b}}, y_{1\mathsf{a}})\}$ produced by taking the Cartesian product of the marginal supports. An object like the average slope of demand (9) can be written more explicitly as a function of the market type: $\overline{\varepsilon}^{\text{D}} \equiv \overline{\varepsilon}^{\text{D}}(M)$.

Economic reasoning can be used to learn about the distribution of market types M from the marginal distributions of $Y(0)$ and $Y(1)$. Consider Figure 1a, which shows one possible configuration of the marginal supports of $Y(0)$ and $Y(1)$. Each of the four possible market types is shown by a line connecting a point in the support of $Y(0)$ with one in the support of $Y(1)$. The line connecting $y_{0\mathsf{b}}$ and $y_{1\mathsf{a}}$ is dashed because it would have to correspond to a value of M with $\overline{\varepsilon}^{\text{D}}(M) > 0$, implying that demand slopes upward. Assuming that the structural demand function is monotonically decreasing means that M must have zero probability of taking the value $(y_{0\mathsf{b}}, y_{1\mathsf{a}})$. In this case, the implication is that the distribution of market types has three points of support and is point identified:

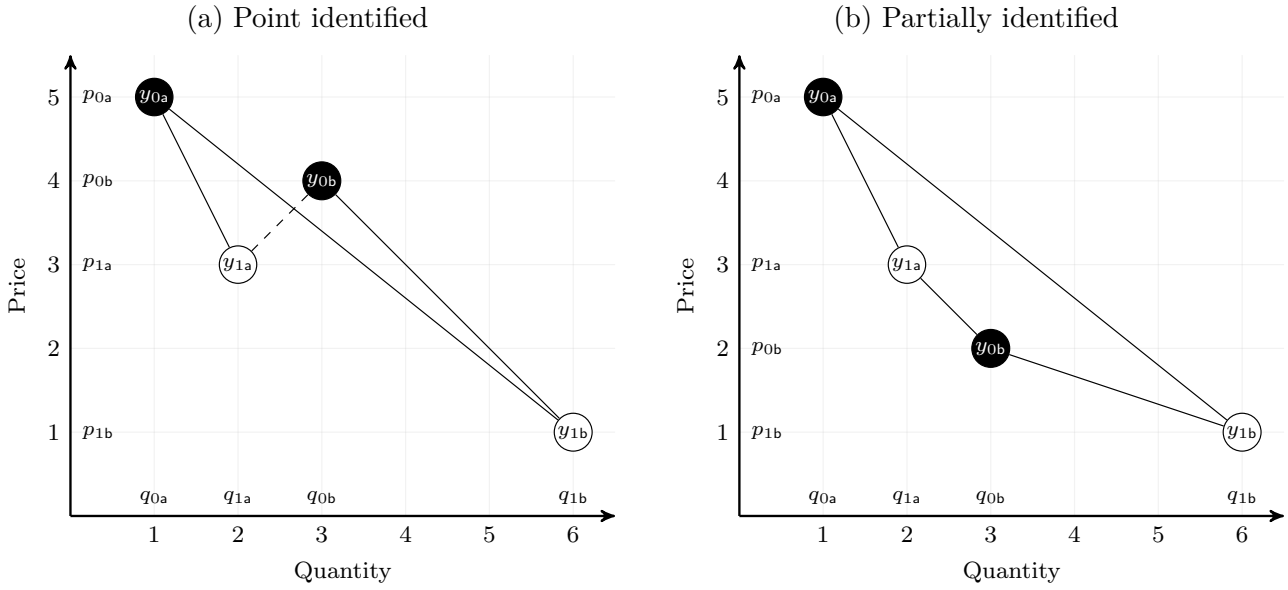
$$\begin{aligned}
&\mathbb{P}[Y = y_{1\mathsf{a}}|Z = 1] = \mathbb{P}[M = (y_{0\mathsf{a}}, y_{1\mathsf{a}})] + \underbrace{\mathbb{P}[M = (y_{0\mathsf{b}}, y_{1\mathsf{a}})]}_{= 0 \text{ (would imply upward-sloping demand)}} \\
\text{and} \quad &\mathbb{P}[Y = y_{0\mathsf{b}}|Z = 0] = \mathbb{P}[M = (y_{0\mathsf{b}}, y_{1\mathsf{b}})] + \overbrace{\mathbb{P}[M = (y_{0\mathsf{b}}, y_{1\mathsf{a}})]} \\
\Rightarrow \quad &\mathbb{P}[M = (y_{0\mathsf{a}}, y_{1\mathsf{b}})] = 1 - \mathbb{P}[Y = y_{0\mathsf{b}}|Z = 0] - \mathbb{P}[Y = y_{1\mathsf{a}}|Z = 1].
\end{aligned} \tag{10}$$

The extent to which this type of reasoning has bite depends on the distribution of the data. In Figure 1b, the assumption of downward-sloping demand doesn't rule out any possible market types. In this case, the distribution of M is only partially identified through the knowledge of its marginals combined with the classic Fréchet-Hoeffding bounds. Taken together, Figure 1 shows that the identified set for the joint distribution of M can have a delicate structure: on the left-hand side it is a singleton, while on the right-hand side it contains nearly all joint distributions compatible with the marginal distributions of $Y|Z = 0$ and $Y|Z = 1$.

This reasoning is easier to work through when the marginal supports of $Y(z)$ each have

Figure 1: Nonparametric identification of the distribution of market types

(a) Point identified (b) Partially identified



Notes: The marginal supports of $Y(0)$ and $Y(1)$ are shown in black and white nodes. Each line connecting two nodes is a potential realization of a market type $M \equiv (Y(0), Y(1))$. Dashed lines indicate market types that must have probability zero due to the assumption that demand slopes down.

two points, but it doesn't depend on that simplification. Figure 2a shows a case in which the supports of $Y(0)$ and $Y(1)$ both have three points. If the marginal distributions of $Y(0)$ and $Y(1)$ are uniform, then the distribution of market types is again point identified because

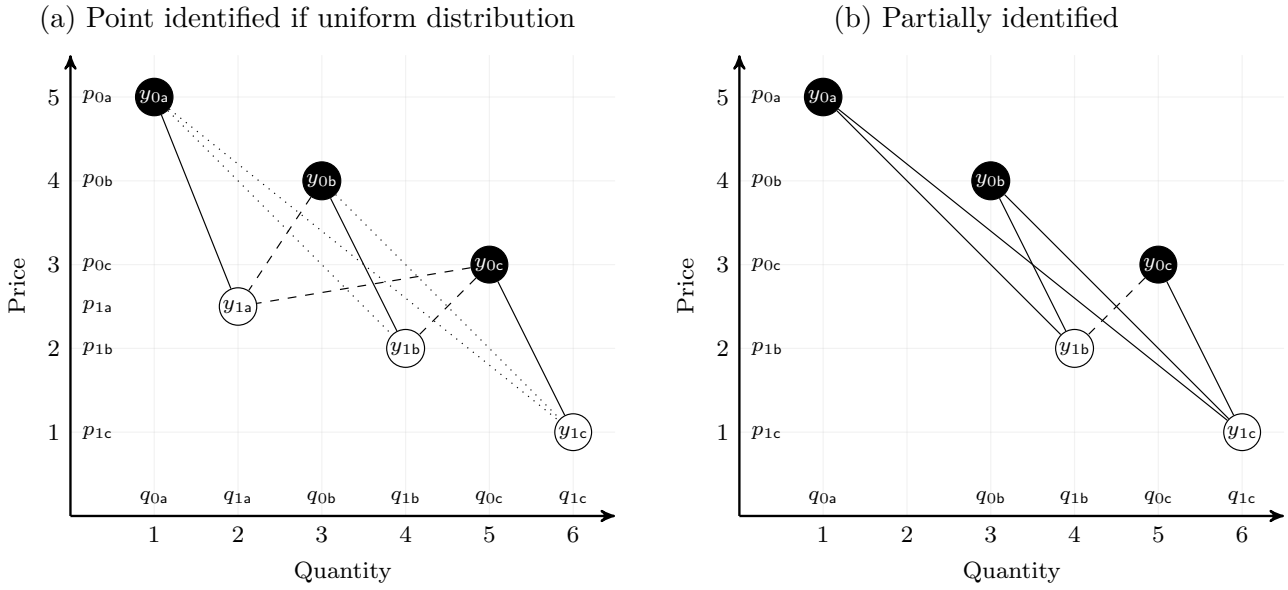
$$\begin{aligned}
& \frac{1}{3} = \mathbb{P}[Y = y_{0\mathsf{c}} \mid Z = 0] \overbrace{= \mathbb{P}[M = (y_{0\mathsf{c}}, y_{1\mathsf{c}})]}^{\text{because } (y_{0\mathsf{c}}, y_{1\mathsf{b}}) \text{ and } (y_{0\mathsf{c}}, y_{1\mathsf{a}}) \text{ imply upward-sloping demand}} \leq \mathbb{P}[Y = y_{1\mathsf{c}} \mid Z = 1] = \frac{1}{3} \\
\Rightarrow \quad & \frac{1}{3} = \mathbb{P}[M = (y_{0\mathsf{c}}, y_{1\mathsf{c}})] \quad \Rightarrow \quad \mathbb{P}[M = (y_{0\mathsf{a}}, y_{1\mathsf{c}})] = 0 \text{ and } \mathbb{P}[M = (y_{0\mathsf{b}}, y_{1\mathsf{c}})] = 0 \\
\Rightarrow \quad & \frac{1}{3} = \mathbb{P}[Y = y_{0\mathsf{b}} \mid Z = 0] = \mathbb{P}[M = (y_{0\mathsf{b}}, y_{1\mathsf{b}})] \leq \mathbb{P}[Y = y_{1\mathsf{b}} \mid Z = 1] = \frac{1}{3} \\
\Rightarrow \quad & \mathbb{P}[M = (y_{0\mathsf{b}}, y_{1\mathsf{b}})] = \frac{1}{3}, \quad \text{so} \quad \mathbb{P}[M = (y_{0\mathsf{a}}, y_{1\mathsf{a}})] = \frac{1}{3}. \qquad (11)
\end{aligned}$$

Figure 2b considers the same marginal distribution for $Y(0)$, but a marginal distribution for $Y(1)$ that has one fewer point, with $y_{1\mathsf{a}}$ omitted. Despite there being fewer possible market types, the distribution of M becomes partially identified, even if the marginals have uniform distributions. This happens because the absence of $y_{1\mathsf{a}}$ makes $\mathbb{P}[Y = y_{1\mathsf{c}} | Z = 1] = 1/2 > 1/3$, which breaks the chain of reasoning in (11).

2.5 Instrument monotonicity

Angrist et al. (2000) augment the baseline IV assumptions with a "monotonicity" condition on how the instrument affects price.

Figure 2: Nonparametric identification with more market types



Notes: See notes for Figure 1. The dotted lines in (a) indicate market types $M \equiv (Y(0), Y(1))$ that can be deduced to have probability zero if the distributions of $Y(0)$ and $Y(1)$ are both uniform.

Instrument monotonicity: $\mathbb{P}[P^{\mathrm{E}}(1) \geq P^{\mathrm{E}}(0)] = 1$ or $\mathbb{P}[P^{\mathrm{E}}(0) \geq P^{\mathrm{E}}(1)] = 1$.

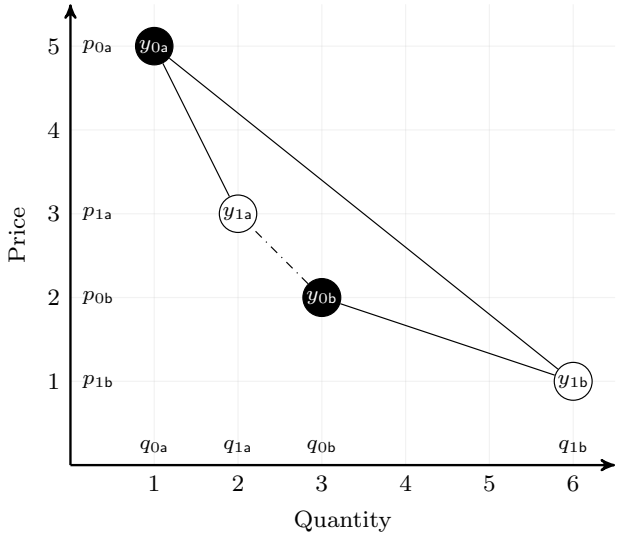
The monotonicity condition says that the instrument shifts equilibrium price in the same direction for all markets. Angrist et al. (2000, Lemma 1) show that if demand and supply slope in the expected direction, and if exclusion is satisfied, then monotonicity is implied by the structural supply curve being monotone in Z.

Monotonicity has different empirical content than downward-sloping demand. For example, monotonicity is satisfied in Figure 1a for all possible market types M, because all of the white nodes for $Z = 0$ lie below the black ones for $Z = 1$. Monotonicity has no identifying power here. Figure 3a modifies Figure 1b to show the opposite case, where downward-sloping demand is satisfied for all possible market types, but monotonicity is not. In this case, imposing the monotonicity assumption point identifies the distribution of market types through a similar argument as used in Figure 1a.

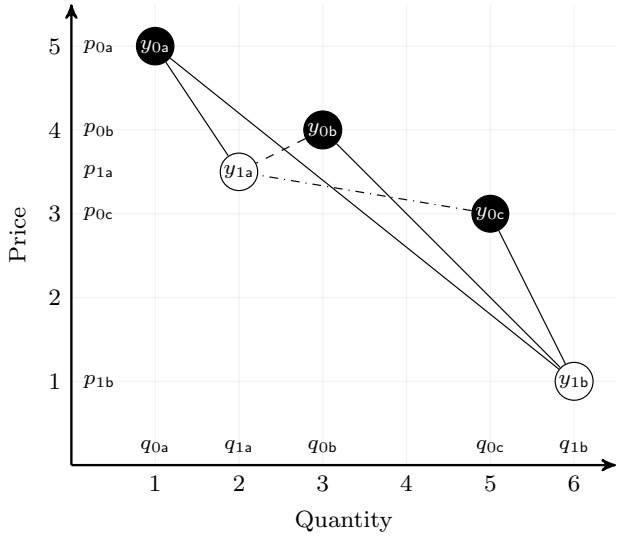
Figure 3b shows an example in which some market types violate downward-sloping demand and others violate monotonicity, assumed to be in the direction $\mathbb{P}[P^{\mathrm{E}}(0) \geq P^{\mathrm{E}}(1)] = 1$. In this case, downward-sloping demand and monotonicity both aid identification. When both are imposed, the distribution of market types is point identified because the assump-

Figure 3: Nonparametric identification with monotonicity

(a) Point identified under monotonicity both with and without downward-sloping demand



(b) Point identified under both monotonicity and downward-sloping demand



Notes: See notes for Figure 1. Dashed lines indicate market types that must have probability zero due to the assumption that demand slopes down. Dot-dashed lines indicate market types that must have probability zero due to the monotonicity condition.

tions leave only four types that could have non-zero probability:

$$\begin{aligned} & \mathbb{P}[M = (y_{0a}, y_{1a})] = \mathbb{P}[Y = y_{1a} | Z = 1] \\ & \mathbb{P}[M = (y_{0b}, y_{1b})] = \mathbb{P}[Y = y_{0b} | Z = 0] \\ \text{and} \quad & \mathbb{P}[M = (y_{0c}, y_{1b})] = \mathbb{P}[Y = y_{0c} | Z = 0] \\ \Rightarrow \quad & \mathbb{P}[M = (y_{0a}, y_{1b})] = \mathbb{P}[Y = y_{0a} | Z = 0] - \mathbb{P}[Y = y_{1a} | Z = 1]. \end{aligned} \tag{12}$$

Point identification would generally be lost in this example if only one of the two assumptions were imposed.

2.6 Tax or subsidy instruments

Zoutman et al. (2018) consider tax or subsidy instruments that affect prices faced by consumers or producers in a known way (see also Dearing, 2022). For example, suppose that producers in markets with $Z = 1$ receive a per-unit subsidy t. A natural assumption, which Zoutman et al. (2018) describe as the "Supply-Side Ramsey Exclusion Restriction," is that producers in a market $Z = 0$ facing a price p would have the same supply as producers in a subsidized market $Z = 1$ facing net-of-subsidy price $p - t$. The assumption can be stated with potential outcomes as follows.

Subsidy instrument: There is a known t such that $Q^{\mathrm{S}}(p, z) = \tilde{Q}^{\mathrm{S}}(p + tz)$ for all p and

z, where $\tilde{Q}^{\text{S}}$ is an increasing function.

The subsidy instrument assumption can be viewed as an index restriction on the structural supply curve. In the classic linear model considered by Zoutman et al. (2018), the subsidy instrument assumption amounts to requiring the coefficients on P and Z in the supply equation to be the same.

Figure 4a shows that the subsidy instrument assumption can also have identifying power without the classic linearity assumption. The supports of $Y|Z=z$ each consist of two points, still shown as black and white nodes. In this case, downward-sloping demand has no identifying power, so the distribution of market types M is partially identified without an additional assumption. Imposing instrument monotonicity would not change this because the instrument shifts all equilibrium prices down.

The points $\tilde{y}_{1\text{a}} \equiv (p_{0\text{a}} - t, q_{0\text{a}})$ and $\tilde{y}_{1\text{b}} \equiv (p_{0\text{b}} - t, q_{0\text{b}})$ shown in blue are additional points on the $Z = 1$ supply curve that can be deduced from the subsidy instrument assumption:

$$Y(0) = y_{0\text{a}} \;\Rightarrow\; Q^{\text{S}}(p_{0\text{a}}, 0) = q_{0\text{a}} \;\Leftrightarrow\; \tilde{Q}^{\text{S}}(p_{0\text{a}}) = q_{0\text{a}} \;\Leftrightarrow\; Q^{\text{S}}(p_{0\text{a}} - t, 1) = q_{0\text{a}}.$$

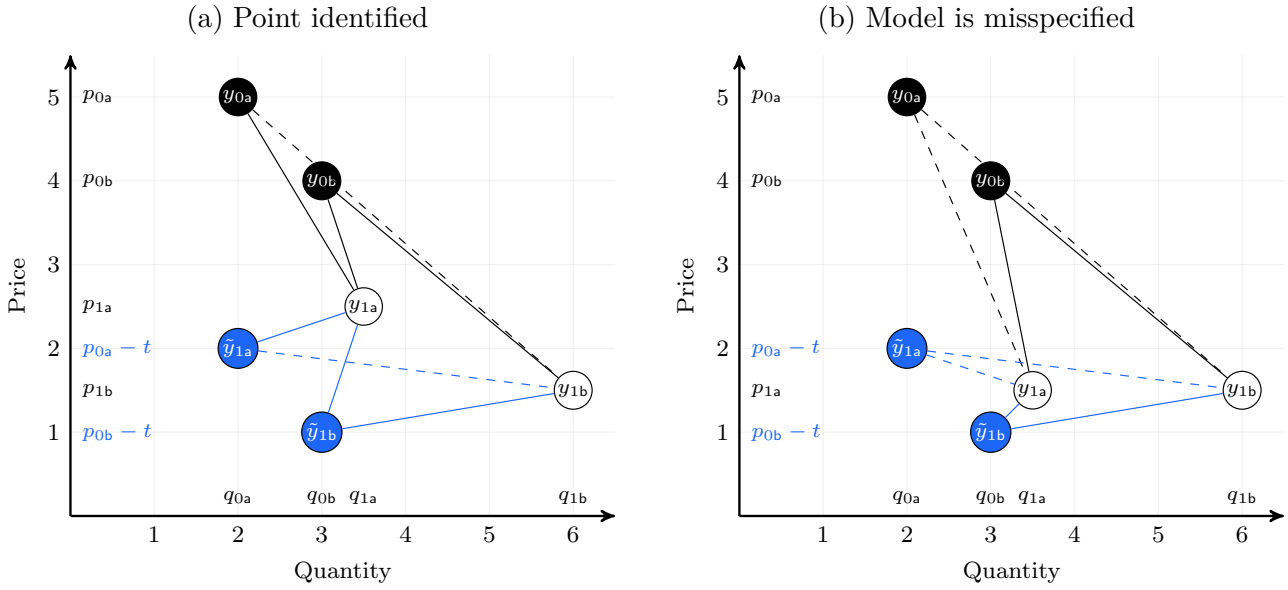
A market type $M = (y_{0\text{a}}, y_{1\text{a}})$ would have $Q^{\text{S}}(p_{0\text{a}} - t, 1) = q_{0\text{a}} < q_{1\text{b}} = Q^{\text{S}}(p_{1\text{b}}, 1)$, implying the downward-sloping supply curve with average slope indicated by the dashed blue line connecting $\tilde{y}_{1\text{a}}$ to $y_{1\text{b}}$. Assuming that supply is upward-sloping requires $\mathbb{P}[M = (y_{0\text{a}}, y_{1\text{b}})] = 0$, leaving only three market types. The distribution of market types is then point identified from its marginals, as in the previous arguments.

Also as with the previous arguments, changing the configuration of the supports of $Y|Z=z$ can change this conclusion. If t were large enough to make $p_{0\text{a}} - t < p_{1\text{b}}$, then upward-sloping supply would not rule out any market types, leaving the distribution of M partially identified. Alternatively, suppose that $p_{1\text{a}}$ were smaller, as shown in Figure 4b. In this case, the point on the $Z = 1$ supply curve that is generated by $y_{0\text{a}}$ would imply downward-sloping supply whether $y_{0\text{a}}$ were paired with $y_{1\text{a}}$ or $y_{1\text{b}}$. There would be no distribution of market types consistent with upward-sloping supply that could also match the marginal distributions of $Y|Z=z$. The identified set would be empty, implying that the model is misspecified.

2.7 Identifying target parameters

Our focus so far has been on identification of the distribution of market types $M \equiv (Y(0), Y(1))$, but our ultimate interest is in identified sets for lower-dimensional target parameters that summarize across market types. For example, the distribution of demand at a hypothetical price, the average change in demand between two hypothetical prices, or the average change in consumer surplus from a marginal tax increase. Target parameters like these depend on the distribution of M, but they typically will not be fully determined

Figure 4: Nonparametric identification with a subsidy instrument



Notes: See notes for Figure 1. The blue nodes represent implied points on the $Z = 1$ curve through the subsidy instrument assumption. Solid blue lines represent upward-sloping supply. Dashed blue lines represent downward-sloping supply. The dashed black lines indicate market pairs that are eliminated through the subsidy instrument assumption.

by it. The reason is that each market type could be consistent with several supply and demand schedules that admit different values of the target parameter.

We formalize this observation by distinguishing between what we call composite and non-composite target parameters.

A non-composite target parameter is one that is fully determined by the distribution of market types. An example is a weighted average (across market types) of the average slope of demand:

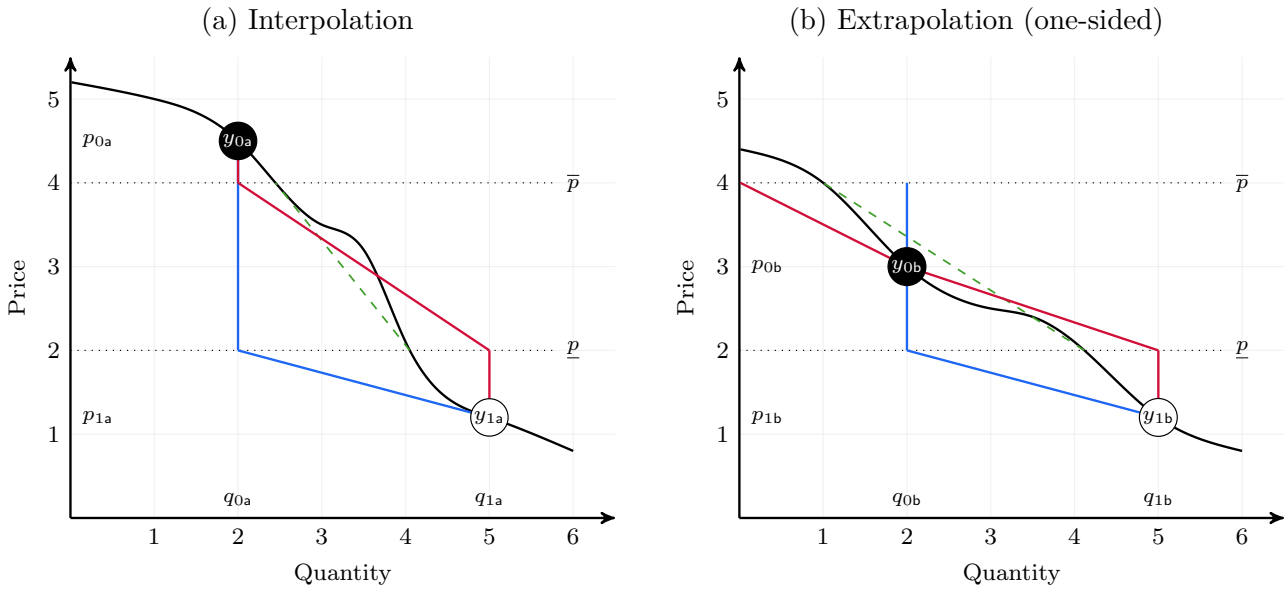
$$\mathbb{E}[\omega(M)\bar{\varepsilon}^{\mathrm{D}}(M)]. \tag{13}$$

Unit weights produce the unconditional average $\bar{\bar{\varepsilon}}^{\mathrm{D}} = \mathbb{E}[\bar{\varepsilon}^{\mathrm{D}}(M)]$. The Wald estimand generated by a linear IV estimator produces weights that are proportional to the change in equilibrium prices:

$$\frac{\mathbb{E}[Q|Z=1] - \mathbb{E}[Q|Z=0]}{\mathbb{E}[P|Z=1] - \mathbb{E}[P|Z=0]} = \mathbb{E}\left[\underbrace{\left(\frac{P^{\mathrm{E}}(1) - P^{\mathrm{E}}(0)}{\mathbb{E}[P^{\mathrm{E}}(1) - P^{\mathrm{E}}(0)]}\right)}_{\omega_{\mathrm{WALD}}(M)} \bar{\varepsilon}^{\mathrm{D}}(M)\right] \equiv \bar{\bar{\varepsilon}}^{\mathrm{D}}_{\mathrm{WALD}}. \tag{14}$$

Both parameters are non-composite because the weighting function ω and average slope of demand function $\bar{\varepsilon}^{\mathrm{D}}$ are both functions of the market types themselves. The only ambiguity is the distribution over market types, M.

Figure 5: Interpolation and extrapolation



Notes: The target parameter is the average slope of demand between two pre-specified prices, $\underline{p}$ and $\overline{p}$. The inverse demand curve is shown in black and the average slope of the inverse demand curve between $\underline{p}$ and $\overline{p}$ is indicated by the dashed green chord. Figure (a) shows a market m_{a} for which this is an interpolation problem because $p_{0\mathsf{a}} > \overline{p} > \underline{p} > p_{1\mathsf{a}}$. The magnitude of the average slope of demand between $\underline{p}$ and $\overline{p}$ for this market type is between zero, indicated by the vertical blue inverse demand curve, and $(q_{0\mathsf{a}} - q_{1\mathsf{a}})/(\overline{p} - \underline{p})$, indicated by the red inverse demand curve. Figure (b) shows a market m_{b} for which this is a one-sided extrapolation problem because $\overline{p} > p_{0\mathsf{b}} > \underline{p} > p_{1\mathsf{b}}$. The average slope of demand could still be zero. The lower bound on the average slope of demand is $-q_{1\mathsf{b}}/(\overline{p} - \underline{p})$, attained by the red curve.

A composite target parameter has a second layer of ambiguity concerning the values it can take for each market type. This ambiguity derives from the definition of a market type, which only encodes its potential equilibrium outcomes $Y(z) \equiv (P^{\mathrm{E}}(z), Q^{\mathrm{E}}(z))$. For example, consider a target parameter that measures the average slope of demand between two hypothetical prices $\underline{p} < \overline{p}$:

$$\overline{\overline{\varepsilon}}^{\mathrm{D}}_{\underline{p},\overline{p}} \equiv \mathbb{E}\left[\frac{Q^{\mathrm{D}}(\overline{p}) - Q^{\mathrm{D}}(\underline{p})}{\overline{p} - \underline{p}}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Q^{\mathrm{D}}(\overline{p}) - Q^{\mathrm{D}}(\underline{p})}{\overline{p} - \underline{p}}\middle| M\right]\right] \equiv \mathbb{E}\left[\overline{\varepsilon}^{\mathrm{D}}_{\underline{p},\overline{p}}(M)\right]. \tag{15}$$

This is a composite target parameter because a given market type m is consistent with multiple values of $\overline{\varepsilon}^{\mathrm{D}}_{\underline{p},\overline{p}}(m)$, creating a second layer of ambiguity on top of the unknown distribution of market types.[1]

We can characterize this ambiguity by solving a composite problem that finds bounds for each market type. If $m_{\mathsf{a}} = ((p_{0\mathsf{a}}, q_{0\mathsf{a}}), (p_{1\mathsf{a}}, q_{1\mathsf{a}}))$ has $p_{1\mathsf{a}} < \underline{p} < \overline{p} < p_{0\mathsf{a}}$, as shown

[1] Kang and Vasserman (2025) show how to derive bounds on welfare parameters from knowledge of two equilibria for the same market. This is an example of resolving the composite layer of ambiguity for a given market type. The fundamental problem that we focus on is that a researcher only observes a single equilibrium for each market.

in Figure 5a, then the assumption of downward-sloping demand implies that $Q^{\mathrm{D}}(p_{0\mathsf{a}}) \leq Q^{\mathrm{D}}(\overline{p}) \leq Q^{\mathrm{D}}(\underline{p}) \leq Q^{\mathrm{D}}(p_{1\mathsf{a}})$. These inequalities imply

$$0 \geq \overline{\varepsilon}^{\mathrm{D}}_{\underline{p},\overline{p}}(m_{\mathsf{a}}) \geq \frac{q_{0\mathsf{a}} - q_{1\mathsf{a}}}{\overline{p} - \underline{p}} = \left(\frac{p_{0\mathsf{a}} - p_{1\mathsf{a}}}{\overline{p} - \underline{p}}\right) \overline{\varepsilon}^{\mathrm{D}}(m_{\mathsf{a}}). \tag{16}$$

On the other hand, for a market type m_{b} that is the same except with $\overline{p} > p_{0\mathsf{b}}$, the problem requires extrapolating to $\overline{p}$ while interpolating to $\underline{p}$, as depicted in Figure 5b. For m_{b}, the most we can conclude is that

$$0 \geq \overline{\varepsilon}^{\mathrm{D}}_{\underline{p},\overline{p}}(m_{\mathsf{b}}) \geq \frac{-q_{1\mathsf{b}}}{\overline{p} - \underline{p}}. \tag{17}$$

The lower bound becomes less informative because we cannot rule out the possibility that demand drops to zero at $\overline{p}$. For a third market type m_{c} with both $\underline{p} < p_{1\mathsf{c}}$ and $\overline{p} > p_{0\mathsf{c}}$, we would need to extrapolate to both $\underline{p}$ and $\overline{p}$. Assuming that demand slopes downward by itself doesn't provide any discipline on the magnitude of the average slope of demand between $\underline{p}$ and $\overline{p}$ for such a market.

2.8 Computing bounds

Deriving analytic bounds by systematically applying the type of reasoning laid out in the previous sections would be complicated in typical empirical settings in which there are many observed equilibria. This is true even for simple, non-composite target parameters like $\overline{\overline{\varepsilon}}^{\mathrm{D}}$. Composite target parameters make the problem even more complex. In this section, we develop a computational approach for computing sharp nonparametric bounds that can be applied in more typical settings.

We begin with an abstract definition of the identified set. The model is about the structural supply and demand *functions* across all prices. In our notation, these are the random functions Q^{D} and Q^{S}, which collect random variables $Q^{\mathrm{D}}(p)$, $Q^{\mathrm{S}}(p, z)$ across evaluation points (p, z). We use bold font $\boldsymbol{Q} \equiv (Q^{\mathrm{D}}, Q^{\mathrm{S}})$ to denote the pair of structural functions. This pair is required to live in some admissible set $\boldsymbol{\mathcal{Q}}^{\dagger}$ that encodes the researcher's assumptions, such as downward-sloping demand or instrument monotonicity. The exclusion restriction, which we always maintain, is embedded in the definition of Q^{D}.

The primitive parameter in the model is a distribution F over $\boldsymbol{Q}$.[2] This distribution must also satisfy the researcher's assumptions by living in a set $\mathcal{F}^{\dagger}$. For now, we assume that $\mathcal{F}^{\dagger} \equiv \{F : \mathbb{P}_F[\boldsymbol{Q} \in \boldsymbol{\mathcal{Q}}^{\dagger}] = 1\}$ is simply the set of distributions that put all of their mass on admissible demand and supply functions. Later, in Section 4, we will consider models in which F is further restricted.

[2]Throughout the following, we ignore technical distractions about measurability, which might become more salient if viewing Q^{D} and Q^{S} as infinite-dimensional objects defined over a continuum of prices. The concerned reader can assume that Q^{D} and Q^{S} are defined over a finite set of prices.

Every realization of $\boldsymbol{Q} \equiv (Q^{\mathrm{D}}, Q^{\mathrm{S}})$ produces realizations of potential equilibria $Y(z)$ for every instrument value z. A distribution F over $\boldsymbol{Q}$ therefore implies a distribution of observed equilibria Y conditional on any value of Z. The identified set for F is the subset of $\mathcal{F}^\dagger$ where this implied distribution matches the population distribution:

$$\mathcal{F}^\star \equiv \left\{ F \in \mathcal{F}^\dagger : \mathbb{P}_F[Y(z) = y] = \mathbb{P}[Y = y | Z = z] \quad \text{for all } y \text{ and } z \right\}.$$

The instrument exogeneity assumption, which we always maintain, is embedded in the definition of $\mathcal{F}^\star$.

The researcher's object of interest is a lower-dimensional target parameter that takes the form $\mathbb{E}_F[\tau(\boldsymbol{Q})]$, where τ is some lower-dimensional function of the structural supply and/or demand curves. The identified set for the target parameter is

$$\mathcal{T}^\star \equiv \{ \mathbb{E}_F[\tau(\boldsymbol{Q})] : F \in \mathcal{F}^\star \}.$$

The goal is to characterize $\mathcal{T}^\star$, or at least its extremal points, $\tau^\star_{\text{lb}} \equiv \inf \mathcal{T}^\star$ and $\tau^\star_{\text{ub}} \equiv \sup \mathcal{T}^\star$, which constitute the sharp bounds on the target parameter. For example, τ could be the average slope of demand between two hypothetical prices, the level of demand at a given price, or the deadweight loss created by a sales tax. Then $\mathbb{E}_F[\tau(\boldsymbol{Q})]$ is the average of $\tau(\boldsymbol{Q})$ over the distribution of supply and demand functions.

The distribution F describes a probability over two functions, so it is a high-dimensional object that is difficult to work with directly without further restrictions. In this section, we avoid placing further restrictions by working with distributions π for the lower-dimensional market type $M \equiv \{Y(z)\}_z$, which we assume has finite support. It turns out that calculations with these lower-dimensional distributions π are sufficient for calculating the sharp bounds on target parameters that depend on F in many interesting cases. Characterizing these cases requires thinking about the mapping from realizations of the structural functions, $\boldsymbol{q}$, to realizations of the market type vector of equilibrium outcomes, m. We denote this mapping by $\mu : \boldsymbol{\mathcal{Q}}^\dagger \to \mathcal{M}$, where $\mathcal{M} \equiv \otimes_z \mathcal{Y}_z$ is the set of potential market types, and $\mathcal{Y}_z$ is the support of $Y|Z = z$. Each F then generates a π defined as $\pi_F(m) \equiv \mathbb{P}_F[\mu(\boldsymbol{Q}) = m]$.

Let Π denote the set of all probability mass functions over market types, that is, the set of all functions $\pi : \mathcal{M} \to [0, 1]$ such that $\sum_{m \in \mathcal{M}} \pi(m) = 1$. Let $\Pi^\dagger \equiv \{\pi_F \in \Pi : F \in \mathcal{F}^\dagger\}$ denote the subset of Π generated by an admissible distribution over structural functions, $F \in \mathcal{F}^\dagger$. Because we have defined $\mathcal{F}^\dagger$ purely in terms of the set of admissible structural functions, $\boldsymbol{\mathcal{Q}}^\dagger$, we can describe $\Pi^\dagger$ by determining the subset of $\mathcal{M}$ that could be produced by some element of $\boldsymbol{\mathcal{Q}}^\dagger$:

$$\mathcal{M}^\dagger \equiv \{m \in \mathcal{M} : \mu(\boldsymbol{q}) = m \text{ for some } \boldsymbol{q} \in \boldsymbol{\mathcal{Q}}^\dagger\}.$$

This is essentially the task that has been illustrated through the preceding graphical examples. The following proposition formalizes this observation.

Proposition 1. Suppose that $\mathcal{F}^\dagger \equiv \{F : \mathbb{P}_F[\boldsymbol{Q} \in \boldsymbol{\mathcal{Q}}^\dagger] = 1\}$. Then

$$\Pi^\dagger = \left\{\pi \in \Pi : \pi(m) = 0 \text{ for all } m \notin \mathcal{M}^\dagger\right\}.$$

The final ingredient to computing bounds with π involves determining bounds on $\tau(\boldsymbol{q})$ across all structural function realizations $\boldsymbol{q}$ that produce a given market type m. That is,

$$\underline{t}(m)/\overline{t}(m) \equiv \inf/\sup_{\boldsymbol{q} \in \boldsymbol{\mathcal{Q}}^\dagger} \tau(\boldsymbol{q}) \quad \text{s.t.} \quad \mu(\boldsymbol{q}) = m. \tag{18}$$

For non-composite target parameters, such as the average slope of demand between equilibrium prices, we can deduce $\underline{t}(m) = \overline{t}(m)$ directly for each market type m. In the previous section, we saw that the average slope of demand between hypothetical prices was a composite target parameter with $\underline{t}(m) < \overline{t}(m)$, but we were able to solve for $\underline{t}(m)$ and $\overline{t}(m)$ through inspection and economic reasoning. In Section 3, we will consider cases in which $\boldsymbol{\mathcal{Q}}^\dagger$ has a parametric structure that makes it tricky to solve for $\underline{t}(m)$ and $\overline{t}(m)$ analytically, but straightforward to solve for them numerically.

Once the market type bounds in (18) have been computed, the following proposition can then be used to construct sharp bounds on $\mathbb{E}_F[\tau(\boldsymbol{Q})]$.

Proposition 2. For any $m \in \mathcal{M} \equiv \otimes_z \mathcal{Y}_z$, let m_z denote the component of m corresponding to $\mathcal{Y}_z$. Define

$$\begin{aligned} \overline{t}^\star \equiv \max_{\pi \in \Pi} \quad & \sum_{m \in \mathcal{M}} \overline{t}(m)\pi(m) \\ \text{s.t.} \quad & \mathbb{P}[Y = y | Z = z] = \sum_{m \in \mathcal{M}} \mathbb{1}[m_z = y]\pi(m) \quad \text{for all } y \in \mathcal{Y}_z \text{ and } z = 0, 1 \\ & \pi(m) = 0 \quad \text{for all } m \notin \mathcal{M}^\dagger, \end{aligned} \tag{NP-LP}$$

Then $\overline{t}^\star = \tau^\star_{\text{ub}}$. Similarly, if $\underline{t}^\star$ is the optimal value of the analogous minimization problem, then $\underline{t}^\star = \tau^\star_{\text{lb}}$.[3]

Proposition 2 shows that we can compute sharp nonparametric bounds on many target parameters by solving linear programs. This type of characterization of an identified set through linear programming has been used by many authors dating back to Balke and Pearl (1994) and Hansen et al. (1995). More recent applications include Lafférs (2013), Freyberger and Horowitz (2015), Mogstad et al. (2018), and Torgovitsky (2019a). The sharpness argument in the proof of Proposition 2 involves using an optimal solution of

[3]We use the usual convention that the maximum or supremum over an empty set is $-\infty$ and the minimum or infimum over an empty set is $+\infty$. So, if (NP-LP) is infeasible, the proposition implies that $\mathcal{T}^\star$ and $\mathcal{F}^\star$ are empty.

(NP-LP) to construct an $F \in \mathcal{F}^\star$ that yields the same value of the target parameter, similar to arguments used in Torgovitsky (2019b) and Tebaldi et al. (2023).

2.9 Nonparametric bounds in the Fulton fish market

In this section, we apply Proposition 2 to the Fulton fish market data assembled by Graddy (1995) and used by Angrist et al. (2000). The data consists of 111 daily observations of the quantity and transacted price of a low-quality white fish. Price and quantity are measured in logs, so that $\bar{\varepsilon}^{\text{D}}(m)$ can be interpreted as approximately the elasticity of demand for market type m. The instrument is a binary indicator for stormy weather at sea, which Angrist et al. (2000) argue is an excluded and exogenous supply shifter.

Table 1 reports bounds on $\overline{\bar{\varepsilon}}^{\text{D}} \equiv \mathbb{E}[\bar{\varepsilon}^{\text{D}}(M)]$, the average across market types of the average slope of demand between equilibrium prices. The row labeled empirical distribution shows bounds under the baseline IV assumptions together with three additional sets of assumptions. With only the baseline IV assumptions, the average slope of demand could be anywhere from strongly downward-sloping to strongly upward-sloping. Requiring demand to slope downward rules out the latter possibility and allows one to conclude that demand is anywhere from mildly inelastic ($\overline{\bar{\varepsilon}}^{\text{D}} = -0.62$) to implausibly elastic ($\overline{\bar{\varepsilon}}^{\text{D}} = -69.69$). Instrument monotonicity by itself yields a tighter lower bound, but allows for the possibility of upward-sloping demand. Column (4) shows that the empirical distribution cannot be matched exactly by any π that satisfies both downward-sloping demand and instrument monotonicity; the constraint set of (NP-LP) is empty. This might represent a rejection of the model, but it could also be due to statistical noise.

In the second row of Table 1, we shut down the statistical noise channel by solving for a π that satisfies both assumptions and comes as close as possible to matching the observed distribution. We then compute bounds using the distribution of observables implied by this π, rather than the empirical distribution. This makes the constraint set of (NP-LP) non-empty by construction because it contains the best-fitting π. Columns (1)–(3) show that bounds under this DGP are quite close to those under the empirical distribution. Column (4) shows that imposing both monotonicity and downward-sloping demand leads to bounds that are substantially tighter than imposing either assumption separately. The bounds rule out inelastic demand ($\overline{\bar{\varepsilon}}^{\text{D}} \leq -1.32$), but still allow for demand to be implausibly elastic. Wide as they are, these nonparametric bounds still don't contain the Wald estimand, $\overline{\bar{\varepsilon}}^{\text{D}}_{\text{WALD}}$, which is only -1.22.[4] This implies, in particular, that the slope

[4] It is easy to see that the identified set for the average slope of demand need not contain the Wald estimand in general. For example, consider Figure 3a, in which point identification was attained under instrument monotonicity. Suppose that each marginal distribution is uniform, so that $\mathbb{P}[Y = y_{0\text{a}}|Z = 0] = \mathbb{P}[Y = y_{1\text{a}}|Z = 1] = 1/2$. Then two potential market types have probability zero because

$$\mathbb{P}[M = (y_{0\text{a}}, y_{1\text{b}})] = \mathbb{P}[Y(0) = y_{0\text{a}}] - \mathbb{P}[M = (y_{0\text{a}}, y_{1\text{a}})] = \mathbb{P}[Y(0) = y_{0\text{a}}] - \mathbb{P}[Y(1) = y_{1\text{a}}] = 0.$$

The two remaining market types have equal probability and average (inverse) slopes of $\bar{\varepsilon}^{\text{D}}((y_{0\text{a}}, y_{1\text{a}})) = -1/2$

Table 1: Nonparametric bounds in the Fulton fish market

	(1)	(2)	(3)	(4)	Wald
Empirical distribution	$[-70.89,\ 66.52]$	$[-70.84,\ -0.63]$	$[-55.22,\ 35.69]$	$\emptyset$	-1.08
Fit distribution	$[-69.75,\ 66.78]$	$[-69.69,\ -0.62]$	$[-56.60,\ 34.86]$	$[-50.51,\ -1.32]$	-1.22
Instrument monotonicity			✓	✓	
Downward-sloping demand		✓		✓	

Notes: Bounds are for the average across markets of the average slope of demand, $\overline{\overline{\varepsilon}}^{D}$. The instrument is a binary indicator for stormy weather at sea, which is one for 32 of the 111 observations. The fit distribution is constructed by finding the π that minimizes the squared deviation of the data-matching constraint in (NP-LP)*. We assume in all cases that market types with equal prices but unequal quantities have probability zero, as this would imply a violation of exclusion and would render the average slope of demand undefined. This rules out two of the $(111-32)\times 32 = 2528$ possible market types in the Fulton fish market data. We also assume that market types with equal quantities have probability zero, as this would imply perfectly inelastic demand and would allow for non-existence of equilibria if supply were also perfectly inelastic. This eliminates one possible market in the Fulton data.*

of demand cannot be constant across markets.

3 Linear Models with Unrestricted Heterogeneity

In this section, we assume that one or both of the structural functions are linear, while still allowing for unrestricted heterogeneity in these linear functions across markets.

3.1 Linear demand

Return to (1) but now parameterize the structural demand function as

$$q^{\mathrm{D}}(p, B) = B_1^{\mathrm{D}} - B_{\mathrm{P}}^{\mathrm{D}} p, \tag{19}$$

where the unobservable B contains the random coefficients B_1^{D} and $B_{\mathrm{P}}^{\mathrm{D}}$, as well as potentially other components that determine supply. We continue to assume that q^{D} is weakly decreasing in price, which now amounts to the assumption that $B_{\mathrm{P}}^{\mathrm{D}} \geq 0$ with probability one. The reduced form functions $Q^{\mathrm{E}}(z) \equiv q^{\mathrm{E}}(z, B)$ and $P^{\mathrm{E}}(z) \equiv p^{\mathrm{E}}(z, B)$ in (3) still exist. The question is what empirical content is provided by the linearity in (19). The answer depends on both the target parameter under consideration and the number of values that the instrument takes.

If the instrument is binary, then linearity provides no additional information about the distribution of market types $M \equiv (P^{\mathrm{E}}(0), Q^{\mathrm{E}}(0), P^{\mathrm{E}}(1), Q^{\mathrm{E}}(1))$. Linear demand in any given market type can be rationalized by a line connecting $(P^{\mathrm{E}}(0), Q^{\mathrm{E}}(0))$ and $(P^{\mathrm{E}}(1), Q^{\mathrm{E}}(1))$.

and $\overline{\varepsilon}^{\mathrm{D}}((y_{0\mathtt{b}}, y_{1\mathtt{b}})) = -3$. So, $\overline{\overline{\varepsilon}}^{\mathrm{D}} = -3\times 1/2 - 1/2\times 1/2 = -7/4$. The Wald estimand, by contrast, places twice as much weight on the more inelastic market, $(y_{0\mathtt{a}}, y_{1\mathtt{a}})$, because that market experiences a larger price change from shifts in the instrument. This produces the inelastic summary measure $\overline{\overline{\varepsilon}}^{\mathrm{D}}_{\mathrm{WALD}} = -3\times 1/3 - 1/2\times 2/3 = -4/3$. The Wald estimand lies outside of the (singleton) identified set of the average slope of demand. We thank Josh Angrist for alerting us to the possibility that this point may not be obvious to all readers.

This implies that the identified set $\mathcal{F}^\star$ of the distribution F of market types M remains the same under the assumption of a linear demand curve, because linear demand does not change $\mathcal{M}^\dagger$. The identified set of any non-composite target parameter like $\overline{\overline{\varepsilon}}^{\mathrm{D}}$ also remains the same.

If the instrument has more than two points of support, then linearity starts to provide identifying content for the distribution of market types and so also for non-composite target parameters like $\overline{\overline{\varepsilon}}^{\mathrm{D}}$. Figure 6a illustrates a case in which downward-sloping demand together with (19) yield point identification of the joint distribution of market types $M \equiv (Y(0), Y(1), Y(2))$, where $Y(z) \equiv (P^{\mathrm{E}}(z), Q^{\mathrm{E}}(z))$. Linearity implies that, for example, both $M = (y_{0\mathsf{a}}, y_{1\mathsf{a}}, y_{2\mathsf{b}})$ and $M = (y_{0\mathsf{b}}, y_{1\mathsf{a}}, y_{2\mathsf{b}})$ have zero probability, because the slope between $y_{1\mathsf{a}}$ and $y_{2\mathsf{b}}$ is different from the slopes between either $y_{0\mathsf{a}}$ and $y_{1\mathsf{a}}$ or $y_{0\mathsf{b}}$ and $y_{1\mathsf{a}}$. Only two market types have constant slopes connecting them: $(y_{0\mathsf{a}}, y_{1\mathsf{a}}, y_{2\mathsf{a}})$ and $(y_{0\mathsf{b}}, y_{1\mathsf{b}}, y_{2\mathsf{b}})$. This example is also one in which the joint distribution of market types M would be partially identified under only downward-sloping demand, with or without instrument monotonicity, because these assumptions are satisfied for all eight possible market types.

Figure 6b provides an example of how partial identification can remain under linear, downward-sloping demand, even with multiple instrument values. Linear demand requires the slopes connecting any potential market type to be constant, meaning that they must lie on a straight line, which also must be decreasing under downward-sloping demand. The configuration shown in Figure 6b has seven market types that are consistent with downward-sloping demand. Mass can be placed on these seven types in multiple different ways that still match the distributions of $Y|Z = z$. For example, suppose that the distributions of $Y|Z = z$ are each uniform, placing probability $1/3$ on each of the points in their supports. One distribution of market types that matches these marginal distributions places equal weight on the types represented by blue lines:

$$\mathbb{P}[M = (y_{0\mathsf{a}}, y_{1\mathsf{a}}, y_{2\mathsf{b}})] = \mathbb{P}[M = (y_{0\mathsf{b}}, y_{1\mathsf{c}}, y_{2\mathsf{c}})] = \mathbb{P}[M = (y_{0\mathsf{c}}, y_{1\mathsf{b}}, y_{2\mathsf{a}})] = \frac{1}{3}.$$

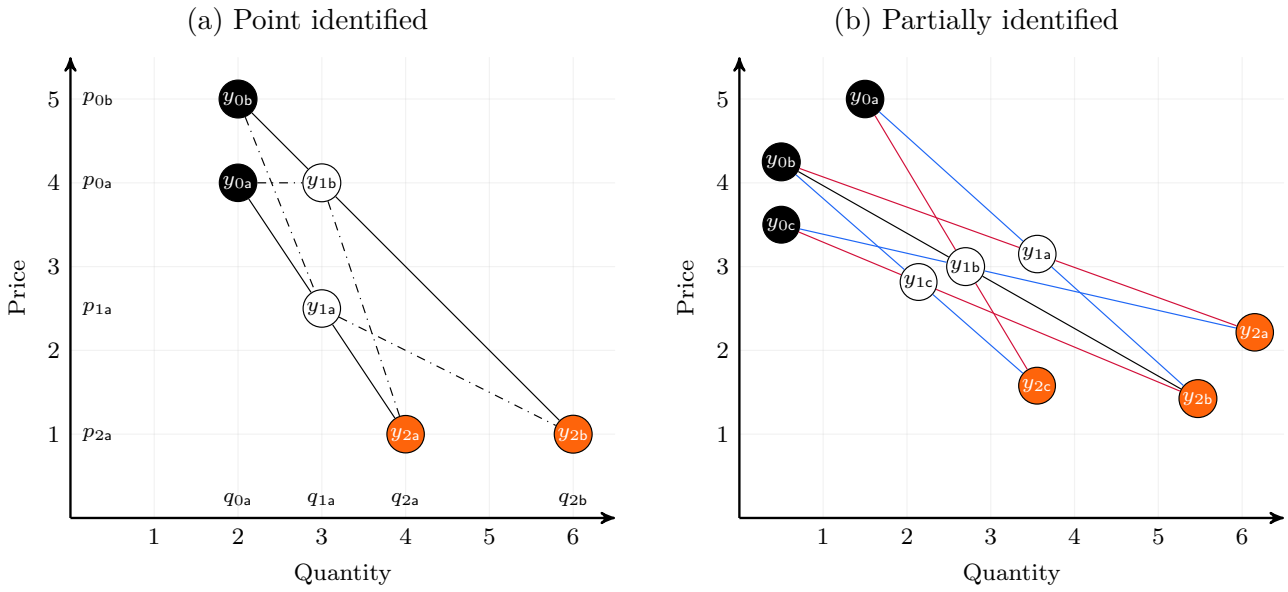
An alternative distribution of market types places equal weight on the types connected by red lines:

$$\mathbb{P}[M = (y_{0\mathsf{a}}, y_{1\mathsf{b}}, y_{2\mathsf{c}})] = \mathbb{P}[M = (y_{0\mathsf{b}}, y_{1\mathsf{a}}, y_{2\mathsf{a}})] = \mathbb{P}[M = (y_{0\mathsf{c}}, y_{1\mathsf{c}}, y_{2\mathsf{b}})] = \frac{1}{3}.$$

Other distributions of market types that match the distributions of $Y|Z = z$ can be generated by subtracting mass from the blue types and adding it to the red types.

If the target parameter is composite, so that $\underline{t}(m) < \overline{t}(m)$ for some type m, then linearity can still have identifying content even if the instrument is binary. For example, consider bounds on average demand at a hypothetical price, $\mathbb{E}[Q^{\mathrm{D}}(p)]$, which fits into the framework of Section 2.8 with $\tau(\boldsymbol{Q}) = Q^{\mathrm{D}}(p)$, where the admissible set $\boldsymbol{\mathcal{Q}}^\dagger$ now incorpo-

Figure 6: The identifying power of linear demand



Notes: See notes for Figure 1. The orange nodes indicate points of the support of $Y|Z = 2$. On the left-hand side, dot-dashed lines indicate pairs that must have probability zero if demand is linear; any market type (a triple) that includes these pairs must also have probability zero. These points are not drawn on the right-hand side to avoid clutter. The colored edges on the right-hand side indicate two collections of triples that can jointly rationalize the observed distribution when it is uniform.

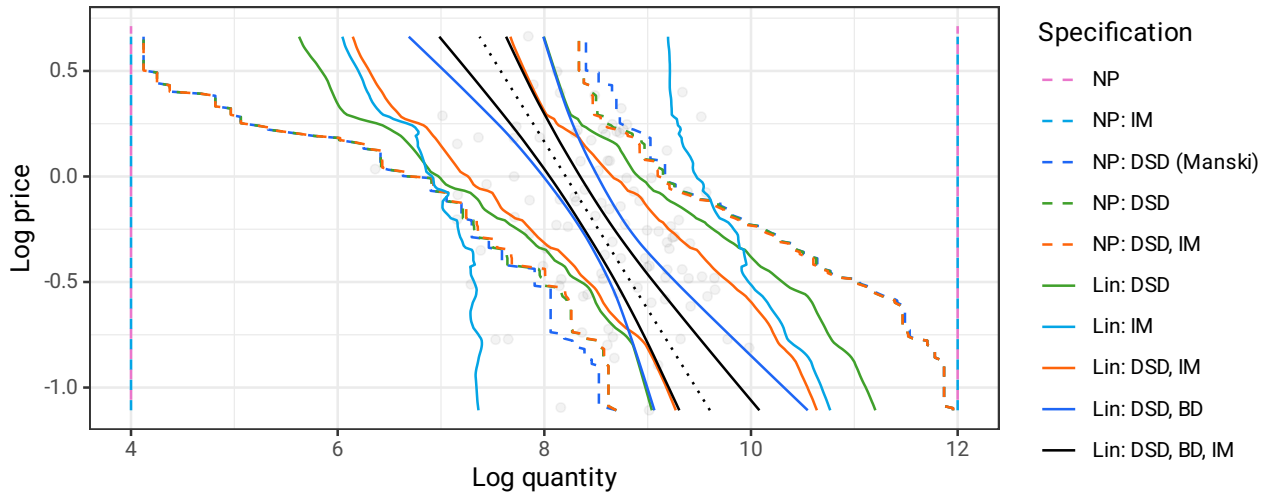
rates linearity. Linear demand implies that for market type $m_{\mathsf{a}} \equiv ((p_{0\mathsf{a}}, q_{0\mathsf{a}}), (p_{1\mathsf{a}}, q_{1\mathsf{a}}))$, average demand satisfies

$$\begin{aligned}&\mathbb{E}[\tau(\boldsymbol{Q})|\mu(\boldsymbol{Q}) = m_{\mathsf{a}}] \\ &\quad = \mathbb{E}[B_1^{\mathrm{D}}|M = m_{\mathsf{a}}] - \mathbb{E}[B_{\mathrm{P}}^{\mathrm{D}}|M = m_{\mathsf{a}}]p = q_{0\mathsf{a}} + \underbrace{\left(\frac{q_{1\mathsf{a}} - q_{0\mathsf{a}}}{p_{1\mathsf{a}} - p_{0\mathsf{a}}}\right)}_{\equiv \bar{\varepsilon}^{\mathrm{D}}(m_{\mathsf{a}})}(p - p_{0\mathsf{a}}).\end{aligned}$$

This implies that the average slope of demand between any two points—in particular, the average slope of demand $\bar{\varepsilon}^{\mathrm{D}}(m)$ between the two equilibrium prices induced by the instrument—is sufficient to identify average demand at any price. Even so, heterogeneity across markets means the distribution of M might not be point identified, which means that $\mathbb{E}[Q^{\mathrm{D}}(p)]$ will still generally not be point identified, even with linearity.

Figure 7 shows several sets of bounds for average demand, $\mathbb{E}[Q^{\mathrm{D}}(p)]$, using a distribution of observables fit to the Fulton fish market data. The dashed lines are nonparametric bounds. With only the baseline IV assumptions, or with both the baseline IV assumptions and instrument monotonicity, the level of average demand can effectively only be bounded between whatever prior bounds the researcher is willing to impose. In Figure 7, we take these bounds to be $[4, 12]$, which is roughly two orders of magnitude lower than the smallest (6.22) and largest (9.98) values of log quantity observed in the data. Adding

Figure 7: Bounds on average demand in the Fulton fish market



Notes: Bounds on average demand, $\mathbb{E}[Q^D(p)]$*, as a function of* p *under various assumptions. The light gray dots show the original Fulton fish market data. We fit a distribution of observables to this data under the strongest set of assumptions (Lin: DSD, BD, IM). This distribution is then used to construct bounds under all sets of assumptions. The dotted black line shows the true value of* $\mathbb{E}[Q^D(p)]$ *in the data generating process. All bounds impose the baseline IV assumptions (with the Manski bounds only using mean independence) and the prior bounds that* $\mathbb{E}[Q^D(p)] \in [4, 12]$*. Other acronyms and abbreviations: nonparametric (NP), linear (Lin), downward-sloping demand (DSD), instrument monotonicity (IM), bounded demand curve (BD). The bounded demand curve assumption (BD) requires a linear demand curve to stay within* $[4, 12]$ *over the support of log prices in the fit DGP, approximately* $[-1.11, .66]$*.*

downward-sloping demand improves these nonparametric bounds considerably, but they are still quite wide. The nonparametric Manski bounds are mean versions of the intersection bounds (8), which are slightly wider than our bounds because they maintain the weaker mean-independence form of exogeneity on the instrument.

The solid lines in Figure 7 are bounds derived assuming linear demand. Linearity with downward-sloping demand is already tighter than the tightest nonparametric bound. Linearity with instrument monotonicity by itself provides much less information. Linearity with downward-sloping demand and instrument monotonicity ("DSD, IM") starts to provide fairly informative bounds. To tighten them further, we impose the assumption that the entire linear demand curve must be contained within the prior quantity range $[4, 12]$ across the observed support of log prices (roughly $[-1.11, .66]$). The solid dark blue line ("DSD, BD") shows the impact that this has with just downward-sloping demand. The black line adds on instrument monotonicity, resulting in bounds on average demand that are quite informative.

3.2 Linear demand with linear supply

The classical model has both linear demand and linear supply. With random coefficients, this becomes

$$\begin{aligned} q^{\mathrm{D}}(p, Z, B) &= B_1^{\mathrm{D}} - B_{\mathrm{P}}^{\mathrm{D}} p \\ \text{and} \quad q^{\mathrm{S}}(p, Z, B) &= B_1^{\mathrm{S}} + B_{\mathrm{Z}}^{\mathrm{S}} Z + B_{\mathrm{P}}^{\mathrm{S}} p, \end{aligned} \tag{20}$$

where now $B \equiv (B_1^{\mathrm{D}}, B_{\mathrm{P}}^{\mathrm{D}}, B_1^{\mathrm{S}}, B_{\mathrm{Z}}^{\mathrm{S}}, B_{\mathrm{P}}^{\mathrm{S}})$. Assuming that supply is upward-sloping means assuming that $B_{\mathrm{P}}^{\mathrm{S}}$ is non-negative. The reduced form functions can now be characterized analytically as

$$q^{\mathrm{E}}(Z, B) = \frac{B_{\mathrm{P}}^{\mathrm{S}} B_1^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{D}} B_1^{\mathrm{S}} + Z B_{\mathrm{P}}^{\mathrm{D}} B_{\mathrm{Z}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}} \quad \text{and} \quad p^{\mathrm{E}}(Z, B) = \frac{B_1^{\mathrm{D}} - B_1^{\mathrm{S}} - Z B_{\mathrm{Z}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}. \tag{21}$$

These reduced forms imply reduced form potential outcomes that are linear in the instrument, with random coefficients that are nonlinear functions of B.

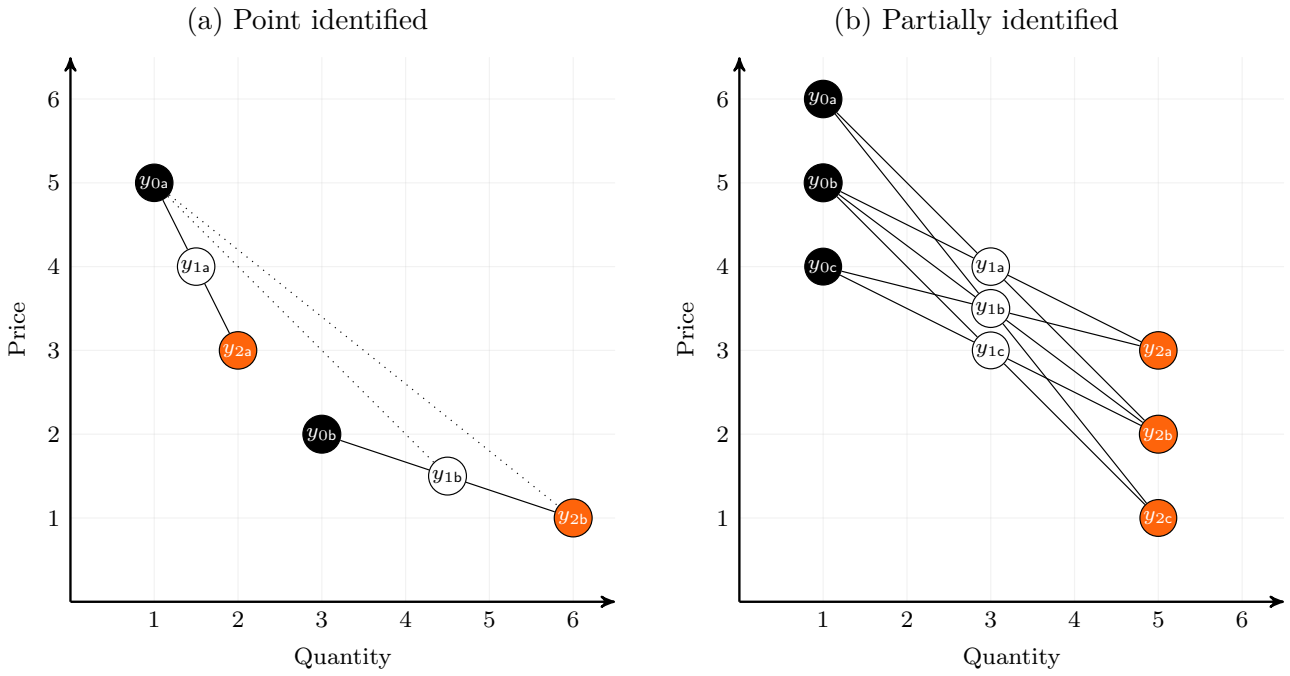
Although interest in a random coefficients model like (20) dates back to Hurwicz (1950), concrete results have only recently been derived by Masten (2018), who provided sufficient conditions for point identification of the marginal distributions of $B_{\mathrm{P}}^{\mathrm{D}}$ and $B_{\mathrm{P}}^{\mathrm{S}}$.[5] Masten's conditions require Z to be continuously distributed (see also Hoderlein et al., 2017, for related results). In addition, Z either needs to vary across the entire real line (have "large support") or an additional tail condition needs to be imposed on the distribution of the random coefficients. Either way, continuous variation in the instrument is essential for Masten's point identification results. Masten (2018, pg. 1199) conjectured that relaxing continuity could lead to informative partial identification.

Our analysis supports Masten's conjecture. If the instrument is binary, then assuming linear supply doesn't have any identifying power for the distribution of market types M. This can be seen from the reduced form (21), which is linear without assumption when Z is binary. As we've seen, the distribution of market types—and therefore of objects like the cross-market average of the average slope of demand, $\overline{\overline{\varepsilon}}^{\mathrm{D}} = \mathbb{E}[B_{\mathrm{P}}^{\mathrm{D}}]$—is partially identified under linear demand if the instrument is binary. It follows that it remains partially identified if both demand and supply are linear and the instrument is binary.

When the instrument takes more values, the linear supply assumption starts to provide

[5] As Masten (2018) explains in a thorough literature review, Hurwicz (1950) pointed to the importance of random coefficients, but did not provide any identification results, while Kelejian (1974) and Hahn (2001) conducted analyses that imposed self-contradictory assumptions. There is a much larger literature on random coefficients models with exogenous variables (for example, Beran and Hall, 1992; Hoderlein et al., 2010; Lewbel and Pendakur, 2017; Gaillac and Gautier, 2022; Hermann and Holzmann, 2024) or with endogenous variables but a triangular structure (for example, Heckman and Vytlacil, 1998; Florens et al., 2008; Masten and Torgovitsky, 2016). A recent contribution to modeling demand along these lines is made by Chernozhukov et al. (2025), who consider a panel random coefficients model of consumer demand with time-invariant individual-specific coefficients, making use of a long panel, but without instruments.

Figure 8: The identifying power of linear reduced forms



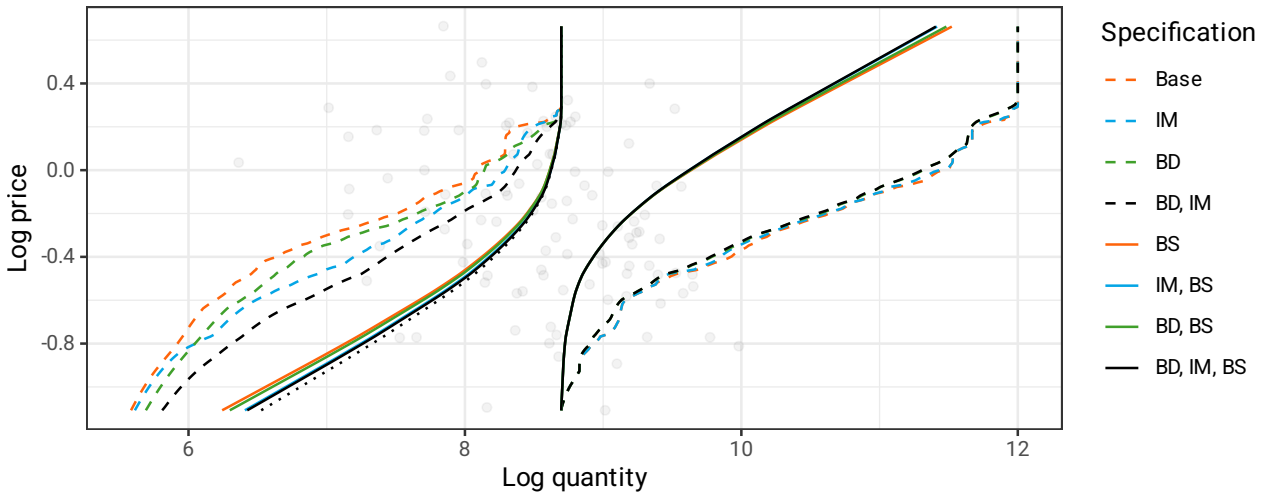
Notes: Solid lines represent markets that are consistent with linear reduced forms. The dotted lines in panel (a) indicate pairs that can be deduced to have joint probability zero because of linearity (only two such pairs are shown).

additional identifying power beyond linear demand. The configuration in Figure 6b, which led to partial identification under only linear demand, is not consistent with linear supply if $Z \in \{0, 1, 2\}$ is interpreted as a cardinal-valued variable. The reason is that the changes in price and quantity between different nodes in each market type triple are not necessarily equal. For example, the market type (y_{0c}, y_{1c}, y_{2b}) has a larger price change between y_{0c} and y_{1c} than between y_{1c} and y_{2b}. This market doesn't have a reduced form that is linear in Z, so it must have probability zero when assuming (21).

Figure 8a shows a distribution that is consistent with a linear reduced form. The point (y_{0a}, y_{1b}, y_{2b}) satisfies downward-sloping demand and is consistent with instrument monotonicity in the direction $P^{E}(0) \geq P^{E}(1) \geq P^{E}(2)$. But the slopes connecting (y_{0a}, y_{1b}) and (y_{1b}, y_{2b}) are different, which would contradict (21). Maintaining linearity implies that the distribution of market types must put zero mass on this type, as well as on several others. In this case, there are only two possible market types that are consistent with linearity, so the distribution of M is point identified.

Figure 8b shows how partial identification can still arise when the distribution of Y has more points of support. Both price and quantity change by the same amount from the black to white to orange nodes of each triple, consistent with a linear reduced form. As in Figure 6b, there are an infinite number of ways to spread mass across the market types while still matching the observed distribution $Y|Z = z$ for each of the three instrument

Figure 9: Bounds on average supply in the Fulton fish market



Notes: Bounds on average supply setting $z = 0$, $\mathbb{E}[Q^s(p,0)]$, as a function of p under various assumptions. Bounds are constructed using the same data generating process as in Figure 7. The true value of $\mathbb{E}[Q^s(p,0)]$ in the data generating process lies between the bounds shown with the dotted black lines, which depict the true values of $\mathbb{E}[\underline{t}(M)]$ and $\mathbb{E}[\bar{t}(M)]$. The baseline assumptions are a linear system, downward-sloping demand, upward-sloping supply, and the prior bound that $\mathbb{E}[Q^s(p,z)] \in [4,12]$ for both $z = 0, 1$. Instrument monotonicity (IM) is the assumption that $B^s_z \leq 0$. The bounded demand and supply curve assumptions (BD and BS) require these curves to stay within $[4,12]$ over the support of log prices in the fit DGP, approximately $[-1.11, .66]$.

values.

The assumption that supply is linear also expands the menu of target parameters for which informative conclusions can potentially be drawn. For example, we can bound the average supply curve at given values of the instrument:

$$\mathbb{E}[Q^{\mathrm{S}}(p,z)] = \mathbb{E}[B^{\mathrm{S}}_{1}] + \mathbb{E}[B^{\mathrm{S}}_{\mathrm{Z}}]z + \mathbb{E}[B^{\mathrm{S}}_{\mathrm{P}}]p. \tag{22}$$

This is a composite target parameter because the instrument is not excluded: a given market type $M = m$ does not uniquely pin down a value of $\mathbb{E}[Q^{\mathrm{S}}(p,z)|M = m]$, which depends on three unknowns. We need to solve (18) for $\underline{t}(m)$ and $\bar{t}(m)$, which can be done numerically for every m through linear programming by searching for linear supply curves that would produce the equilibria characterized by market type m. The values of $\underline{t}(m)$ and $\bar{t}(m)$ are infinite without further restrictions. They can be made finite by placing prior bounds on the magnitude of $\mathbb{E}[Q^{\mathrm{S}}(p,z)]$.

Figure 9 illustrates this point with the Fulton fish market DGP. All bounds in the figure are for $\mathbb{E}[Q^{\mathrm{S}}(p,0)]$ and are derived under the assumption of a linear system, downward-sloping demand, and upward-sloping supply, with $\mathbb{E}[Q^{\mathrm{S}}(p,0)]$ restricted to lie within the same prior bounds $[4,12]$ that were used in Figure 7. By themselves, these assumptions produce bounds that are wide, but non-trivial. Adding instrument monotonicity tightens the bounds somewhat. So does assuming that demand curves must be entirely bounded

within $[4, 12]$ over the support of prices. Making the analogous assumption about the supply curves provides the most information. Combining all three auxiliary assumptions produces the bounds shown with the solid black curve. While the bounds contain perfectly inelastic supply curves, they rule out highly elastic supply curves.

The results in Figure 9 are interesting in the context of the classical, constant coefficients model. The usual analysis of that model would conclude that the constant slope of supply is completely unidentified without a demand shifter. Here, we have weakened the assumption of a constant slope, but imposed additional—albeit quite conservative—prior bounds on the supply curve that are not usually entertained in the classical analysis. The result is a much less pessimistic conclusion. Obtaining this conclusion required entertaining the possibility of partial identification.

The implications for empirical practice are important. For example, consider the change in deadweight loss from changing sales taxes. Assuming no income effects, deadweight loss can be summarized with target parameters that depend only on the market-level supply and demand curves (Harberger, 1964; Chetty, 2009). In Appendix SA.2, we show that if price and quantity are measured in logs, then a marginal increase in ad valorem tax from a base rate of r leads to a relative change in deadweight loss of

$$\frac{\Delta_{\text{DWL}}}{\Delta_{\text{REV}}} = \frac{r B_{\text{P}}^{\text{D}} B_{\text{P}}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}} + r\left(1 - B_{\text{P}}^{\text{D}}\right) B_{\text{P}}^{\text{S}}}, \tag{23}$$

where Δ_{REV} is the change in government revenue. Deadweight loss depends on both supply and demand. A classical analysis would conclude that it is unidentified, while our analysis produces informative bounds, even while allowing for unrestricted heterogeneity.[6]

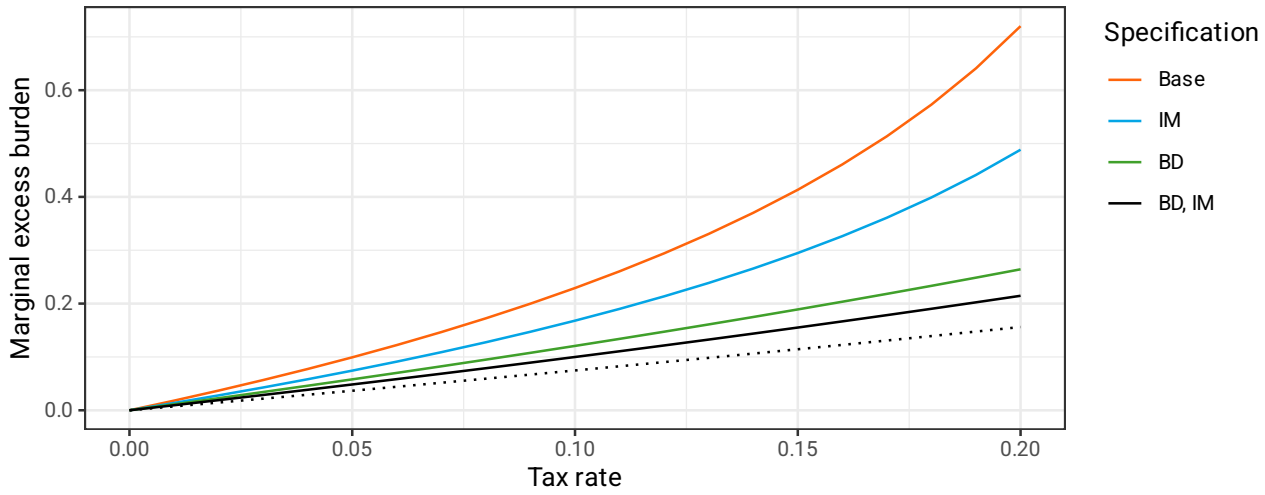
Figure 10 shows upper bounds on the average of (23) across markets using the Fulton fish market DGP. The upper bound is informative and increasing as a function of the base tax rate, reflecting the widening of Harberger's triangle. The lower bound is uninformative (zero) because the assumptions considered so far cannot rule out the possibility that supply is perfectly inelastic.

3.3 Reverse engineering linear IV estimands with random coefficients

An influential literature initiated by Imbens and Angrist (1994) takes a very different approach to accounting for heterogeneity. Instead of developing new methods to infer natural target parameters, this literature starts with classical IV estimators and reverse-engineers interpretations that allow for heterogeneity. See Mogstad and Torgovitsky (2024) for a survey. Analyzing the consequences of misspecification in simultaneous equations models is an exercise with a long history (for example, Bronfenbrenner, 1953). The novelty of the

[6] If the instrument is a tax (Section 2.6), then deadweight loss can be identified in the classical analysis through the reduced form (Harberger, 1964; Chetty, 2009; Zoutman et al., 2018). This approach utilizes economic structure that doesn't necessarily apply to other instruments, such as the weather shock instrument used in the Fulton fish market.

Figure 10: Bounds on marginal relative deadweight loss in the Fulton fish market



Notes: Bounds on marginal relative deadweight loss, as defined in (23) for different values of r. Bounds are constructed using the same data generating process as in Figures 7 and 9. The true upper bound in the data generating process is indicated with a dotted black line. The baseline assumptions are a linear, downward-sloping demand with linear, upward-sloping supply curves that satisfy the bounded supply assumption. Instrument monotonicity (IM), bounded demand (BD), and bounded supply (BS) are the same as defined in Figure 9.

reverse engineering proposal is the suggestion that the reverse-engineered interpretations actually provide useful conclusions.

This suggestion turns out to be particularly hard to support in a model of supply and demand. In Section 2, we saw that the Wald estimand overweights markets for which the instrument moves prices more and that it can potentially be an attenuated measure of demand. This theoretical possibility was borne out in the Fulton fish market data, which had the Wald estimand lying outside of the nonparametric sharp identified set for average slope of demand. In this section, we provide a theoretical explanation for this phenomenon in the context of the linear random coefficients system (20).

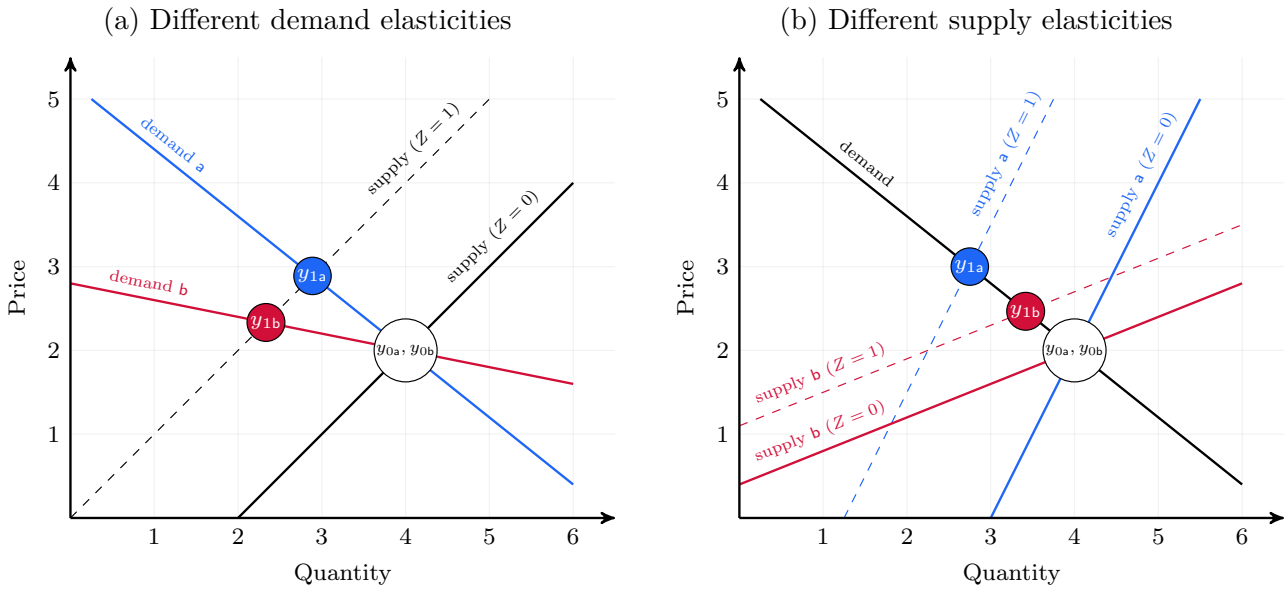
Our starting point is a weighted average interpretation of a class of two-stage least squares (2SLS) estimands. In Appendix SA.3, we build on the analysis of Angrist et al. (2000) to show that

$$-\beta_{\text{2SLS}} = \mathbb{E}[\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})B_{\text{P}}^{\text{D}}] = \mathbb{E}[B_{\text{P}}^{\text{D}}] + \mathbb{C}[B_{\text{P}}^{\text{D}}, \omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})], \tag{24}$$

where $\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})$ are weights that have unit mean. Angrist et al. (2000) showed that these weights will be non-negative under the instrument monotonicity condition considered in Section 2.5. We show that the non-negativity of ω_{2SLS} can also be ensured without instrument monotonicity by assuming that the coefficient on the supply shifter is independent of the coefficients on price.

Non-negative weighting interpretations like these are commonly found in reverse engineering analyses of linear IV estimators under heterogeneous treatment effects (Angrist

Figure 11: Instruments create larger price changes in more inelastic markets



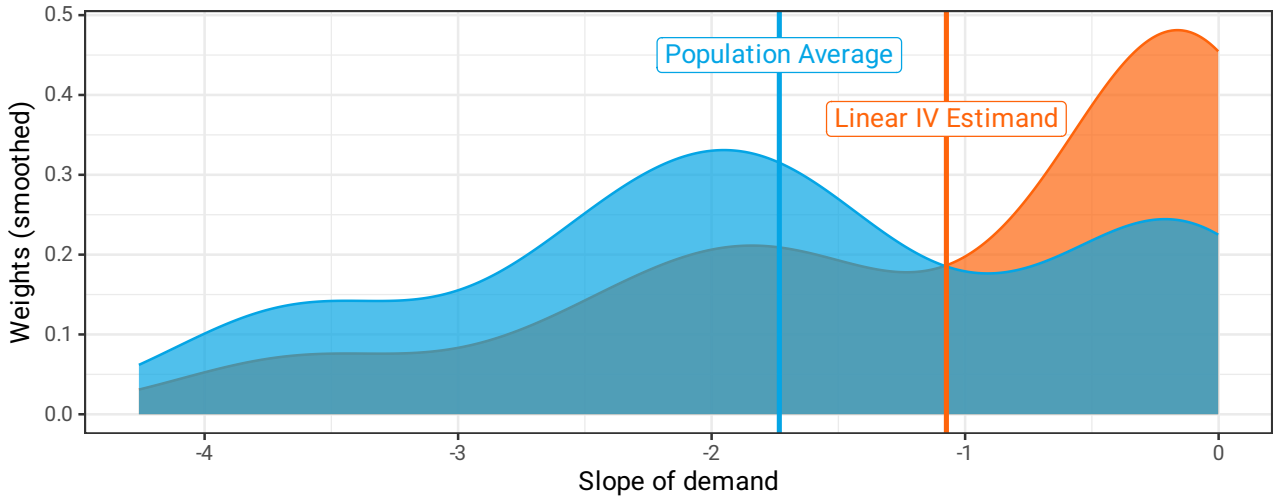
Notes: Figure (a) shows two markets with the same supply, but different demand curves. When $Z_1 = 0$, both markets have the same equilibrium, $y_{0a} = y_{0b}$. When $Z_1 = 1$, prices change more in market a, which has a steeper inverse demand curve and therefore more inelastic demand. Figure (b) shows two markets with the same demand, but different supply curves. When $Z_1 = 1$, prices change more in market a, which has more inelastic supply.

and Pischke, 2009; Mogstad and Torgovitsky, 2024). Their attraction is in ensuring that if the underlying causal effect takes a single sign with probability one, then the IV estimand also has that same sign, a property Blandhol et al. (2025) described as "weakly causal." In the supply and demand context, the underlying causal effect of price on demand is already assumed to be negative, so accurately reflecting this assumed sign is not much of an achievement.

Moreover, the particular weighting scheme used by the 2SLS estimand will typically make it an attenuated measure of demand. For the binary instrument case, Angrist et al. (2000, pg. 507) noted that if the coefficients in the supply equation on price and the excluded instrument are both constant, then $|\beta_{2SLS}| \leq \mathbb{E}[B_P^D]$. In Appendix SA.3, we show that this conclusion holds under more general conditions. The central requirement is that B_P^D and B_P^S are not strongly negatively correlated. Similar attenuation results continue to hold even if the demand and supply equations are nonlinear (Section SA.3.3).

Figure 11a illustrates how attenuation arises by comparing the impacts of an equal additive supply shift on the equilibria of two markets with the same supply curve but different demand curves. In both markets the shift leads equilibrium prices to increase and equilibrium quantities to decline. Prices increase more in market a because its demand is less elastic. This means that the instrument (supply shifter) has a larger impact on the endogenous variable (price) in markets with more inelastic demand. As we explain in

Figure 12: The linear IV estimand is attenuated



Notes: The blue density is a kernel density estimate of the population density of the average slope of demand, $\bar{\varepsilon}^D(M)$. The orange density is a kernel density estimate of the contribution to the Wald estimand, $\omega_{2SLS}(M)\bar{\varepsilon}^D(M)$, where in this case $\omega_{2SLS} = \omega_{WALD}$ are as defined in (14). The vertical lines show $\bar{\bar{\varepsilon}}^D \equiv \mathbb{E}[\bar{\varepsilon}^D(M)]$ and $\beta_{2SLS} = \mathbb{E}[\omega_{2SLS}(M)\bar{\varepsilon}^D(M)]$. The data generating process is the same DGP fit to the Fulton fish market data that was used in Figures 7 and 9.

Appendix SA.3, the statistical weighting used by the 2SLS estimator overweights markets that experience larger price changes. This means that ω_{2SLS} overweights more inelastic markets, leading β_{2SLS} to understate the average elasticity.

This reasoning depends on the slope of supply as well. Figure 11b shows the same additive supply shift in two markets with the same demand curve but different supply curves. Prices change more in the market with less elastic supply. Together, Figure 11 shows that markets that are more inelastic—both in supply and demand—receive larger weight in the linear IV estimand. Only if markets that have inelastic demand also tend to have elastic supply, so that B_P^D and B_P^S are negatively dependent, is it possible for the IV estimand to overstate the average demand slope.

Figure 12 illustrates the magnitude of the attenuation using the Fulton fish market DGP. The linear IV estimand is about 60% as large in magnitude as the average of the demand coefficient, reflecting substantial attenuation. The weights show that this happens because the IV estimand overweights markets that are highly inelastic, consistent with the intuition just developed. These results illustrate how the linear IV estimand can fail to be a useful measure of demand when there is unobserved heterogeneity across markets.

4 Linear Random Coefficients with Smooth Heterogeneity

The results in the previous section showed that useful conclusions can be drawn from a linear system even while allowing for unrestricted heterogeneity. However, the approach taken would be challenging to apply to settings richer than the Fulton fish market. The

dimension of market types M grows quickly with the number of observed equilibria, which explodes the computational burden. Multivalued instruments and covariates further exacerbate this burden.

In this section, we address these problems by using a random coefficients model with smooth heterogeneity. This model solves the dimensionality problem by imposing some discipline on the distribution of market types. We do this using a mixture model formulated as a linear sieve (see, for example, Chen, 2007), which allows us to explore both parsimonious and flexible specifications in a single, computationally tractable framework. We develop two related approaches that differ in what features of the observable variables they use and in what portion of the random coefficients distribution must be modeled. In Section 4.2, we consider a "full information" approach that uses the entire distribution of prices and quantities. Then, in Section 4.3, we consider a "limited information" approach that uses only selected coefficients from reduced form regressions. The terminology is intended to echo the distinction laid out by Anderson and Rubin (1949) in the context of models with constant effects.

4.1 Model

We generalize the random coefficients system (20) to

$$\begin{aligned} q^{\text{D}}(p, Z, B) &= Z' B_{\text{Z}}^{\text{D}} - p B_{\text{P}}^{\text{D}}, \\ q^{\text{S}}(p, Z, B) &= Z' B_{\text{Z}}^{\text{S}} + p B_{\text{P}}^{\text{S}}, \end{aligned} \tag{RC}$$

where Z is a vector that now includes both instruments and covariates—including a constant term—and the unobservable $B \equiv (B_{\text{Z}}^{\text{D}}, B_{\text{P}}^{\text{D}}, B_{\text{Z}}^{\text{S}}, B_{\text{P}}^{\text{S}})$ now combines the random scalars, B_{P}^{D} and B_{P}^{S}, with the random vectors, B_{Z}^{D} and B_{Z}^{S}. We continue to assume that $B_{\text{P}}^{\text{D}} \geq 0$ and $B_{\text{P}}^{\text{S}} \geq 0$ with probability one and that a unique equilibrium exists. The equilibrium can be characterized through reduced form equations that generalize (21):

$$p^{\text{E}}(Z, B) = Z' \left(\frac{B_{\text{Z}}^{\text{D}} - B_{\text{Z}}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}} \right) \quad \text{and} \quad q^{\text{E}}(Z, B) = Z' \left(\frac{B_{\text{P}}^{\text{S}} B_{\text{Z}}^{\text{D}} + B_{\text{P}}^{\text{D}} B_{\text{Z}}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}} \right). \tag{25}$$

We continue to maintain the exogeneity condition that Z and B are independent.

The vector Z appears in both equations of (RC). Exclusion restrictions can be imposed by assuming that certain components of B_{Z}^{D} or B_{Z}^{S} are constant and equal to zero. Other components of Z may have non-zero coefficients in both B_{Z}^{D} and B_{Z}^{S}, in which case these variables can be interpreted as covariates. The tax or subsidy instrument assumptions considered in Section 2.6 can be imposed by setting components of B_{Z}^{D} or B_{Z}^{S} to be equal in magnitude to B_{P}^{D} or B_{P}^{S}.

4.2 Full information

In this section, we model the distribution F of the entire random coefficients vector, B. We assume that F lives in the set

$$\mathcal{F}^{\dagger} = \left\{ F : F(b) = \sum_{h=1}^{d_\pi} \pi_h F_h(b;\alpha) \quad \text{for some } (\alpha,\pi) \in \mathcal{A}\Pi \right\}, \tag{26}$$

where $\{F_h(b;\alpha)\}_{h=1}^{d_\pi}$ is a collection of known basis functions and $\mathcal{A}\Pi$ is a set of permissible parameters for (α,π). In the specifications we consider, each $F_h(\cdot;\alpha)$ is a distribution function for any fixed α, so that F is a mixture distribution over F_h with mixing weights π_h that reside in the simplex. We don't use α in our simulation or empirical results, so we suppress it in the remainder of the main text and replace $\mathcal{A}\Pi$ by simply Π. As we discuss in Section SA.4, the leading way that α could be used is to enforce the constraint that some coefficients are deterministic.

We focus our simulation and empirical work on mixtures of B-spline densities, where each component F_h is an integrated B-spline component. We find B-splines attractive because they are easy to parameterize and can approximate a large class of densities arbitrarily well.[7] The mixture components F_h should be chosen so that all $F \in \mathcal{F}^{\dagger}$ have $\mathbb{P}_F[B_{\mathrm{P}}^{\mathrm{D}} \geq 0, B_{\mathrm{P}}^{\mathrm{S}} \geq 0] = 1$. They can also optionally be chosen to enforce an instrument monotonicity condition (Section 2.5) on the sign of a random coefficient for a component of Z that is an excluded instrument.

Each distribution F implies a conditional distribution of Y given Z via the reduced form (25):

$$\mathbb{P}_F[Y \leq y | Z = z] \equiv \int \mathbb{1}[y^{\mathrm{E}}(z,b) \leq y] \, dF(b), \tag{27}$$

where $y^{\mathrm{E}}(z,b) \equiv (p^{\mathrm{E}}(z,b), q^{\mathrm{E}}(z,b))$. The definition of the identified set $\mathcal{F}^{\star}$ remains the same as in Section 2.8, except that now we allow for Y to be continuous, so we match its distribution function (27) rather than its mass function. The target parameter and its identified set $\mathcal{T}^{\star}$ also remain the same except that we write it as $\mathbb{E}_F[\tau(Z,B)]$ because the structural functions $\boldsymbol{Q} \equiv (Q^{\mathrm{D}}, Q^{\mathrm{S}})$ are now fully determined by (Z,B) given the random coefficients specification. With these changes, we again pursue a strategy of computing sharp bounds on the target parameter by solving the problems

$$\tau_{\mathrm{lb}}^{\star}/\tau_{\mathrm{ub}}^{\star} \equiv \inf_{F \in \mathcal{F}^{\dagger}} / \sup_{F \in \mathcal{F}^{\dagger}} \mathbb{E}_F[\tau(Z,B)] \quad \text{s.t. (27) for all } y \text{ and } z. \tag{28}$$

[7] See, for example, Gehringer and Redner (1992), Cai and Meyer (2011), or Matsuda and Iwafuchi (2025). Other choices that also have universal approximation properties within large classes of distributions are mixtures of beta distributions (for example, Diaconis and Ylvisaker, 1985; Perron and Mengersen, 2001) and mixtures of Erlang distributions (for example, Lee and Lin, 2010, and the references cited within).

The linear basis specification (26) helps with solving (28) because it implies that

$$\mathbb{P}_F[Y \leq y | Z = z] = \sum_{h=1}^{d_\pi} \pi_h \overbrace{\int \mathbb{1}[y^{\mathrm{E}}(z,b) \leq y] \, dF_h(b)}^{\equiv \bar{g}_h(y,z)} \equiv \pi' \bar{g}(y,z)$$
$$\text{and} \quad \mathbb{E}_F[\tau(Z,B)] = \sum_{h=1}^{d_\pi} \pi_h \underbrace{\mathbb{E}\left[\int \tau(Z,b) \, dF_h(b)\right]}_{\equiv \bar{\tau}_h} \equiv \pi' \bar{\tau}, \tag{29}$$

which are both linear functions of π with coefficients that can be computed and/or estimated. The linearity means (28) becomes a linear program:

$$\max_{\pi \in \Pi} \pi' \bar{\tau} \quad \text{s.t.} \quad \pi' \bar{g}(y,z) = \mathbb{P}[Y \leq y | Z = z] \text{ for all } y, z, \tag{30}$$

as long as Π is a polyhedral set, such as the simplex in the leading case of interest.

Because we are allowing $Y \equiv (P,Q)$ to be continuously distributed, it may not be possible to impose the constraints on the inner problem of (30) for all y. A straightforward solution is to impose the constraints for a large but finite grid of y and all z, although this can lead the resulting bounds to be potentially non-sharp. This is, however, only an issue for the simulation exercise of computing an identified set from a known population distribution. In Appendix SA.5, we develop estimators of $\tau^\star_{\text{lb}}$ and $\tau^\star_{\text{ub}}$ using an approach similar to that in Mogstad et al. (2018). We construct the estimators to be of the sharp population bounds by using a criterion function that asymptotically incorporates all of the information in the distribution of Y, whether discrete or continuous.

4.3 Limited information

The full information approach allows us to estimate sharp bounds because it matches the entire distribution of observables. This is also its disadvantage. Modeling the full distribution of (P,Q) requires modeling the full distribution of B, which is still a relatively large vector even in simple cases. For example, with only a single excluded instrument and no covariates, B has five components: two intercepts, two price slopes, and the slope of the instrument. The distribution functions F are five-dimensional objects, which can be challenging to model flexibly in the linear sieve form (26). Covariates add to this challenge, unless they are assumed to have constant coefficients (see Appendix SA.4). Even so, having to model the distribution of the intercepts seems like a step backwards relative to the classical constant coefficients model.

These problems can be solved by using a limited information approach in which only certain features of the conditional distribution of $(P,Q)|Z$ are matched. We choose these features so that they only depend on the vector $\tilde{B} \equiv (B^{\mathrm{D}}_{\mathrm{P}}, B^{\mathrm{S}}_{\mathrm{P}}, B^{\mathrm{S}}_{\mathrm{Z},1})$ composed of the price coefficients and the coefficient on the excluded instrument, focusing here on the case of a single supply shifter, Z_1. The distribution $\tilde{F}$ of this three-dimensional vector is much

easier to model in the linear mixture form (26) because we can be agnostic about the distribution of the other random coefficients not in $\tilde{B}$.

There are three costs of focusing on $\tilde{B}$. One is that we are limited to considering target parameters that only depend on $\tilde{B}$. The second is that the resulting bounds may not be sharp. The third is that, as will become clear ahead, we need the excluded instrument to have three or more points of support to secure interesting bounds. None of these costs turn out to be important in the sales tax application in Section 5.

The features of the observable distribution that we match are coefficients in linear regressions implied by the reduced form equations (25). To see how this works, let Z_1 denote the excluded supply shifter, which has demand coefficient $B^{\text{D}}_{\text{Z},1} = 0$. Due to the exogeneity condition, the coefficients on Z in a regression of either P or Q onto Z are the averages of the reduced form random coefficients given in (25). The coefficients $\rho_{\text{P},1}$ and $\rho_{\text{Q},1}$ on Z_1 only depend on $\tilde{B} \equiv (B^{\text{D}}_{\text{P}}, B^{\text{S}}_{\text{P}}, B^{\text{S}}_{\text{Z},1})$:

$$\rho_{\text{P},1} = \mathbb{E}\left[\frac{-B^{\text{S}}_{\text{Z},1}}{B^{\text{D}}_{\text{P}} + B^{\text{S}}_{\text{P}}}\right] \quad \text{and} \quad \rho_{\text{Q},1} = \mathbb{E}\left[\frac{B^{\text{S}}_{\text{Z},1} B^{\text{D}}_{\text{P}}}{B^{\text{D}}_{\text{P}} + B^{\text{S}}_{\text{P}}}\right],$$

assuming that these moments exist.[8] Any full distribution F for B that is in the identified set ($F \in \mathcal{F}^\star$) produces a three-dimensional distribution $\tilde{F}$ for $\tilde{B}$ that needs to match these reduced form coefficients through the equations

$$\mathbb{E}_{\tilde{F}}\left[\frac{-B^{\text{S}}_{\text{Z},1}}{B^{\text{D}}_{\text{P}} + B^{\text{S}}_{\text{P}}}\right] = \sum_{h=1}^{d_\pi} \pi_h \left[\int \left(\frac{-b^{\text{S}}_{\text{Z},1}}{b^{\text{D}}_{\text{P}} + b^{\text{S}}_{\text{P}}}\right) d\tilde{F}_h(b)\right] = \rho_{\text{P},1} \tag{31}$$

$$\text{and} \quad \mathbb{E}_{\tilde{F}}\left[\frac{B^{\text{S}}_{\text{Z},1} B^{\text{D}}_{\text{P}}}{B^{\text{D}}_{\text{P}} + B^{\text{S}}_{\text{P}}}\right] = \sum_{h=1}^{d_\pi} \pi_h \left[\int \left(\frac{b^{\text{S}}_{\text{Z},1} b^{\text{D}}_{\text{P}}}{b^{\text{D}}_{\text{P}} + b^{\text{S}}_{\text{P}}}\right) d\tilde{F}_h(b)\right] = \rho_{\text{Q},1}. \tag{32}$$

Minimizing or maximizing $\mathbb{E}_F[\tau(Z, B)] = \pi'\bar{\tau}$ over $\pi \in \Pi$ subject to (31)–(32) is a linear program that produces valid outer bounds.

The simulations in the next section show that these outer bounds are quite wide when only matching the two reduced form coefficients (31)–(32). However, we can find other quantities that only depend on $\tilde{B}$ by considering powers and products of P and Q. For any non-negative integers j and k, one can show using a bit of algebraic accounting (Appendix

[8] Masten (2018) observed that the moments of the reduced form coefficients might not exist because of the random price coefficients in the denominator. In the supply and demand setting with linear functions of price and reduced form (3), this could happen if B^{D}_{P} and/or B^{S}_{P} have densities that put a large amount of mass near zero. We assume in the following that all of the reduced form coefficients that we analyze exist both under the population distribution and under any distribution $F \in \mathcal{F}^\dagger$.

SA.6) that

$$P^j Q^k = \sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|=j+k} Z^{\boldsymbol{\ell}} \left(\sum_{\substack{|\boldsymbol{j}|=j \\ \boldsymbol{j}+\boldsymbol{k}=\boldsymbol{\ell}}} \frac{j!\, k!}{\boldsymbol{j}!\, \boldsymbol{k}!} \prod_{i=1}^{d_z} \left(\frac{B^{\mathrm{D}}_{\mathrm{Z},i} - B^{\mathrm{S}}_{\mathrm{Z},i}}{B^{\mathrm{D}}_{\mathrm{P}} + B^{\mathrm{S}}_{\mathrm{P}}} \right)^{j_i} \left(\frac{B^{\mathrm{S}}_{\mathrm{P}} B^{\mathrm{D}}_{\mathrm{Z},i} + B^{\mathrm{D}}_{\mathrm{P}} B^{\mathrm{S}}_{\mathrm{Z},i}}{B^{\mathrm{D}}_{\mathrm{P}} + B^{\mathrm{S}}_{\mathrm{P}}} \right)^{k_i} \right), \tag{33}$$

where $\boldsymbol{\ell}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are d_z-dimensional vectors of non-negative integers (multi-indices) with the standard notation $\boldsymbol{\ell} \equiv (\ell_1, \ldots, \ell_{d_z})$, $|\boldsymbol{\ell}| \equiv \sum_{i=1}^{d_z} \ell_i$, and $Z^{\boldsymbol{\ell}} \equiv \prod_{i=1}^{d_z} Z_i^{\ell_i}$. One of the multi-indices in the sum is $\boldsymbol{\ell} = (j+k, 0, \ldots, 0)$, which corresponds to the regressor $Z^{\boldsymbol{\ell}} = Z_1^{j+k}$. From (33), the coefficient on Z_1^{j+k} from regressing $P^j Q^k$ onto $\{Z^{\boldsymbol{\ell}} : \boldsymbol{\ell} \text{ s.t. } |\boldsymbol{\ell}| = j+k\}$ is

$$\rho_{j,k} \equiv (-1)^j \, \mathbb{E}\left[\frac{(B^{\mathrm{S}}_{\mathrm{Z},1})^{j+k} (B^{\mathrm{D}}_{\mathrm{P}})^k}{(B^{\mathrm{D}}_{\mathrm{P}} + B^{\mathrm{S}}_{\mathrm{P}})^{j+k}} \right],$$

where the notation nests the two coefficients in (31)–(32) as $\rho_{1,0} \equiv \rho_{\mathrm{P},1}$ and $\rho_{0,1} \equiv \rho_{\mathrm{Q},1}$.

Each $\rho_{j,k}$ also only depends on $\tilde{B}$, but in a different way from the $(j,k) = (1,0)$ and $(0,1)$ cases shown in (31)–(32). This means that tighter bounds can be found by optimizing the target parameter under the constraints

$$(-1)^j \, \mathbb{E}_{\tilde{F}}\left[\frac{(B^{\mathrm{S}}_{\mathrm{Z},1})^{j+k} (B^{\mathrm{D}}_{\mathrm{P}})^k}{(B^{\mathrm{D}}_{\mathrm{P}} + B^{\mathrm{S}}_{\mathrm{P}})^{j+k}} \right] = \sum_{h=1}^{d_\pi} \pi_h \left[(-1)^j \int \left(\frac{(b^{\mathrm{S}}_{\mathrm{Z},1})^{j+k} (b^{\mathrm{D}}_{\mathrm{P}})^k}{(b^{\mathrm{D}}_{\mathrm{P}} + b^{\mathrm{S}}_{\mathrm{P}})^{j+k}} \right) d\tilde{F}_h(b) \right] = \rho_{j,k} \tag{34}$$

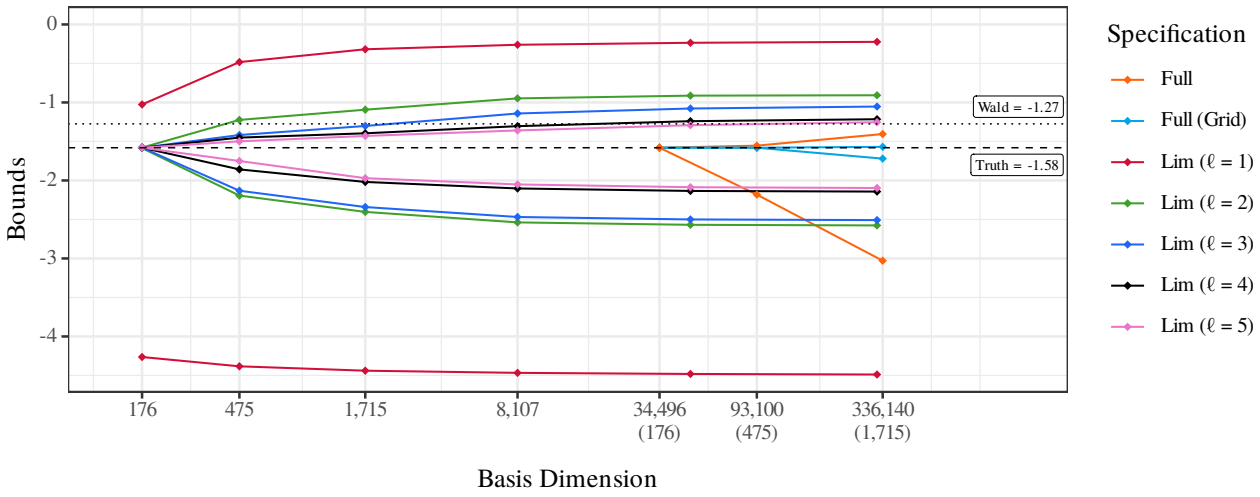
for any non-negative integers j and k. In the next section, we find that adding all (j,k) combinations with $j + k \leq 3$ is enough to produce tight bounds in a DGP constructed from the Fulton fish market data. However, we need Z_1 to be cardinal-valued for the regressions that produce $\rho_{j,k}$ to make sense, and we need Z_1 to have at least $j + k + 1$ values for $\rho_{j,k}$ to exist.

In Appendix SA.5, we construct estimators of the outer bounds produced by the limited information approach. We also show that inference in the limited information approach can be cast as inference in a linear system of equations with known coefficients. This allows us to apply the inference procedure developed by Fang et al. (2023). In our application, we use a simpler projection inference approach that is also discussed in Appendix SA.5.

4.4 Smooth heterogeneity in the Fulton fish market

Figure 13 reports bounds on the average slope of demand, $\mathbb{E}[B^{\mathrm{D}}_{\mathrm{P}}]$, in a DGP fit to the Fulton fish market data. The x-axis shows the number of terms in the basis, which is five-dimensional in the full information approach and three-dimensional in the limited information approach. A tick with a parenthetical is a five-dimensional basis, with the parenthetical reporting the dimension of the implied three-dimensional basis of $\tilde{B}$. We construct the DGP by fitting a full information basis to the original support of 111 price

Figure 13: Bounds on the average slope of demand using a smooth basis



Notes: Bounds on $-\mathbb{E}[B_P^D]$ *using cubic B-spline bases. "Full" is the full information approach constrained by the original* 111 *equilibria observed in the Fulton fish market data, while "Full (Grid)" is the full information approach constrained by a grid of* 450 *equilibria that contains the original equilibria. "Lim" is the limited information approach with all reduced form coefficients* (j,k) *in* (34) *having* $j+k \leq \ell$*. The x-axis is in log scale. A single number on the x-axis indicates the size of a three-dimensional basis for* $\tilde{B}$ *using the limited information approach. When there are two numbers, the larger one is the size of a five-dimensional basis for* B *using the full information approach, while the one in parentheses is the size of the implied three-dimensional basis for* $\tilde{B}$*. The bounds are constructed using a distribution of observables fit to the original Fulton fish market data using the five-dimensional cubic B-spline with* 34,496 *components.*

and quantity pairs. The basis is a tensor of univariate cubic B-splines; it has 34,496 terms overall.[9] The true value of the average slope of demand ($-\mathbb{E}[B_P^D]$) under the DGP is -1.58, while the Wald estimand is -1.27.

The full information approach leads to point identification when fitting the baseline basis to the original support. Expanding the full information basis by bisecting all of the knots for the slope coefficients creates a basis with 93,100 terms, which leads to partial identification with the original support, but still retains point identification on a richer grid of 450 equilibria. Bisecting once more produces a basis with 336,140 terms, leading to partial identification for the fine grid as well, although the bounds are quite tight. These results show the primary trade-offs involved in the full information approach: a great deal of information can be harnessed by fitting the joint distribution of prices and quantity, but doing so requires modeling the full distribution of random coefficients, which is difficult due to the curse of dimensionality.

The curse of dimensionality is much less pronounced in the limited information approach, which only requires modeling a three-dimensional density. Figure 13 shows bounds that use up to 317,275 terms, representing a highly flexible three-dimensional B-spline.

[9]For both intercepts, there are twelve equally spaced points between 4 and 12. For the demand slope, there are nine equally spaced points between 0 and 4.5. For the supply slope, there are two knots: 0.01 and 1.5. For the instrument slope, there are also two knots: -1 and 0.

The bounds expand somewhat at first, but then quickly stabilize, suggesting that they are approaching the nonparametric bounds. Figure SA.15 shows that this expansion is driven by the flexibility of the instrument slope, rather than the demand or supply slopes.

The cost for this dimension reduction is that the limited information approach only harnesses information from the reduced form moments $\rho_{j,k}$ defined in (34). Figure 13 shows bounds for all (j,k) pairs with $j+k \leq \ell$, where ℓ ranges between one and five. The bounds are quite wide with $\ell = 1$. Including second-order terms ($\ell = 2$) adds a great deal of information, but requires an instrument with three points of support instead of just two.[10] Increasing to $\ell = 3$ provides a modest tightening, as does $\ell = 4$. With $\ell = 5$ there is almost no additional tightening. In practice, it may be statistically challenging to use large values of ℓ as this requires estimating the coefficient on Z_1^ℓ in a linear regression including all terms up to the ℓth power. In our sales tax application, we find that setting $\ell = 2$ already provides tight and precisely estimated bounds even with a flexible basis.

5 The Welfare Impacts of Sales Taxes

In this section, we apply our methodology to study a core question in public finance: how large is the efficiency loss from sales taxes and who bears the loss? Recent research on this question includes Kroft et al. (2024a,b) and Gaarder and Henry de Frahan (2025), all of which use models without unobserved heterogeneity across markets. We estimate similar constant coefficients models before incorporating unobserved heterogeneity with a linear random coefficients model.

5.1 Institutions and data

Sales taxes in the United States are imposed on most goods and some services in 45 states and the District of Columbia. Tax rates vary substantially across regions. We focus on counties as the regional unit.

Our data follows the construction used by Gaarder and Henry de Frahan (2025), who combine NielsenIQ scanner data with county-level sales tax information from the Thomson Reuters OneSource Sales Tax database. The authors construct a balanced panel spanning fourteen semiannual periods between January 2008 and December 2014 with data on quantities, prices, and sales tax rates in 2,111 counties and 249 product modules. The data is used to construct quantity and price indices for each county, module, and time period triple. Appendix SA.8.1 contains more details on this construction, as well as summary statistics of the resulting data; see Gaarder and Henry de Frahan (2025) for full details.

[10] In our simulation, we assume that we know the reduced form population coefficients up to order ℓ, so we don't need to specify the marginal distribution of the instrument. In practice, one would need the instrument to have enough support points to be able to identify these OLS estimands.

We define a market as a county-module pair. Consumer prices P and quantities Q are log indices, which are then averaged over two-year periods and residualized against module-region-time period fixed effects; see Appendix SA.8.1. The instrument Z is the two-year change in $\log(1+r)$, where r is the ad valorem sales tax for a county-module pair, which we residualize in the same way as price and quantity.

5.2 Estimates without market heterogeneity

Panel B of Table SA.3 reports the reduced form effects of the sales tax instrument on quantity and prices. We find that a one percent increase in sales taxes reduces quantity by 0.56 percent. It increases the tax-inclusive price faced by consumers by 0.90, which means that it lowers the before-tax price by 0.10.

Panel A of Table 2 turns these reduced form estimates into structural estimates of the slopes of supply and demand using the classical reasoning for a linear model with constant coefficients. The demand slope estimate of $0.62 \approx 0.56/0.90$ comes from using sales tax as an instrument for the consumer-facing tax-inclusive price. The supply slope estimate of $5.81 \approx 0.56/0.10$ comes from using sales tax as an instrument for the before-tax price, which is justified under the Supply-Side Ramsey Exclusion Restriction (RER) introduced by Zoutman et al. (2018) and discussed in Section 2.6.

We plug these elasticities into the expressions in Appendix SA.2 to produce estimates of consumer incidence and marginal deadweight loss. At the baseline average tax rate of 7.2 percent, we estimate the deadweight loss from a marginal increase in sales taxes to be \$0.04 per dollar of revenue raised. The share borne by consumers (the consumer incidence) is estimated to be 91 percent.

5.3 Estimates with market heterogeneity

In Appendix SA.7, we show how to derive linear (in logs) supply and demand equations from a microfounded model of consumers and firms. The derivation shows how unobserved market heterogeneity in the price coefficient arises naturally. For the demand equation, heterogeneity across markets can come from variation in the composition of consumers. For the supply equation, heterogeneity across markets can come from variation in firm technology and in the costs of inputs. The estimates in Panel A of Table 2 rely on ruling out these sources of heterogeneity.

Panel B of Table 2 reports results from a linear random coefficients model that allows for unobserved market heterogeneity in demand and supply elasticities. We use the limited information approach discussed in Section 4.3 with a rich B-spline basis for the random coefficients. Appendix SA.8.2 provides more information on how this basis is specified. Confidence intervals, which are reported in square brackets, are constructed using the projection approach described in Appendix SA.5.5, with variance estimators that are clustered at the module-by-county level.

Table 2: Estimates of average elasticities and welfare

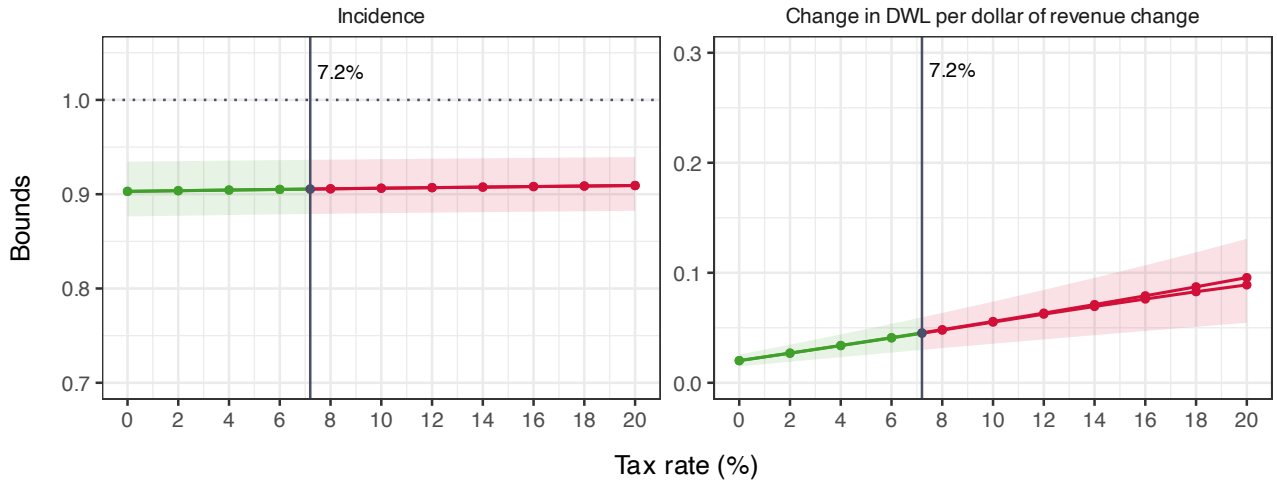
Specification		Target parameters			
Exclusions	Moments	$E[B_p^d]$	$E[B_p^s]$	ICD	DWL
Panel A. Constant coefficient benchmark					
Standard + Ramsey	IV	0.62 (0.52, 0.72)	5.81 (4.07, 7.55)	0.91 (0.89, 0.93)	0.04 (0.03, 0.05)
Panel B. Random coefficient bounds					
Standard only	$(j+k \leq 1)$	[0.31, 3.71] (0.27, 3.75)	[3.37*, 13.31] (3.37*, 13.42)	[0.56, 0.96] (0.55, 0.97)	[0.02, 0.25] (0.02, 0.25)
Standard + Ramsey	$(j+k \leq 1)$	[0.62, 0.99] (0.51, 1.13)	[3.98, 12.38] (3.52, 13.02)	[0.91, 0.91] (0.88, 0.93)	[0.04, 0.05] (0.03, 0.06)
Standard only	$(j+k \leq 2)$	[0.70, 0.96] (0.39, 1.37)	[3.50, 12.81] (3.37, 13.33)	[0.89, 0.96] (0.81, 0.97)	[0.03, 0.07] (0.02, 0.09)
Standard + Ramsey	$(j+k \leq 2)$	[0.71, 0.96] (0.49, 1.15)	[3.91, 11.60] (3.46, 13.09)	[0.91, 0.91] (0.88, 0.94)	[0.05, 0.05] (0.03, 0.06)

Notes: Panel A reports linear IV estimates and implied consumer incidence (ICD) and deadweight loss (DWL). Panel B reports estimated bounds under the linear random coefficients model; see Appendix SA.8 for details on implementation. In Panel A, round brackets denote 95% cluster-robust confidence intervals. In Panel B, square brackets denote estimated bounds and round brackets denote 95% cluster-robust confidence intervals. Consumer incidence and deadweight loss are evaluated at a baseline total sales tax rate of 7.2 percent. An asterisk indicates a bound that is equal to the logical bound constructed without using the data.

The first row of Panel B reports bounds that only use the standard reduced form coefficients. These bounds are informative and, except for the lower bound on average supply, they are tighter than the logical bounds produced by the model without fitting the data at all. However, they are still quite wide, with the average elasticity of demand being anywhere from highly inelastic (0.31) to highly elastic (3.71). The bounds on supply are similarly quite wide. The combined result is that the bounds on deadweight loss are consistent with a range of possibilities, including moderate efficiency losses of \$0.25 per dollar raised. The bounds on incidence run the gamut from being roughly equally shared (0.56) to being almost fully borne by the consumer (0.96).

The second row of Panel B adds the RER, which requires taxes to affect supply only through the before-tax price. In the random coefficients model, this is imposed by setting $B_{\text{Z}}^{\text{S}} = -B_{\text{P}}^{\text{S}}$. The RER leads to the much tighter set estimate on average demand of $[0.62, 0.99]$. It also produces noticeably tighter bounds on the average supply elasticity, although the tightening is less than one might have anticipated given the identifying power of the RER for the constant coefficients case. The lack of information about supply doesn't matter much for estimates of incidence and deadweight loss, which are nearly identical to the ones produced by the constant coefficients model. The implication is that the way

Figure 14: Incidence and change in deadweight loss for discrete tax changes



Notes: The x-axis shows the counterfactual tax rate, with the vertical line at 7.2 *denoting the baseline rate. The left panel shows the level of consumer incidence under the new tax rate. The right panel shows the change in deadweight loss compared to the baseline rate divided by the change in revenue compared to the baseline rate. These estimates are positive because deadweight loss and revenue change in the same direction as tax rates change. The specification is the same as in the final row of Panel B. Solid lines report point estimates and shaded regions report 95% confidence intervals.*

in which Zoutman et al. (2018) use the RER for identification implicitly relies on market homogeneity in an important way to extract the single estimate 5.81 in Panel A from the range of possibilities represented by our bounds in Panel B.

A more formal explanation of this point can be seen from analyzing the pass-through of taxes to consumer prices, which is given by

$$\frac{\partial}{\partial Z} p^{\mathrm{E}}(Z, B) = \frac{-B_{\mathrm{Z}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}} = \frac{B_{\mathrm{P}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}},$$

where the final equality imposes the RER by setting $B_{\mathrm{Z}}^{\mathrm{S}} = -B_{\mathrm{P}}^{\mathrm{S}}$. Suppose that the pass-through in a market is between .86 and .94, similar in magnitude to our reduced form estimate of 0.90 in Table SA.4, which averages across markets. Even if the demand elasticity $B_{\mathrm{P}}^{\mathrm{D}}$ were known to be exactly .62 in this market, such a narrow range of pass-through would still imply that supply elasticities must lie between 3.81 and 9.71. This suggests that even small amounts of heterogeneity can cause RER-based point identification arguments to rapidly break down.

The RER may be a questionable assumption if there are tax-specific compliance costs. In the third row of Panel B, we replace the RER by second-order moments of the reduced form, which provide a different source of additional information without requiring additional assumptions (see Section 4.3). The resulting bounds are broadly similar to the second row under the RER; the average demand elasticity bound is somewhat tighter with a larger lower endpoint, while the average supply elasticity bound is somewhat wider.

Bounds on incidence and deadweight loss are also slightly wider, but still quite similar to estimates under constant coefficients. Note that our bounds on the average elasticity of demand exclude the point estimate in the constant coefficients case, consistent with the attenuation results developed in Section 3.3.

The final row of Panel B combines the RER with higher-order moments of the reduced form. The estimates remain roughly the same, but the identified sets tighten, as expected from including more information. All of the estimates point to small efficiency losses from sales tax with a large majority of the losses being borne by consumers.

Figure 14 shows the estimated welfare impacts of larger, discrete reforms to the sales tax rate, using the derivations in Appendix SA.2. Points on the x-axis to the left of 7.2 correspond to tax cuts, while points to the right correspond to tax increases. Consumer incidence remains tightly estimated at around 0.91, regardless of the size of the reform. The change in deadweight loss is also tightly estimated, although some ambiguity emerges in our bounds when considering large tax increases. Even for these large tax increases, the results suggest that the efficiency costs of sales taxes are modest.

6 Conclusion

We analyzed the classic simultaneity problem modified to allow for rich unobserved heterogeneity across markets. This modification is natural in the context of the modern literature on instrumental variables with heterogeneous treatment effects and arises naturally from microfounded models. However, it raises substantial complications for identification, computation, estimation, and inference. As we showed, it also implies that linear IV estimators designed for a constant coefficients model will tend to be attenuated. In light of this finding, it may be worth re-examining empirical results that used the linear IV estimator for evidence of attenuation and its downstream implications.

While it is known that most interesting target parameters will not be point identified (Masten, 2018), we showed that it is still possible to provide remarkably informative bounds on several interesting target parameters while using an empirically tractable methodology. Our application to the welfare impacts of sales taxes showed that while some features of a system, such as the average supply elasticity, may be partially identified with relatively uninformative bounds, one can still obtain tight inference on other features of the system, such as deadweight loss and consumer incidence.

It is worth emphasizing the importance of a partial identification approach for obtaining these findings. The necessary results for point identification developed by Masten (2018) are quite stark. They are certainly not satisfied in the Fulton fish market example and likely not satisfied in our sales tax application either. However, Masten's results do not say by how "much" point identification fails. Partial identification emerges as an essential tool for quantifying the extent to which a failure of point identification really indicates a lack of information, rather than simply a lack of uniqueness. For further discussion, see

Manski (2007), Tamer (2010), Ho and Rosen (2017), or Molinari (2020).

Other commonly used models of demand (and supply) place strong restrictions on market heterogeneity. A prominent example in recent literature is the discrete choice model for differentiated goods developed by Berry (1994) and Berry et al. (1995). These models have limited unobserved heterogeneity across markets but are otherwise more complicated than the single-good linear model that we studied. They are likely point identified (Berry and Haile, 2014) and always implemented under this premise. It stands to reason that incorporating unobserved heterogeneity into these models will also render them partially identified. Whether they remain informative requires further research.

References

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67, 499–527.

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

Balke, A. and J. Pearl (1994): "Counterfactual Probabilities: Computational Methods, Bounds, and Applications," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-94)*, ed. by R. Lopez de Mantras and D. Poole, 46–54.

Beran, R. and P. Hall (1992): "Estimating Coefficient Distributions in Random Coefficient Regressions," *The Annals of Statistics*, 20, 1970–1984.

Bergquist, L. F. and M. Dinerstein (2020): "Competition and Entry in Agricultural Markets: Experimental Evidence from Kenya," *American Economic Review*, 110, 3705–3747.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Berry, S. T. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 25, 242–262.

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

——— (2018): "Identification of Nonparametric Simultaneous Equations Models With a Residual Index Structure," *Econometrica*, 86, 289–315.

——— (2021): "Chapter 1 - Foundations of Demand Estimation," in *Handbook of Industrial Organization*, ed. by K. Ho, A. Hortaçsu, and A. Lizzeri, Elsevier, vol. 4 of *Handbook of Industrial Organization, Volume 4*, 1–62.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2025): "When Is TSLS Actually LATE?" Tech. Rep. w29709, National Bureau of Economic Research, Cambridge, MA.

Bronfenbrenner, J. (1953): "Sources and Size of Least-Squares Bias in a Two-Equation Model," *Studies in Econometric Method, Cowles Commission Monograph*, 14, 221–235.

Cai, B. and R. Meyer (2011): "Bayesian Semiparametric Modeling of Survival Data Based on Mixtures of -Spline Distributions," *Computational Statistics & Data Analysis*, 55, 1260–1272.

Chen, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 5549–5632.

Chernozhukov, V., B. Deaner, Y. Gao, J. Hausman, and W. K. Newey (2025): "Linear Estimation of Structural and Causal Effects for Nonseparable Panel Data," .

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

Chetty, R. (2009): "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Review of Economics*, 1, 451–488.

Coglianese, J., L. W. Davis, L. Kilian, and J. H. Stock (2017): "Anticipation, Tax Avoidance, and the Price Elasticity of Gasoline Demand," *Journal of Applied Econometrics*, 32, 1–15.

Dearing, A. (2022): "Estimating Structural Demand and Supply Models Using Tax Rates as Instruments," *Journal of Public Economics*, 205, 104561.

Diaconis, P. and D. Ylvisaker (1985): "Quantifying Prior Opinion," in *Bayesian Statistics 2*, ed. by J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Amsterdam: North-Holland, 133–156.

Dubé, J.-P. (2019): "Microeconometric Models of Consumer Demand," in *Handbook of the Economics of Marketing*, Elsevier, vol. 1, 1–68.

Fang, Z., A. Santos, A. M. Shaikh, and A. Torgovitsky (2023): "Inference for Large-Scale Linear Systems With Known Coefficients," *Econometrica*, 91, 299–327.

Fisher, F. M. (1966): *The Identification Problem in Econometrics*, New York: McGraw-Hill.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206.

Frangakis, C. E. and D. B. Rubin (2002): "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Freyberger, J. and J. L. Horowitz (2015): "Identification and Shape Restrictions in Nonparametric Instrumental Variables Estimation," *Journal of Econometrics*, 189, 41–53.

Gaarder, I. and L. Henry de Frahan (2025): "The Welfare Effect of Marginal and Nonmarginal Changes in Sales Taxes in the U.S." .

Gaillac, C. and E. Gautier (2022): "Adaptive Estimation in the Linear Random Coefficients Model When Regressors Have Limited Variation," *Bernoulli*, 28, 504–524.

Gandhi, A. and A. Nevo (2021): "Empirical Models of Demand and Supply in Differentiated Products Industries," in *Handbook of Industrial Organization*, Elsevier, vol. 4, 63–139.

Gehringer, K. R. and R. A. Redner (1992): "Nonparametric Probability Density Estimation Using Normalized b–Splines," *Communications in Statistics - Simulation and Computation*, 21, 849–878.

Goldberger, A. S. (1972): "Structural Equation Methods in the Social Sciences," *Econometrica*, 40, 979.

Graddy, K. (1995): "Testing for Imperfect Competition at the Fulton Fish Market," *The RAND Journal of Economics*, 26, 75–92.

Haavelmo, T. (1943): "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 11, 1–12.

Hackmann, M. B., J. T. Kolstad, and A. E. Kowalski (2015): "Adverse Selection and an Individual Mandate: When Theory Meets Practice," *American Economic Review*, 105, 1030–1066.

Hahn, J. (2001): "Consistent Estimation of the Random Structural Coefficient Distribution from the Linear Simultaneous Equations System," *Economics Letters*, 73, 227–231.

Handbury, J. and D. E. Weinstein (2015): "Goods Prices and Availability in Cities," *The Review of Economic Studies*, 82, 258–296.

Hansen, L. P., J. Heaton, and E. G. J. Luttmer (1995): "Econometric Evaluation of Asset Pricing Models," *The Review of Financial Studies*, 8, 237–274.

Harberger, A. C. (1964): "The Measurement of Waste," *The American Economic Review*, 54, 58–76.

Hazell, J., J. Herreño, E. Nakamura, and J. Steinsson (2022): "The Slope of the Phillips Curve: Evidence from U.S. States," *The Quarterly Journal of Economics*, 137, 1299–1344.

Heckman, J. and E. Vytlacil (1998): "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return Is Correlated with Schooling," *The Journal of Human Resources*, 33, 974–987.

Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*.

——— (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *The Journal of Political Economy*, 109, 673–748.

Heckman, J. J. and R. Pinto (2018): "Unordered Monotonicity," *Econometrica*, 86, 1–35.
Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389–432.
Hermann, P. and H. Holzmann (2024): "Bounded Support in Linear Random Coefficient Models: Identification and Variable Selection," *Econometric Theory*, 1–30.
Ho, K. and A. M. Rosen (2017): "Partial Identification in Applied Research: Benefits and Challenges," in *Advances in Economics and Econometrics*, ed. by B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, 307–359.
Hoderlein, S., H. Holzmann, and A. Meister (2017): "The Triangular Model with Random Coefficients," *Journal of Econometrics*, 201, 144–169.
Hoderlein, S., J. Klemelä, and E. Mammen (2010): "Analyzing the Random Coefficient Model Nonparametrically," *Econometric Theory*, 26, 804–837.
Hughes, J. E., C. R. Knittel, and D. Sperling (2008): "Evidence of a Shift in the Short-Run Price Elasticity of Gasoline Demand," *The Energy Journal*, 29, 113–134.
Hurwicz, L. (1950): "Systems with Nonadditive Disturbances," in *Cowles 10*, ed. by T. Koopmans, no. 10 in Cowles Commission Monographs, 410–418.
Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.
Kang, Z. Y. and S. Vasserman (2025): "Robustness Measures for Welfare Analysis," *American Economic Review*, 115, 2449–2487.
Kelejian, H. H. (1974): "Random Parameters in a Simultaneous Equation Framework: Identification and Estimation," *Econometrica*, 42, 517–527.
Kilian, L. (2009): "Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market," *American Economic Review*, 99, 1053–1069.
Koopmans, T. C. and W. C. Hood (1953): "The Estimation of Simultaneous Linear Economic Relationships," *Studies in Econometric Method, Cowles Commission Monograph*, 14, 112–199.
Kroft, K., J.-W. P. Laliberté, R. Leal-Vizcaíno, and M. J. Notowidigdo (2024a): "Efficiency and Incidence of Taxation with Free Entry and Love-of-Variety Preferences," *American Economic Journal: Economic Policy*, 16, 300–334.
——— (2024b): "Salience and Taxation with Imperfect Competition," *Review of Economic Studies*, 91, 403–437.
Laffférs, L. (2013): "Essays in Partial Identification," Ph.D. thesis, Norwegian School of Economics.
Lee, S. C. K. and X. S. Lin (2010): "Modeling and Evaluating Insurance Losses Via Mixtures of Erlang Distributions," *North American Actuarial Journal*, 14, 107–130.
Lewbel, A. and K. Pendakur (2017): "Unobserved Preference Heterogeneity in Demand Using Generalized Random Coefficients," *Journal of Political Economy*, 125, 1100–1148.
Manski, C. F. (1994): *Simultaneity With Downward Sloping Demand*, University of Wisconsin-Madison. Social Systems Research Institute.
——— (1995): *Identification Problems in the Social Sciences*, Harvard University Press.
——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.
——— (2007): *Identification for Prediction and Decision*, Harvard University Press.
Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.
Masten, M. A. (2018): "Random Coefficients on Endogenous Variables in Simultaneous Equations Models," *The Review of Economic Studies*, 85, 1193–1250.
Masten, M. A. and A. Torgovitsky (2016): "Identification of Instrumental Variable Correlated Random Coefficients Models," *Review of Economics and Statistics*, 98, 1001–1005.
Matsuda, Y. and R. Iwafuchi (2025): "Density-Valued ARMA Models by Spline Mixtures," *Working Paper*.
Matzkin, R. L. (2008): "Identification in Nonparametric Simultaneous Equations Models," *Econometrica*, 76, 945–978.
——— (2015): "Estimation of Nonparametric Models With Simultaneity," *Econometrica*, 83, 1–66.
McLeay, M. and S. Tenreyro (2020): "Optimal Inflation and the Identification of the Phillips

Curve," *NBER Macroeconomics Annual*, 34, 199–255.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters," *Econometrica*, 86, 1589–1619.

Mogstad, M. and A. Torgovitsky (2024): "Instrumental Variables with Unobserved Heterogeneity in Treatment Effects," in *Handbook of Labor Economics*, Elsevier, vol. 5, 1–114.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2021): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *American Economic Review*, 111, 3663–3698.

——— (2024): "Policy Evaluation with Multiple Instrumental Variables," *Journal of Econometrics*, 105718.

Molinari, F. (2020): "Microeconometrics with Partial Identification," *arXiv:2004.11751 [econ]*.

Perron, F. and K. Mengersen (2001): "Bayesian Nonparametric Modeling Using Mixtures of Triangular Distributions," *Biometrics*, 57, 518–528.

Robins, J. M. and S. Greenland (1992): "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143.

Saez, E., M. Matsaganis, and P. Tsakloglou (2012): "Earnings Determination and Taxes: Evidence From a Cohort-Based Payroll Tax Reform in Greece," *The Quarterly Journal of Economics*, 127, 493–533.

Saiz, A. (2010): "The Geographic Determinants of Housing Supply*," *The Quarterly Journal of Economics*, 125, 1253–1296.

Schultz, H. (1938): *The Theory and Measurement of Demand*, University of Chicago Press Chicago.

Suárez Serrato, J. C. and O. Zidar (2016): "Who Benefits from State Corporate Tax Cuts? A Local Labor Markets Approach with Heterogeneous Firms," *American Economic Review*, 106, 2582–2624.

Tamer, E. (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195.

Tebaldi, P., A. Torgovitsky, and H. Yang (2023): "Nonparametric Estimates of Demand in the California Health Insurance Exchange," *Econometrica*, 91, 107–146.

Thorisson, H. (1995): "Coupling Methods in Probability Theory," *Scandinavian Journal of Statistics*, 22, 159–182.

Tinbergen, J. (1930/1995): "Determination and Interpretation of Supply Curves: An Example," in *The Foundations of Econometric Analysis*, ed. by D. F. Hendry and M. S. Morgan, Cambridge: Cambridge University Press, 669–679.

Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83, 1185–1197.

——— (2019a): "Nonparametric Inference on State Dependence in Unemployment," *Econometrica*, 87, 1475–1505.

——— (2019b): "Partial Identification by Extending Subdistributions," *Quantitative Economics*, 10, 105–144.

Windmeijer, F. (2019): "Two-Stage Least Squares as Minimum Distance," *The Econometrics Journal*, 22, 1–9.

Wright, P. G. (1915): "Review of Economic Cycles by Henry Moore," *Quarterly Journal of Economics*, 29, 631–41.

——— (1928): *The Tariff on Animal and Vegetable Oils*, 26, Macmillan.

Zoutman, F. T., E. Gavrilova, and A. O. Hopland (2018): "Estimating Both Supply and Demand Elasticities Using Variation in a Single Tax Rate," *Econometrica*, 86, 763–771.

Supplemental Appendix

Table of Contents

SA.1 Proofs for Section 2 ..... S2

SA.2 Derivation of Welfare Target Parameters S4

SA.3 Reverse Engineering Linear IV Estimators S6

SA.3.1 Weighting expressions for the 2SLS estimand S6

SA.3.2 Attenuation of the 2SLS estimand S10

SA.3.3 Nonparametric attenuation of the Wald estimand S11

SA.4 Incorporating Additional Nonlinear Parameters S13

SA.4.1 Profiled optimization S13

SA.4.2 Constant coefficients S13

SA.4.3 Overidentification with constant coefficients S15

SA.5 Estimation and Statistical Inference S15

SA.5.1 Criterion function for the full information approach S16

SA.5.2 Criterion function for the limited information approach S16

SA.5.3 Estimation S17

SA.5.4 Choosing the tuning parameter S18

SA.5.5 Statistical inference S19

SA.6 Algebra for the Limited Information Approach S20

SA.7 Microfounding a Linear Random Coefficients Model S21

SA.7.1 Demand S21

SA.7.2 Supply S23

SA.8 Application Details S25

SA.8.1 Data construction S25

SA.8.2 Basis specification S26

SA.8.3 Reduced form moments S26

SA.9 Additional Exhibits S27

SA.1 Proofs for Section 2

Proof of Proposition 1. First, suppose that $\pi \in \Pi^\dagger$. Then there exists an $F \in \mathcal{F}^\dagger$ such that $\pi = \pi_F$. Suppose that $m \notin \mathcal{M}^\dagger$. The definition of $\mathcal{M}^\dagger$ implies that if there is a $\boldsymbol{q}$ such that $\mu(\boldsymbol{q}) = m$, then $\boldsymbol{q} \notin \boldsymbol{\mathcal{Q}}^\dagger$. The definition of $\mathcal{F}^\dagger$ as $\mathbb{P}_F[\boldsymbol{Q} \in \boldsymbol{\mathcal{Q}}^\dagger] = 1$ then implies that

$$\pi_F(m) \equiv \mathbb{P}_F[\mu(\boldsymbol{Q}) = m] \leq \mathbb{P}_F[\boldsymbol{Q} \notin \boldsymbol{\mathcal{Q}}^\dagger] = 0,$$

and therefore that $\pi_F(m) = 0$ for all $m \notin \mathcal{M}^\dagger$.

Conversely, suppose that $\pi \in \Pi$ is such that $\pi(m) = 0$ for all $m \notin \mathcal{M}^\dagger$. Then for each $m \in \mathcal{M}^\dagger$, there exists at least one $\boldsymbol{q} \in \boldsymbol{\mathcal{Q}}^\dagger$ such that $\mu(\boldsymbol{q}) = m$; select a unique one $\tilde{\boldsymbol{q}}(m)$ for each m. Let F be a discrete distribution defined as $\mathbb{P}_F[\boldsymbol{Q} = \boldsymbol{q}] = \sum_{m \in \mathcal{M}} \mathbb{1}[\boldsymbol{q} = \tilde{\boldsymbol{q}}(m)]\pi(m)$. Then for all $m \in \mathcal{M}^\dagger$,

$$\pi_F(m) \equiv \mathbb{P}_F[\mu(\boldsymbol{Q}) = m] = \mathbb{P}_F[\boldsymbol{Q} = \tilde{\boldsymbol{q}}(m)] = \pi(m),$$

while $\pi_F(m) = \pi(m) = 0$ for all $m \notin \mathcal{M}^\dagger$. We conclude that $\pi \in \Pi^\dagger$. Q.E.D.

Proof of Proposition 2. Let $\Pi^\star$ denote the feasible region of (NP-LP). We begin by showing that $\mathcal{F}^\star$ is non-empty if and only if $\Pi^\star$ is non-empty.

Suppose first that $\mathcal{F}^\star$ is non-empty and let $F^\star \in \mathcal{F}^\star$. Let $\pi^\star(m) \equiv \pi_{F^\star}(m) \equiv \mathbb{P}_{F^\star}[\mu(\boldsymbol{Q}) = m]$. Then $\pi^\star \in \Pi^\dagger$ by definition of $\mathcal{F}^\dagger$ and because $F^\star \in \mathcal{F}^\star \subseteq \mathcal{F}^\dagger$. In addition, because $F^\star \in \mathcal{F}^\star$,

$$\begin{aligned}\sum_{m \in \mathcal{M}} \mathbb{1}[m_z = y]\pi^\star(m) &= \sum_{m \in \mathcal{M}} \mathbb{1}[m_z = y]\, \mathbb{P}_{F^\star}[\mu(\boldsymbol{Q}) = m] \\ &= \mathbb{P}_{F^\star}[\mu(\boldsymbol{Q}) \in \{m : m_z = y\}] \\ &= \mathbb{P}_{F^\star}[Y(z) = y] = \mathbb{P}[Y = y | Z = z],\end{aligned}$$

so that $\pi^\star \in \Pi^\star$. Conclude that $\Pi^\star$ is non-empty.

Conversely, suppose that $\Pi^\star$ is non-empty and let $\pi^\star \in \Pi^\star$. Then let $F^\star$ be a discrete distribution defined as

$$\mathbb{P}_{F^\star}[\boldsymbol{Q} = \boldsymbol{q}] = \sum_{m \in \mathcal{M}^\dagger} \mathbb{1}[\boldsymbol{q} = \boldsymbol{q}^\star(m)]\pi^\star(m), \tag{35}$$

where $\boldsymbol{q}^\star(m)$ is any element of $\boldsymbol{\mathcal{Q}}^\dagger$ such that $\mu(\boldsymbol{q}^\star(m)) = m$. The existence of $\boldsymbol{q}^\star(m)$ for $m \in \mathcal{M}^\dagger$ is ensured by $\pi^\star \in \Pi^\star \subseteq \Pi^\dagger$ and the definition of $\mathcal{M}^\dagger$. Then $F^\star \in \mathcal{F}^\dagger$ because

$\boldsymbol{q}^\star(m) \in \boldsymbol{\mathcal{Q}}^\dagger$ for all m. In addition, $F^\star \in \mathcal{F}^\star$ because

$$\begin{aligned}\mathbb{P}_{F^\star}[Y(z) = y] &= \mathbb{P}_{F^\star}[\mu(\boldsymbol{Q}) \in \{m : m_z = y\}] \\ &= \sum_{m \in \mathcal{M}} \mathbb{1}[m_z = y]\, \mathbb{P}_{F^\star}[\boldsymbol{Q} = \boldsymbol{q}^\star(m)] \\ &= \sum_{m \in \mathcal{M}} \mathbb{1}[m_z = y]\pi^\star(m) = \mathbb{P}[Y = y | Z = z],\end{aligned}$$

where the final equality follows from $\pi^\star \in \Pi^\star$. Conclude that $\mathcal{F}^\star$ is non-empty.

Now we show that $\tau^\star_{\text{ub}} = \bar{t}^\star$ by showing that $\bar{t}^\star \geq \tau^\star_{\text{ub}}$ and $\bar{t}^\star \leq \tau^\star_{\text{ub}}$. Symmetric arguments apply to the lower bounds. If either $\mathcal{F}^\star$ or $\Pi^\star$ is empty, then both are empty, as just shown, so $\tau^\star_{\text{ub}} = \bar{t}^\star = -\infty$, and the claim is established. So, assume that both $\mathcal{F}^\star$ and $\Pi^\star$ are non-empty in the following.

We begin by showing that $\bar{t}^\star \geq \tau^\star_{\text{ub}}$. Suppose first that $\tau^\star_{\text{ub}}$ is finite and let $\epsilon > 0$ be arbitrary. By the definition of a supremum and the assumption that $\mathcal{F}^\star$ is non-empty, there exists an $F^\star \in \mathcal{F}^\star$ such that $\mathbb{E}_{F^\star}[\tau(\boldsymbol{Q})] \geq \tau^\star_{\text{ub}} - \epsilon$. Let $\pi^\star(m) \equiv \pi_{F^\star}(m)$, which we established above was an element of $\Pi^\star$. As such, it is a feasible solution of (NP-LP), which implies that

$$\begin{aligned}\bar{t}^\star \geq \sum_{m \in \mathcal{M}} \bar{t}(m)\pi^\star(m) &= \sum_{m \in \mathcal{M}} \bar{t}(m)\, \mathbb{P}_{F^\star}[\mu(\boldsymbol{Q}) = m] \\ &\geq \sum_{m \in \mathcal{M}} \mathbb{E}_{F^\star}[\tau(\boldsymbol{Q}) | \mu(\boldsymbol{Q}) = m]\, \mathbb{P}_{F^\star}[\mu(\boldsymbol{Q}) = m] = \mathbb{E}_{F^\star}[\tau(\boldsymbol{Q})],\end{aligned} \tag{36}$$

where the second inequality follows from the definition of $\bar{t}(m)$ as a supremum. We conclude that $\bar{t}^\star \geq \mathbb{E}_{F^\star}[\tau(\boldsymbol{Q})] \geq \tau^\star_{\text{ub}} - \epsilon$. Because $\epsilon > 0$ was arbitrary, this implies that $\bar{t}^\star \geq \tau^\star_{\text{ub}}$.

Now consider the case that $\tau^\star_{\text{ub}} = \infty$ is infinite. Then for any $K > 0$, there exists an $F^\star_K \in \mathcal{F}^\star$ such that $\mathbb{E}_{F^\star_K}[\tau(\boldsymbol{Q})] > K$. Let $\pi^\star_K(m) \equiv \pi_{F^\star_K}(m)$, which is again an element of $\Pi^\star$, so feasible in (NP-LP), and satisfies (36) with $\pi^\star_K$ and $F^\star_K$ in place of $\pi^\star$ and $F^\star$. So, $\bar{t}^\star \geq \mathbb{E}_{F^\star_K}[\tau(\boldsymbol{Q})] > K$. Because $K > 0$ was arbitrary, we conclude that $\bar{t}^\star = \infty$, so that again $\bar{t}^\star \geq \tau^\star_{\text{ub}}$.

For the other inequality, start by considering the case that there exists an optimal solution $\pi^\star$ to (NP-LP), so that $\bar{t}^\star$ is finite. Let $\mathcal{M}^\dagger_+ \equiv \{m : \pi^\star(m) > 0\}$ denote the set of market types to which $\pi^\star$ assigns strictly positive mass. Then $\bar{t}(m)$ is finite for all $m \in \mathcal{M}^\dagger_+$. Let $\epsilon > 0$ be arbitrary. For each $m \in \mathcal{M}^\dagger_+$, take any $\boldsymbol{q}^\star(m) \in \boldsymbol{\mathcal{Q}}^\dagger$ that satisfies both $\mu(\boldsymbol{q}^\star(m)) = m$ and $\tau(\boldsymbol{q}^\star(m)) \geq \bar{t}(m) - \epsilon$. The existence of $\boldsymbol{q}^\star(m)$ follows from $\bar{t}(m)$ being the supremum in the market type problems (18), which is finite for $m \in \mathcal{M}^\dagger_+$. Let $F^\star$ be the discrete distribution defined from $\boldsymbol{q}^\star(m)$ and $\pi^\star(m)$ as in (35). Then $F^\star \in \mathcal{F}^\star$, as established above, so that $\tau^\star_{\text{ub}} \geq \mathbb{E}_{F^\star}[\tau(\boldsymbol{Q})]$. In addition, we have also constructed $F^\star$

so that

$$\begin{aligned}\tau^\star_{\text{ub}} \geq \mathbb{E}_{F^\star}[\tau(\boldsymbol{Q})] &= \sum_{m\in\mathcal{M}^\dagger_+} \mathbb{E}_{F^\star}[\tau(\boldsymbol{Q})|M=m]\,\mathbb{P}_{F^\star}[M=m] \\ &= \sum_{m\in\mathcal{M}^\dagger_+} \tau(\boldsymbol{q}^\star(m))\pi^\star(m) \\ &\geq \sum_{m\in\mathcal{M}^\dagger_+} \bar{t}(m)\pi^\star(m) - \epsilon \sum_{m\in\mathcal{M}^\dagger_+} \pi^\star(m) \\ &= \sum_{m\in\mathcal{M}} \bar{t}(m)\pi^\star(m) - \epsilon = \bar{t}^\star - \epsilon.\end{aligned}$$

Because $\epsilon > 0$ was arbitrarily small, we conclude that $\tau^\star_{\text{ub}} \geq \bar{t}^\star$.

Now consider the case when (NP-LP) has no solution. Because $\Pi^\star$ is non-empty by assumption, compact by construction, and the objective of (NP-LP) is linear, this can only happen if at least one of the coefficients $\bar{t}(m)$ in the linear objective is infinite. Divide $\mathcal{M}^\dagger \equiv \mathcal{M}^\dagger_0 \cup \mathcal{M}^\dagger_\infty$, where $\mathcal{M}^\dagger_0$ contains all $m \in \mathcal{M}^\dagger$ such that $\bar{t}(m)$ is finite and $\mathcal{M}^\dagger_\infty$ contains all $m \in \mathcal{M}^\dagger$ such that $\bar{t}(m) = \infty$. Let $K > 0$ be arbitrary. Then there exists a $\pi^\star_K \in \Pi^\star$ such that $\sum_{m\in\mathcal{M}^\dagger} \bar{t}(m)\pi^\star_K(m) > K$. For all $m \in \mathcal{M}^\dagger$, let $\boldsymbol{q}^\star_K(m)$ be such that $\mu(\boldsymbol{q}^\star_K(m)) = m$. Additionally, for $m \in \mathcal{M}^\dagger_\infty$, choose this $\boldsymbol{q}^\star_K(m)$ such that

$$\tau(\boldsymbol{q}^\star_K(m)) > \frac{K - K_0}{\sum_{m'\in\mathcal{M}^\dagger_\infty} \pi^\star_K(m')} \quad \text{where} \quad K_0 \equiv \sum_{m\in\mathcal{M}^\dagger_0} \tau(\boldsymbol{q}^\star_K(m))\pi^\star_K(m).$$

Set $F^\star_K$ to be the discrete distribution that places mass $\pi^\star_K(m)$ on $\boldsymbol{q}^\star_K(m)$, the same as in (35). Then $F^\star_K \in \mathcal{F}^\star$, as previously established. In addition,

$$\begin{aligned}\mathbb{E}_{F^\star_K}[\tau(\boldsymbol{Q})] &= \sum_{m\in\mathcal{M}^\dagger} \mathbb{E}_{F^\star_K}[\tau(\boldsymbol{Q})|M=m]\,\mathbb{P}_{F^\star_K}[M=m] \\ &= \sum_{m\in\mathcal{M}^\dagger_0} \tau(\boldsymbol{q}^\star_K(m))\pi^\star_K(m) + \sum_{m\in\mathcal{M}^\dagger_\infty} \tau(\boldsymbol{q}^\star_K(m))\pi^\star_K(m) \\ &> K_0 + \sum_{m\in\mathcal{M}^\dagger_\infty} \left(\frac{K - K_0}{\sum_{m'\in\mathcal{M}^\dagger_\infty} \pi^\star_K(m')}\right) \pi^\star_K(m) = K.\end{aligned}$$

Because $K > 0$ was arbitrary, this shows that $\tau^\star_{\text{ub}} = \infty$, so that $\tau^\star_{\text{ub}} \geq \bar{t}^\star = \infty$.

Q.E.D.

SA.2 Derivation of Welfare Target Parameters

We focus on the case where Q and P are specified in logs, because this is what we use in our application. Similar expressions can be derived when Q and P are specified in levels. For notation, we let $q^{\text{E}} \equiv \exp(Q^{\text{E}})$ and $p^{\text{E}} \equiv \exp(P^{\text{E}})$ be equilibrium quantity and price in levels. Note that in this section p^{E} is the before-tax price.

We consider an ad valorem tax with rate r and let $\theta \equiv \log(1+r)$. Assuming the tax θ is paid by consumers, the equilibrium price and quantity as a function of the tax are

$$p^{\mathrm{E}}(\theta) = \exp\left(Z'\left(\frac{B_{\mathrm{Z}}^{\mathrm{D}} - B_{\mathrm{Z}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right) - \frac{B_{\mathrm{P}}^{\mathrm{D}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\theta\right) \tag{37}$$

$$q^{\mathrm{E}}(\theta) = \exp\left(Z'\left(\frac{B_{\mathrm{P}}^{\mathrm{S}} B_{\mathrm{Z}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{D}} B_{\mathrm{Z}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right) - \frac{B_{\mathrm{P}}^{\mathrm{D}} B_{\mathrm{P}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\theta\right). \tag{38}$$

Equilibrium sales as a function of θ are therefore:

$$\textsc{sal}(\theta) \equiv p^{\mathrm{E}}(\theta)\, q^{\mathrm{E}}(\theta) = \exp\left(Z'\left(\frac{B_{\mathrm{Z}}^{\mathrm{D}}(1 + B_{\mathrm{P}}^{\mathrm{S}}) - B_{\mathrm{Z}}^{\mathrm{S}}(1 - B_{\mathrm{P}}^{\mathrm{D}})}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right) - \frac{B_{\mathrm{P}}^{\mathrm{D}}(1 + B_{\mathrm{P}}^{\mathrm{S}})}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\theta\right). \tag{39}$$

And government revenue is:

$$\textsc{rev}(\theta) = r \times \textsc{sal}(\theta) = (\exp(\theta) - 1)\textsc{sal}(\theta). \tag{40}$$

The differences in consumer and producer surplus from a tax of r relative to no taxes are given by

$$\textsc{cs}(\theta) \equiv \int_{(1+r)p^{\mathrm{E}}(\theta)}^{p^{\mathrm{E}}(0)} \exp(Z' B_{\mathrm{Z}}^{\mathrm{D}} - B_{\mathrm{P}}^{\mathrm{D}} \log(x))\, dx \tag{41}$$

$$\textsc{ps}(\theta) \equiv \int_{p^{\mathrm{E}}(0)}^{p^{\mathrm{E}}(\theta)} \exp(Z' B_{\mathrm{Z}}^{\mathrm{S}} + B_{\mathrm{P}}^{\mathrm{S}} \log(x))\, dx. \tag{42}$$

Deadweight loss from the tax is then given by

$$\textsc{dwl}(\theta) = -\left(\textsc{cs}(\theta) + \textsc{ps}(\theta) + \textsc{rev}(\theta)\right). \tag{43}$$

Using the above expressions, the change in consumer and producer surplus and revenue from a marginal tax increase can be shown through Leibniz's rule to be given by

$$\begin{aligned}
\textsc{cs}'(\theta) &\equiv -(1+r)\textsc{sal}(\theta)\frac{B_{\mathrm{P}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}} \\
\textsc{ps}'(\theta) &\equiv -\textsc{sal}(\theta)\frac{B_{\mathrm{P}}^{\mathrm{D}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}} \\
\textsc{rev}'(\theta) &\equiv \textsc{sal}(\theta)\left(1 + r\frac{B_{\mathrm{P}}^{\mathrm{S}}(1 - B_{\mathrm{P}}^{\mathrm{D}})}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right).
\end{aligned}$$

The corresponding change in deadweight loss is then

$$\textsc{dwl}'(\theta) \equiv -\left(\textsc{cs}'(\theta) + \textsc{ps}'(\theta) + \textsc{rev}'(\theta)\right) = r\textsc{sal}(\theta)\frac{B_{\mathrm{P}}^{\mathrm{D}} B_{\mathrm{P}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}.$$

Usually, we normalize the change in deadweight loss by the corresponding change in

revenue and consider the quantity

$$\overline{\text{DWL}}'(\theta) \equiv \frac{\text{DWL}'(\theta)}{\text{REV}'(\theta)} = \frac{r B_{\text{P}}^{\text{D}} B_{\text{P}}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}} + r\,(1 - B_{\text{P}}^{\text{D}})\,B_{\text{P}}^{\text{S}}}. \tag{44}$$

We also consider the marginal incidence on consumers of the tax, which is defined as the relative consumer surplus impact:

$$\text{ICD}(\theta) \equiv \frac{\text{CS}'(\theta)}{\text{CS}'(\theta) + \text{PS}'(\theta)} = \frac{(1+r)B_{\text{P}}^{\text{S}}}{B_{\text{P}}^{\text{D}} + (1+r)B_{\text{P}}^{\text{S}}}. \tag{45}$$

We can also compute the impacts of a non-marginal change in the tax from θ_0 to θ_1. This results in a non-marginal change in deadweight loss of

$$\text{DWL}(\theta_0 \to \theta_1) \equiv \text{DWL}(\theta_1) - \text{DWL}(\theta_0) = \int_{\theta_0}^{\theta_1} \text{DWL}'(\theta)\,d\theta. \tag{46}$$

Normalizing this against the change in government revenue gives

$$\overline{\text{DWL}}(\theta_0 \to \theta_1) \equiv \frac{\text{DWL}(\theta_1) - \text{DWL}(\theta_0)}{\text{REV}(\theta_1) - \text{REV}(\theta_0)} = \frac{\int_{\theta_0}^{\theta_1} \text{DWL}'(\theta)\,d\theta}{\int_{\theta_0}^{\theta_1} \text{REV}'(\theta)\,d\theta}, \tag{47}$$

which is a complicated function of both supply and demand slopes, as well as the two tax levels. Similar arguments can be used to derive expressions for discrete changes in other quantities, such as consumer and producer surplus.

SA.3 Reverse Engineering Linear IV Estimators

In this section, we provide details on the results discussed in Section 3.3 about interpreting a linear IV estimator through the lens of a model with unobserved heterogeneity in price elasticities. Sections SA.3.1 and SA.3.2 provide conditions under which a general 2SLS estimand is a non-negatively weighted average of heterogeneous demand slopes when the actual model is a linear random coefficients model. We express the linear random coefficients model using the notation (RC) introduced in Section 4. Section SA.3.3 shows that similar results continue to hold for the Wald estimand if the actual model is nonparametric.

SA.3.1 Weighting expressions for the 2SLS estimand

Let β_{2SLS} denote the two-stage least squares estimand with outcome variable Q, endogenous variable P, and $Z \equiv (Z_1, Z_2)$ divided into excluded variables (instruments), Z_1, and included covariates, Z_2. An application of the Frisch-Waugh Theorem shows that

$$\beta_{\text{2SLS}} = \frac{\mathbb{E}[Q(\dot{P} - \mathbb{L}[\dot{P}|Z_2])]}{\mathbb{E}[P(\dot{P} - \mathbb{L}[\dot{P}|Z_2])]}, \tag{48}$$

where $\mathbb{L}[\cdot|\cdot]$ denotes the linear projection (population fitted values) from regressing the first argument onto the second, so that $\dot{P} \equiv \mathbb{L}[P|Z]$ are the population fitted values from regressing P onto Z. The next proposition shows that β_{2SLS} will be a weighted average of the demand slope if Z_1 are supply shifters that are excluded from the demand equation.

Proposition SA.1. Divide $Z = (Z_1, Z_2)$ and assume that $B_{\text{Z}}^{\text{D}} = (0, B_{\text{Z},2}^{\text{D}})$, so that Z_1 corresponds to excluded supply shifters. Let $\tilde{Z}_1 \equiv Z_1 - \mathbb{L}[Z_1|Z_2]$ denote the vector of population residuals from linear regressions of each component of Z_1 onto Z_2. Then

$$-\beta_{\text{2SLS}} = \mathbb{E}\left[B_{\text{P}}^{\text{D}} \omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})\right]$$
$$\text{where} \quad \omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}}) \equiv \mathbb{E}\left[\frac{B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}} \middle| B_{\text{P}}^{\text{D}}\right]' \frac{\mathbb{E}[\tilde{Z}_1 \tilde{Z}_1' \delta_1]}{\mathbb{E}[(\tilde{Z}_1'\delta_1)^2]}, \quad \text{with} \quad \delta_1 \equiv \mathbb{E}\left[\frac{B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right].$$

The weights $\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})$ satisfy $\mathbb{E}[\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})] = 1$.

Note that the interpretation in Proposition SA.1 is for $-\beta_{\text{2SLS}}$ rather than β_{2SLS} simply because of our normalization in (RC) that B_{P}^{D} is non-negative.

Proof of Proposition SA.1. Because Z and B are independent and $B_{\text{Z},1}^{\text{D}} = 0$, we get from (25) that

$$\mathbb{E}[P|Z] = Z_1' \,\mathbb{E}\left[\frac{-B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right] + Z_2' \,\mathbb{E}\left[\frac{B_{\text{Z},2}^{\text{D}} - B_{\text{Z},2}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right] \equiv -Z_1'\delta_1 + Z_2'\delta_2. \tag{49}$$

This conditional mean is linear in Z, so the first-stage fitted values are $\dot{P} = \mathbb{E}[P|Z]$ and the residuals from projecting off Z_2 are

$$\dot{P} - \mathbb{L}[\dot{P}|Z_2] = -\tilde{Z}_1'\delta_1,$$

where $\tilde{Z}_1 \equiv Z_1 - \mathbb{L}[Z_1|Z_2]$. Then from (48),

$$\beta_{\text{2SLS}} = \frac{\mathbb{E}[Q(\dot{P} - \mathbb{L}[\dot{P}|Z_2])]}{\mathbb{E}[P(\dot{P} - \mathbb{L}[\dot{P}|Z_2])]} = \frac{\mathbb{E}[Q\tilde{Z}_1'\delta_1]}{\mathbb{E}[P\tilde{Z}_1'\delta_1]}.$$

Because Z_2 and $\tilde{Z}_1$ are orthogonal, the denominator simplifies to

$$\mathbb{E}[P\tilde{Z}_1'\delta_1] = \mathbb{E}\left[(-\delta_1'Z_1 + \delta_2'Z_2)\tilde{Z}_1'\delta_1\right] = -\,\mathbb{E}\left[\delta_1'Z_1\tilde{Z}_1'\delta_1\right] = -\,\mathbb{E}\left[(\tilde{Z}_1'\delta_1)^2\right].$$

For the numerator, first note that from (25), the independence of B and Z, and $B_{\text{Z},1}^{\text{D}} = 0$,

$$\mathbb{E}[Q|Z] = Z_1' \,\mathbb{E}\left[\frac{B_{\text{P}}^{\text{D}} B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right] + Z_2' \,\mathbb{E}\left[\frac{B_{\text{P}}^{\text{S}} B_{\text{Z},2}^{\text{D}} + B_{\text{P}}^{\text{D}} B_{\text{Z},2}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right] \equiv Z_1'\eta_1 + Z_2'\eta_2, \tag{50}$$

where

$$\eta_1 \equiv \mathbb{E}\left[\frac{B_{\text{P}}^{\text{D}} B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right] = \mathbb{E}\left[B_{\text{P}}^{\text{D}}\, \mathbb{E}\left[\frac{B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\Big| B_{\text{P}}^{\text{D}}\right]\right].$$

Using the orthogonality of $\tilde{Z}_1$ with Z_2, the numerator of β_{2SLS} can be written as

$$\mathbb{E}[(Z_1'\eta_1 + Z_2'\eta_2)\tilde{Z}_1'\delta_1] = \eta_1'\, \mathbb{E}[Z_1\tilde{Z}_1'\delta_1] = \mathbb{E}\left[B_{\text{P}}^{\text{D}}\, \mathbb{E}\left[\frac{B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\Big| B_{\text{P}}^{\text{D}}\right]'\, \mathbb{E}[\tilde{Z}_1\tilde{Z}_1'\delta_1]\right].$$

Combining numerator and denominator, we arrive at

$$-\beta_{\text{2SLS}} = \mathbb{E}\left[B_{\text{P}}^{\text{D}}\, \mathbb{E}\left[\frac{B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\Big| B_{\text{P}}^{\text{D}}\right]'\, \frac{\mathbb{E}[\tilde{Z}_1\tilde{Z}_1'\delta_1]}{\mathbb{E}[(\tilde{Z}_1'\delta_1)^2]}\right] \equiv \mathbb{E}[B_{\text{P}}^{\text{D}}\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})],$$

which is the claimed expression. To see that $\mathbb{E}[\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})] = 1$, note that

$$\mathbb{E}\left[\mathbb{E}\left[\frac{B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\Big| B_{\text{P}}^{\text{D}}\right]\right] = \mathbb{E}\left[\frac{B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right] \equiv \delta_1.$$

Q.E.D.

Interpreting the weights $\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})$ in Proposition SA.1 is easier when Z_1 is scalar, so that δ_1 is also scalar. In this case,

$$\frac{\mathbb{E}[\tilde{Z}_1\tilde{Z}_1'\delta_1]}{\mathbb{E}[(\tilde{Z}_1'\delta_1)^2]} = \frac{\mathbb{E}[\tilde{Z}_1^2]\delta_1}{\mathbb{E}[\tilde{Z}_1^2]\delta_1^2} = \frac{1}{\delta_1} \equiv \frac{1}{\mathbb{E}[B_{\text{Z},1}^{\text{S}}/(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})]},$$

and the weights reduce to

$$\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}}) = \frac{\mathbb{E}\left[B_{\text{Z},1}^{\text{S}}/(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})|B_{\text{P}}^{\text{D}}\right]}{\mathbb{E}[B_{\text{Z},1}^{\text{S}}/(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})]}. \tag{51}$$

This expression shows algebraically why β_{2SLS} is attenuated. The weights are proportional to the expected shift in prices for a market with demand slope B_{P}^{D}. The size of the shift depends directly on the impact of the supply shifter, $B_{\text{Z},1}^{\text{S}}$, but inversely on the total slopes of the supply and demand curves, $B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}$. If B_{P}^{D} is independent of both $B_{\text{Z},1}^{\text{S}}$ and B_{P}^{S}, then this implies that markets with larger values of B_{P}^{D} receive smaller weights. Deviations from independence can potentially overturn this conclusion.

Proposition SA.1 is interesting in the context of more general recent results on interpreting 2SLS estimands developed by Blandhol et al. (2025). Those authors show that 2SLS estimands for specifications that include covariates cannot in general be expressed as a weighted average of causal effects because they have "level dependence," meaning that they also depend on the levels of potential values of the outcome variable. They show that a necessary and sufficient condition to avoid level dependence is that the covariate

specification is rich enough to ensure that $\mathbb{L}[Z_1|Z_2] = \mathbb{E}[Z_1|Z_2]$. Proposition SA.1 shows that the random coefficients structure breaks this necessary condition, ensuring that β_{2SLS} is no longer level dependent. This finding reinforces a general theme of the results in Blandhol et al. (2025) that positive interpretations of the 2SLS estimand require some parametric structure. The random coefficients model provides that structure by implying reduced form equations ((49) and (50)) that are linear in Z.

While Proposition SA.1 shows that β_{2SLS} is a weighted average of B_{P}^{D}, it doesn't say anything about the sign of the weights. The next proposition provides two sufficient conditions for the weights to be non-negative.

Proposition SA.2. Suppose that the conditions of Proposition SA.1 are satisfied. Then $\mathbb{P}[\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}}) \geq 0] = 1$ if either

(a) Z_1 is scalar and either $\mathbb{P}[B_{\text{Z},1}^{\text{S}} \geq 0] = 1$ or $\mathbb{P}[B_{\text{Z},1}^{\text{S}} \leq 0] = 1$.

(b) $B_{\text{Z},1}^{\text{S}}$ is independent of $(B_{\text{P}}^{\text{D}}, B_{\text{P}}^{\text{S}})$.

Proof of Proposition SA.2. (a) If Z_1 is scalar, then the weights reduce to (51). Because $B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}$ is always non-negative, $\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})$ is also non-negative if $B_{\text{Z},1}^{\text{S}}$ only takes one sign.
(b) If $B_{\text{Z},1}^{\text{S}}$ is independent of $(B_{\text{P}}^{\text{D}}, B_{\text{P}}^{\text{S}})$, then $\delta_1 = \mathbb{E}[B_{\text{Z},1}^{\text{S}}]\,\mathbb{E}[1/(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})]$, so that

$$\mathbb{E}\left[\frac{B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\middle| B_{\text{P}}^{\text{D}}\right] = \mathbb{E}[B_{\text{Z},1}^{\text{S}}]\,\mathbb{E}\left[\frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\middle| B_{\text{P}}^{\text{D}}\right] = \delta_1 \frac{\mathbb{E}\left[1/(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})|B_{\text{P}}^{\text{D}}\right]}{\mathbb{E}\left[1/(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})\right]}.$$

It follows that

$$\omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}}) = \mathbb{E}\left[\frac{B_{\text{Z},1}^{\text{S}}}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\middle| B_{\text{P}}^{\text{D}}\right]' \frac{\mathbb{E}[\tilde{Z}_1\tilde{Z}_1'\delta_1]}{\mathbb{E}[(\tilde{Z}_1'\delta_1)^2]} = \frac{\mathbb{E}[1/(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})|B_{\text{P}}^{\text{D}}]}{\mathbb{E}[1/(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})]}, \tag{52}$$

which is always positive because $B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}$ is always positive. Q.E.D.

The condition in the first part of Proposition SA.2 is a random coefficients version of the instrument monotonicity condition discussed in Section 2.5. Requiring Z_1 to be scalar is important for the monotonicity condition to reflect a sensible ordering, yet vector instruments are commonly used in practice (Mogstad et al., 2021, 2024). For example, when Angrist et al. (2000) analyze the Fulton fish market data, they use three different types of binary weather instruments without commenting on the instrument monotonicity condition. The condition in the second part of Proposition SA.2 allows Z_1 to be a vector. It replaces the monotonicity condition with the assumption that the impact of the supply shifter is independent of the supply and demand slopes.

SA.3.2 Attenuation of the 2SLS estimand

The next proposition shows that if the second condition of Proposition SA.2 is satisfied, then β_{2SLS} will tend to understate the average slope of demand.

Proposition SA.3. Suppose that the conditions of Proposition SA.1 are satisfied and that condition (b) of Proposition SA.2 is satisfied. Then $-\beta_{\text{2SLS}} \leq \mathbb{E}[B_{\text{P}}^{\text{D}}]$ if and only if

$$\mathbb{C}\left[B_{\text{P}}^{\text{D}}, \frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right] \leq 0. \tag{53}$$

In particular, (53) is satisfied under either of the following two conditions:

(a) $\mathbb{E}[B_{\text{P}}^{\text{D}} \mid B_{\text{P}}^{\text{S}}] = \mathbb{E}[B_{\text{P}}^{\text{D}}]$.

(b) $\mathbb{E}[(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})^{-1} \mid B_{\text{P}}^{\text{D}} = b_{\text{P}}^{\text{D}}]$ is a weakly decreasing function of b_{P}^{D}.

Proof of Proposition SA.3. Recall (24):

$$-\beta_{\text{2SLS}} = \mathbb{E}[B_{\text{P}}^{\text{D}}] + \mathbb{C}[B_{\text{P}}^{\text{D}}, \omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})]. \tag{24}$$

We will show that the covariance term is negative under either of the two stated conditions.

Suppose that condition (b) of Proposition SA.2 is satisfied. Then (52) in the proof of Proposition SA.2 is satisfied. So

$$\mathbb{C}[B_{\text{P}}^{\text{D}}, \omega_{\text{2SLS}}(B_{\text{P}}^{\text{D}})] = \mathbb{C}\left[B_{\text{P}}^{\text{D}}, \mathbb{E}\left[\frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}} \middle| B_{\text{P}}^{\text{D}}\right]\right] \mathbb{E}\left[\frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right]^{-1}.$$

Because $B_{\text{P}}^{\text{D}}, B_{\text{P}}^{\text{S}}$ are assumed to be non-negative, the sign of this term is determined by the covariance, which simplifies to

$$\begin{aligned}
\mathbb{C}\left[B_{\text{P}}^{\text{D}}, \mathbb{E}\left[\frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}} \middle| B_{\text{P}}^{\text{D}}\right]\right] &= \mathbb{E}\left[(B_{\text{P}}^{\text{D}} - \mathbb{E}[B_{\text{P}}^{\text{D}}]) \, \mathbb{E}\left[\frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}} \middle| B_{\text{P}}^{\text{D}}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(B_{\text{P}}^{\text{D}} - \mathbb{E}[B_{\text{P}}^{\text{D}}]) \, \frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}} \middle| B_{\text{P}}^{\text{D}}\right]\right] \\
&= \mathbb{E}\left[(B_{\text{P}}^{\text{D}} - \mathbb{E}[B_{\text{P}}^{\text{D}}]) \, \frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right] = \mathbb{C}\left[B_{\text{P}}^{\text{D}}, \frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right].
\end{aligned} \tag{54}$$

Using (24), this shows that (53) is necessary and sufficient for $-\beta_{\text{2SLS}} \leq \mathbb{E}[B_{\text{P}}^{\text{D}}]$.

If $\mathbb{E}[(B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}})^{-1} \mid B_{\text{P}}^{\text{D}} = b_{\text{P}}^{\text{D}}]$ is a weakly decreasing function of b_{P}^{D}, then from (54),

$$\mathbb{C}\left[B_{\text{P}}^{\text{D}}, \frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}}\right] = -\,\mathbb{C}\left[B_{\text{P}}^{\text{D}}, -\,\mathbb{E}\left[\frac{1}{B_{\text{P}}^{\text{D}} + B_{\text{P}}^{\text{S}}} \middle| B_{\text{P}}^{\text{D}}\right]\right] \leq 0,$$

because the covariance of two weakly increasing functions of B_{P}^{D} is non-negative (for ex-

ample, Thorisson, 1995, Section 2). Alternatively, if B_{P}^{D} is mean independent of B_{P}^{S}, then

$$\begin{aligned}\mathbb{C}\left[B_{\text{P}}^{\text{D}}, \frac{1}{B_{\text{P}}^{\text{D}}+B_{\text{P}}^{\text{S}}}\right] &= \mathbb{E}\left[(B_{\text{P}}^{\text{D}}-\mathbb{E}[B_{\text{P}}^{\text{D}}])\frac{1}{B_{\text{P}}^{\text{D}}+B_{\text{P}}^{\text{S}}}\right] \\ &= \mathbb{E}\left[(B_{\text{P}}^{\text{D}}-\mathbb{E}[B_{\text{P}}^{\text{D}}])\left(\frac{1}{B_{\text{P}}^{\text{D}}+B_{\text{P}}^{\text{S}}}-\frac{1}{\mathbb{E}[B_{\text{P}}^{\text{D}}]+B_{\text{P}}^{\text{S}}}\right)\right], \end{aligned} \tag{55}$$

where the second equality follows because

$$\mathbb{E}\left[(B_{\text{P}}^{\text{D}}-\mathbb{E}[B_{\text{P}}^{\text{D}}])\frac{1}{\mathbb{E}[B_{\text{P}}^{\text{D}}]+B_{\text{P}}^{\text{S}}}\right] = \mathbb{E}\left[(\mathbb{E}[B_{\text{P}}^{\text{D}}|B_{\text{P}}^{\text{S}}]-\mathbb{E}[B_{\text{P}}^{\text{D}}])\frac{1}{\mathbb{E}[B_{\text{P}}^{\text{D}}]+B_{\text{P}}^{\text{S}}}\right] = 0.$$

Simplifying (55) then shows that

$$\mathbb{C}\left[B_{\text{P}}^{\text{D}}, \frac{1}{B_{\text{P}}^{\text{D}}+B_{\text{P}}^{\text{S}}}\right] = \mathbb{E}\left[(B_{\text{P}}^{\text{D}}-\mathbb{E}[B_{\text{P}}^{\text{D}}])\left(\frac{\mathbb{E}[B_{\text{P}}^{\text{D}}]-B_{\text{P}}^{\text{D}}}{(B_{\text{P}}^{\text{D}}+B_{\text{P}}^{\text{S}})(\mathbb{E}[B_{\text{P}}^{\text{D}}]+B_{\text{P}}^{\text{S}})}\right)\right] \leq 0.$$

Q.E.D.

Proposition SA.3 illustrates an important difference between simultaneous linear models with random coefficients and their simpler, triangular counterparts. The triangular model would replace the supply equation (for example) in (RC) with an equation like $P = Z'C$ for random coefficients C. Heckman and Vytlacil (1998) observed that if the component of C corresponding to the excluded instrument is in fact constant, then the linear IV estimand is equal to the average partial effect of the endogenous variable on the outcome. This type of reasoning no longer applies when there is simultaneity because the coefficient on the instrument in the reduced form for price is still heterogeneous due to heterogeneity in the slope of the demand equation. In this way, simultaneity with random coefficients can be interpreted as fundamentally creating "essential heterogeneity" in the terminology of Heckman et al. (2006).

SA.3.3 Nonparametric attenuation of the Wald estimand

In this section, we show that the basic argument and intuition behind Proposition SA.3 continues to apply to the Wald estimand even if the actual data generating process does not have a random coefficients structure. Our starting point is the expression (14) for the Wald estimand constructed from a binary instrument contrast:

$$\begin{aligned}\beta_{\text{WALD}} &\equiv \frac{\mathbb{E}[Q|Z=1]-\mathbb{E}[Q|Z=0]}{\mathbb{E}[P|Z=1]-\mathbb{E}[P|Z=0]} \\ &= \mathbb{E}\left[\left(\frac{P^{\text{E}}(1)-P^{\text{E}}(0)}{\mathbb{E}[P^{\text{E}}(1)-P^{\text{E}}(0)]}\right)\bar{\varepsilon}^{\text{D}}\right] = \mathbb{E}\left[\bar{\varepsilon}^{\text{D}}\right] + \frac{\mathbb{C}[P^{\text{E}}(1)-P^{\text{E}}(0), \bar{\varepsilon}^{\text{D}}]}{\mathbb{E}[P^{\text{E}}(1)-P^{\text{E}}(0)]}, \end{aligned}$$

where we suppress the market type (M) in the notation, because it is not relevant for the following argument. Assume that the monotonicity condition $P^{\text{E}}(1) \geq P^{\text{E}}(0)$ is satisfied.

Changing signs to be more comparable to the 2SLS discussion gives

$$-\beta_{\text{WALD}} \leq -\mathbb{E}[\bar{\varepsilon}^{\text{D}}] \quad \Leftrightarrow \quad \mathbb{C}[P^{\text{E}}(1) - P^{\text{E}}(0), -\bar{\varepsilon}^{\text{D}}] \leq 0. \tag{56}$$

We conclude that β_{WALD} will be an attenuated measure of average demand if and only if markets with larger equilibrium price changes have less elastic demand.

This result can be connected to the intuition discussed in the main text under the assumption that the instrument is an additive supply shifter. Suppose that

$$Q^{\text{S}}(p, z) = Q_0^{\text{S}}(p) + B_{\text{Z}}^{\text{S}} z \tag{57}$$

for some function $Q_0^{\text{S}}(p)$ that does not depend on z. Define the average slope of supply to be

$$\bar{\varepsilon}^{\text{S}} \equiv \frac{Q^{\text{S}}(P^{\text{E}}(1), z) - Q^{\text{S}}(P^{\text{E}}(0), z)}{P^{\text{E}}(1) - P^{\text{E}}(0)} = \frac{Q_0^{\text{S}}(P^{\text{E}}(1)) - Q_0^{\text{S}}(P^{\text{E}}(0))}{P^{\text{E}}(1) - P^{\text{E}}(0)},$$

noting that the second equality holds because of the separability in (57). Then

$$\begin{aligned}
\bar{\varepsilon}^{\text{D}} \left(P^{\text{E}}(1) - P^{\text{E}}(0)\right) &= Q^{\text{E}}(1) - Q^{\text{E}}(0) \\
&= Q^{\text{S}}(P^{\text{E}}(1), 1) - Q^{\text{S}}(P^{\text{E}}(0), 0) \\
&= Q_0^{\text{S}}(P^{\text{E}}(1)) - Q_0^{\text{S}}(P^{\text{E}}(0)) + B_{\text{Z}}^{\text{S}} = \bar{\varepsilon}^{\text{S}} \left(P^{\text{E}}(1) - P^{\text{E}}(0)\right) + B_{\text{Z}}^{\text{S}}.
\end{aligned}$$

Rearranging gives an expression for the change in equilibrium prices:

$$P^{\text{E}}(1) - P^{\text{E}}(0) = \frac{-B_{\text{Z}}^{\text{S}}}{\bar{\varepsilon}^{\text{S}} - \bar{\varepsilon}^{\text{D}}}. \tag{58}$$

Substituting into (56), we conclude that

$$-\beta_{\text{WALD}} \leq -\mathbb{E}[\bar{\varepsilon}^{\text{D}}] \quad \Leftrightarrow \quad \mathbb{C}\left[\frac{-B_{\text{Z}}^{\text{S}}}{\bar{\varepsilon}^{\text{S}} - \bar{\varepsilon}^{\text{D}}}, -\bar{\varepsilon}^{\text{D}}\right] \leq 0. \tag{59}$$

If demand and supply slope in the expected direction, then the monotonicity condition $P^{\text{E}}(1) \geq P^{\text{E}}(0)$ can hold if and only if $B_{\text{Z}}^{\text{S}} \leq 0$ (see (58)). If B_{Z}^{S} is constant or more generally independent of $(\bar{\varepsilon}^{\text{D}}, \bar{\varepsilon}^{\text{S}})$, then we can simplify (59) further to conclude that

$$-\beta_{\text{WALD}} \leq -\mathbb{E}[\bar{\varepsilon}^{\text{D}}] \quad \Leftrightarrow \quad \mathbb{C}\left[\frac{1}{\bar{\varepsilon}^{\text{S}} - \bar{\varepsilon}^{\text{D}}}, -\bar{\varepsilon}^{\text{D}}\right] \leq 0. \tag{60}$$

The covariance term in (60) compares a weakly increasing and weakly decreasing function of the non-negative random variable $-\bar{\varepsilon}^{\text{D}}$. If $\bar{\varepsilon}^{\text{S}}$ were constant or independent, then this term would necessarily be negative, leading to attenuation. This is effectively a nonparametric version of (53) and Proposition SA.3. In order for there not to be attenuation, the stochastic dependence between $\bar{\varepsilon}^{\text{D}}$ and $\bar{\varepsilon}^{\text{S}}$ needs to be sufficiently strong to flip the sign of

the covariance term.

SA.4 Incorporating Additional Nonlinear Parameters

In this section, we discuss how to compute the full and limited information approaches when the random coefficients specification (26) has an additional nonlinear parameter α. The leading example in which this flexibility is useful is to require some coefficients to be constant, in which case their constant values become part of α, with the corresponding marginals of $F_h(b;\alpha)$ being set to Dirac point masses characterized by α. This case has a considerable amount of structure that we show how to exploit for computational gains. It also leads to testable implications that shed some light on the random coefficients model.

SA.4.1 Profiled optimization

The optimization problem remains (28) with nonlinear parameters, but we now need to optimize over both π and α. This can be handled by profiling α into an outer optimization. The upper bound problem becomes a nested version of (30):

$$\tau_{\text{ub}}^{\star} = \max_{\alpha \in \mathcal{A}} \left[\max_{\pi \in \Pi(\alpha)} \pi' \bar{\tau}(\alpha) \text{ s.t. } \pi' \bar{g}(y,z;\alpha) = \mathbb{P}[Y \leq y | Z = z] \text{ for all } y, z \right], \tag{61}$$

where $\Pi(\alpha) \equiv \{\pi : (\alpha, \pi) \in \mathcal{A}\Pi\}$. We follow the usual convention here that the maximum of an infeasible program is negative infinity if α is such that the inner problem of (61) is infeasible. The inner problem of (61) is a linear program as long as $\Pi(\alpha)$ is polyhedral, which it will be when π are mixing weights and $\Pi(\alpha) = \Pi$ is the simplex. This means we can solve the inner problem to provable optimality quite quickly, leaving only the outer problem to contend with using less structured algorithms.

SA.4.2 Constant coefficients

Solving the outer problem of (61) can be challenging if α is high-dimensional. In this section, we show that if α is used to keep track of constant coefficients, then the structure of the random coefficients model means that it has effective dimension no larger than two. This makes it feasible to solve the outer problem with a grid search.

Separate $Z \equiv (Z_1, Z_2)$ into two subvectors, where the coefficients on Z_1 are constant in both the supply and the demand equations and the coefficients on Z_2 are random variables. Write the coefficient vectors as $B_{\text{Z}}^{\text{D}} \equiv (b_{\text{Z},1}^{\text{D}}, B_{\text{Z},2}^{\text{D}})$ and $B_{\text{Z}}^{\text{S}} \equiv (b_{\text{Z},1}^{\text{S}}, B_{\text{Z},2}^{\text{S}})$. Substituting this

notation into the reduced form equation (25) and taking expectation produces

$$\mathbb{E}[P|Z] = Z_1' \left(b_{\mathrm{Z},1}^{\mathrm{D}} - b_{\mathrm{Z},1}^{\mathrm{S}}\right) \mathbb{E}\left[\frac{1}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right] + Z_2' \, \mathbb{E}\left[\frac{B_{\mathrm{Z},2}^{\mathrm{D}} - B_{\mathrm{Z},2}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right],$$
$$\mathbb{E}[Q|Z] = Z_1' \left(b_{\mathrm{Z},1}^{\mathrm{D}} \, \mathbb{E}\left[\frac{B_{\mathrm{P}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right] + b_{\mathrm{Z},1}^{\mathrm{S}} \, \mathbb{E}\left[\frac{B_{\mathrm{P}}^{\mathrm{D}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right]\right) + Z_2' \, \mathbb{E}\left[\frac{B_{\mathrm{Z},2}^{\mathrm{D}} B_{\mathrm{P}}^{\mathrm{S}} + B_{\mathrm{Z},2}^{\mathrm{S}} B_{\mathrm{P}}^{\mathrm{D}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right].$$

The coefficients on Z_1 in linear regressions of P and Q onto Z are therefore

$$\rho_{\mathrm{P}} \equiv \beta_{\mathrm{P}} b_{\mathrm{Z},1}^{\mathrm{D}} - \beta_{\mathrm{P}} b_{\mathrm{Z},1}^{\mathrm{S}}, \qquad \text{where} \quad \beta_{\mathrm{P}} \equiv \mathbb{E}\left[\frac{1}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right],$$
$$\text{and} \quad \rho_{\mathrm{Q}} \equiv (1 - \beta_{\mathrm{Q}}) b_{\mathrm{Z},1}^{\mathrm{D}} + \beta_{\mathrm{Q}} b_{\mathrm{Z},1}^{\mathrm{S}}, \qquad \text{where} \quad \beta_{\mathrm{Q}} \equiv \mathbb{E}\left[\frac{B_{\mathrm{P}}^{\mathrm{D}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}}\right].$$

Solving this system of equations for $b_{\mathrm{Z},1}^{\mathrm{D}}$ and $b_{\mathrm{Z},1}^{\mathrm{S}}$ gives

$$b_{\mathrm{Z},1}^{\mathrm{D}} = \rho_{\mathrm{Q}} + \frac{\beta_{\mathrm{Q}}}{\beta_{\mathrm{P}}} \rho_{\mathrm{P}}, \quad \text{and} \quad b_{\mathrm{Z},1}^{\mathrm{S}} = \rho_{\mathrm{Q}} + \left(\frac{\beta_{\mathrm{Q}} - 1}{\beta_{\mathrm{P}}}\right) \rho_{\mathrm{P}}. \tag{62}$$

Equation (62) shows that the constant coefficient vectors $b_{\mathrm{Z},1}^{\mathrm{D}}$ and $b_{\mathrm{Z},1}^{\mathrm{S}}$ are fully determined by the unknown scalars β_{P} and β_{Q} together with the reduced form regression coefficients ρ_{P} and ρ_{Q}. Let $\alpha = (\beta_{\mathrm{P}}, \beta_{\mathrm{Q}}, b_{\mathrm{Z},1}^{\mathrm{D}}, b_{\mathrm{Z},1}^{\mathrm{S}})$. Then (62) shows that α is two-dimensional given knowledge of ρ_{P} and ρ_{Q}, regardless of the dimension of Z_1. This observation means that the grid search over $\mathcal{A}$ in the outer problem of (61) has dimension two. To exploit this dimension reduction, create a grid over $(\beta_{\mathrm{P}}, \beta_{\mathrm{Q}})$, fix a point on the grid, use (62) to solve for $b_{\mathrm{Z},1}^{\mathrm{D}}$ and $b_{\mathrm{Z},1}^{\mathrm{S}}$ from the (point identified) reduced form coefficients ρ_{Q} and ρ_{P}, then solve the inner problem of (61) while adding the deterministic constraints

$$\beta_{\mathrm{P}} = \sum_{h=1}^{d_\pi} \pi_h \int \left(\frac{1}{b_{\mathrm{P}}^{\mathrm{D}} + b_{\mathrm{P}}^{\mathrm{S}}}\right) dF_h(b; \alpha) \quad \text{and} \quad \beta_{\mathrm{Q}} = \sum_{h=1}^{d_\pi} \pi_h \int \left(\frac{b_{\mathrm{P}}^{\mathrm{D}}}{b_{\mathrm{P}}^{\mathrm{D}} + b_{\mathrm{P}}^{\mathrm{S}}}\right) dF_h(b; \alpha)$$

to the definition of $\Pi(\alpha)$. This strategy makes it computationally feasible to include several exogenous shifters or covariates in the full information approach, as long as they enter with constant coefficients.

Exclusion restrictions further reduce the dimension of α. Suppose that the coefficient on the first component of Z_1 in the demand equation is known to be zero ($b_{\mathrm{Z},1,1}^{\mathrm{D}} = 0$), so that this component is an excluded supply shifter. Then the first equation of (62) implies that

$$0 = \rho_{\mathrm{Q},1} + \frac{\beta_{\mathrm{Q}}}{\beta_{\mathrm{P}}} \rho_{\mathrm{P},1} \quad \text{or} \quad \beta_{\mathrm{P}} = -\frac{\rho_{\mathrm{P},1}}{\rho_{\mathrm{Q},1}} \beta_{\mathrm{Q}}, \tag{63}$$

so that β_{P} is determined by β_{Q} and the reduced form regression coefficients. The unknown scalar β_{Q} determines α, so $\mathcal{A}$ is effectively one-dimensional.

Having both an excluded demand and supply shifter makes α point identified. Suppose that the first coefficient on Z_1 in the demand equation is zero, $b^{\mathrm{D}}_{\mathrm{Z},1,1} = 0$, and that the coefficient on the second component of Z_1 is zero in the supply equation, $b^{\mathrm{S}}_{\mathrm{Z},1,2} = 0$. Then (63) still holds, but the second equation of (62) additionally implies that

$$0 = \rho_{\mathrm{Q},2} + \left(\frac{\beta_{\mathrm{Q}} - 1}{\beta_{\mathrm{P}}}\right)\rho_{\mathrm{P},2}, \tag{64}$$

where $\rho_{\mathrm{Q},2}$ and $\rho_{\mathrm{P},2}$ are the second components of the vectors of reduced form coefficients. Combining (63) and (64) shows that β_{P} and β_{Q} are point identified from the reduced form regression coefficients:

$$\beta_{\mathrm{Q}} = \left(1 - \frac{\rho_{\mathrm{P},1}\rho_{\mathrm{Q},2}}{\rho_{\mathrm{Q},1}\rho_{\mathrm{P},2}}\right)^{-1} \quad \text{and} \quad \beta_{\mathrm{P}} = -\frac{\rho_{\mathrm{P},1}}{\rho_{\mathrm{Q},1}}\left(1 - \frac{\rho_{\mathrm{P},1}\rho_{\mathrm{Q},2}}{\rho_{\mathrm{Q},1}\rho_{\mathrm{P},2}}\right)^{-1}.$$

The constant coefficients are then point identified through (62). In this case, the outer problem of (61) disappears.

SA.4.3 Overidentification with constant coefficients

Suppose that both the first and second demand coefficients of Z_1 are known to be zero, so that there are two excluded supply shifters, both with constant supply coefficients. Then (63) applied to both of these components produces the testable implication that $\rho_{\mathrm{Q},1}/\rho_{\mathrm{P},1} = \rho_{\mathrm{Q},2}/\rho_{\mathrm{P},2}$.

This testable implication represents a partial preservation of the familiar overidentification test from the constant coefficients case to the random coefficients case. The difference is that the overidentified quantity is not the slope on price in the demand equation—which is now random—but rather the ratio $\beta_{\mathrm{Q}}/\beta_{\mathrm{P}}$, which is a weighted average of the slope of price in the demand equation. The implication requires the corresponding coefficients of the supply equation to be non-random, which is a restrictive assumption. The implication can be turned into a test using the usual types of overidentification tests, as long as all excluded instruments are supply shifters with constant supply coefficients. Otherwise, one could use the minimum distance interpretation of overidentified linear IV estimators (see Windmeijer, 2019) to construct a test that compares only those coefficients that satisfy these properties.

SA.5 Estimation and Statistical Inference

In this section, we discuss methods for estimation and inference using a finite sample $\{P_i, Q_i, Z_i\}_{i=1}^n$ of n markets. As in Section SA.4, we allow the basis functions $F_h(\cdot;\alpha)$ to be indexed by a nonlinear parameter α, which could correspond to constant coefficients, but could also incorporate some other parameterization.

SA.5.1 Criterion function for the full information approach

In this section, we derive a scalar criterion function $c(F)$ that measures the extent to which the distribution of $Y \equiv (P, Q)$ implied by a given distribution of random coefficients, F, matches the empirical distribution of Y. Let

$$\upsilon(y; F) \equiv \mathbb{E}\left[\left(\mathbb{1}[Y \leq y] - \mathbb{P}_F[Y \leq y|Z]\right)^2\right], \tag{65}$$

where the outer expectation is taken over the joint distribution of (Y, Z) for a fixed y. Then

$$\upsilon(y; F) \geq \mathbb{E}\left[\left(\mathbb{1}[Y \leq y] - \mathbb{P}[Y \leq y|Z]\right)^2\right] \text{ for any } F, \tag{66}$$

by standard least squares arguments, with equality obtained if and only if $\mathbb{P}_F[Y \leq y|Z = z] = \mathbb{P}[Y \leq y|Z = z]$ for almost every z. The population criterion function aggregates over all y:

$$c(F) \equiv \int \upsilon(y; F)\, dG(y),$$

where G is the unconditional population distribution of Y. Then F minimizes c if and only if $\mathbb{P}_F[Y \leq y|Z = z] = \mathbb{P}[Y \leq y|Z = z]$ for every y and almost every z. The criterion function therefore provides an alternative characterization of the identified set for F:

$$\mathcal{F}^\star = \arg\min_{F \in \mathcal{F}^\dagger} c(F).$$

For computation, we continue to restrict the set of admissible distributions $\mathcal{F}^\dagger$ to have the form (26), so that F is parameterized by (α, π), and the criterion function can be written as

$$c(\alpha, \pi) = \int \mathbb{E}\left[\left(\mathbb{1}[Y \leq y] - \pi' \bar{g}(y, Z; \alpha)\right)^2\right] dG(y),$$

where π and $\bar{g}(y, z; \alpha)$ are d_π-dimensional vectors containing π_h and $\bar{g}_h(p, q, z; \alpha)$ for $h = 1, \ldots, d_\pi$. We estimate c from $\{P_i, Q_i, Z_i\}_{i=1}^n$ using sample analogs that replace G by the empirical distribution of Y and the inner expectation in the definition of c by the joint distribution of (Y, Z):

$$c_n(\alpha, \pi) \equiv \frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n \left(\mathbb{1}[Y_i \leq Y_j] - \pi' \bar{g}(Y_j, Z_i; \alpha)\right)^2. \tag{67}$$

SA.5.2 Criterion function for the limited information approach

For the limited information approach, we build a criterion function based on matching the reduced form coefficients $\rho_{j,k}$, as defined in (34), across multiple values of j and k.

Collect $\rho_{j,k}$ across choices of j and k into a vector ρ. Then (34) imposed for all of these j and k can be rewritten as the system of linear equations

$$L(\alpha)\pi = \rho, \tag{68}$$

where $L(\alpha)$ is a matrix whose entries are given by

$$[L(\alpha)]_{(j,k),h} = (-1)^j \int \left(\frac{(b^{\mathrm{S}}_{\mathrm{Z},1})^{j+k}(b^{\mathrm{D}}_{\mathrm{P}})^k}{(b^{\mathrm{D}}_{\mathrm{P}} + b^{\mathrm{S}}_{\mathrm{P}})^{j+k}} \right) d\tilde{F}_h(b;\alpha),$$

with (j,k) indexing the rows and h indexing the columns. We've also added α to the notation, in case a nonlinear parameter is desired. We use the population criterion function

$$c(\alpha,\pi) \equiv \|L(\alpha)\pi - \rho\|^2 .$$

Let $\rho_{j,k,n}$ denote the OLS estimator of $\rho_{j,k}$, with ρ_n the vector formed from these estimators. Then the sample criterion function is

$$c_n(\alpha,\pi) \equiv \|L(\alpha)\pi - \rho_n\|^2. \tag{69}$$

In our application in Section 5, we use a weighted version of this criterion:

$$c_n(\alpha,\pi) \equiv \| \operatorname{diag}(\Sigma_n)^{-1} (L(\alpha)\pi - \rho_n) \|^2, \tag{70}$$

where Σ_n is an estimator of the asymptotic variance of ρ_n.

SA.5.3 Estimation

For either the full or limited information approaches, we use the criterion function to construct estimators of $\tau^\star_{\mathrm{lb}}$ and $\tau^\star_{\mathrm{ub}}$ while accounting for partial identification, following a procedure similar to that in Mogstad et al. (2018). In the first step of the procedure, we determine the best possible fit to the data by solving for

$$c^\star_n \equiv \min_{(\alpha,\pi)\in\mathcal{A}\Pi} c_n(\alpha,\pi). \tag{71}$$

We solve this problem by profiling α as in (61); the inner problem is now a quadratic program in either the full or limited information approach, as can be seen from both (67) and (69). In the second step, we construct estimators of the sharp bounds on the identified set $\mathcal{T}^\star$. The upper bound estimator is

$$\tau_{\mathrm{ub},n} \equiv \max_{(\alpha,\pi)\in\mathcal{A}\Pi} \pi'\bar{\tau}_n(\alpha) - \kappa_n \left(c_n(\alpha,\pi) - c^\star_n\right), \tag{72}$$

where κ_n is a non-negative tuning parameter, and $\bar{\tau}_n(\alpha)$ is a vector of estimators of $\bar{\tau}_h(\alpha)$, only needed if τ depends on Z:

$$\bar{\tau}_{h,n}(\alpha) \equiv \frac{1}{n}\sum_{i=1}^{n}\int \tau(Z_i, b)\, dF_h(b;\alpha). \tag{73}$$

The lower bound estimator is the optimal value of a minimization counterpart to (72) that penalizes in the opposite direction:

$$\tau_{\text{lb},n} \equiv \min_{(\alpha,\pi)\in\mathcal{A}\Pi} \pi'\bar{\tau}_n(\alpha) + \kappa_n\left(c_n(\alpha,\pi) - c_n^\star\right). \tag{74}$$

Both $\tau_{\text{lb},n}$ and $\tau_{\text{ub},n}$ always exist, with $\tau_{\text{lb},n} \leq \tau_{\text{ub},n}$, so that the estimated bounds are always non-empty.

Intuitively, the upper bound estimator $\tau_{\text{ub},n}$ tries to find the largest value of the estimated target parameter $\pi'\bar{\tau}_n(\alpha)$ that can be achieved by a parameter value that "almost" matches the data as well as possible. The notion of "almost" matching the data is controlled by κ_n, which penalizes deviations from the best-matching criterion value, $c_n^\star$. Mogstad et al. (2018, Sections 3.2 and S3) imposed these penalties by requiring (α,π) to satisfy the constraint

$$c_n(\alpha,\pi) \leq c_n^\star(1+\lambda_n) \tag{75}$$

for a non-negative tuning parameter λ_n that shrinks to zero with n. Our approach here simply focuses on the Lagrangian form, with the shrinking slackness parameter λ_n being replaced by the expanding Lagrangian penalty κ_n. We do this because $c_n(\alpha,\pi)$ is a quadratic function of π and adding a quadratic penalty is less burdensome computationally than imposing a quadratic constraint.

SA.5.4 Choosing the tuning parameter

Using a finite penalty (instead of $\kappa_n = +\infty$) is important to ensure consistency because it smooths out potential pathologies that could arise when considering the convergence of sequences of sets (see, for example, Molinari, 2020, Section 4.2.1). In practice, we use data-driven Monte Carlo simulations to choose κ_n so that the resulting bound estimates perform well in a DGP fit to the empirical distribution.

In the first step of this procedure, we find an optimizer $(\alpha_n^\star, \pi_n^\star)$ of the minimum criterion problem (71). We augment the criterion of this problem with a small L_2 penalty for π, to help ensure that $\pi_n^\star$ is away from the boundary of the simplex. Let $F_n^\star$ denote the distribution of B produced by $(\alpha_n^\star, \pi_n^\star)$:

$$F_n^\star(b) \equiv \sum_{h=1}^{d_\pi} \pi_{nh}^\star F_h(b;\alpha_n^\star). \tag{76}$$

Then let

$$\tau^{\star}_{\text{ub},n} \equiv \max_{\alpha\in\mathcal{A}} \left[\max_{\pi\in\Pi(\alpha)} \pi'\bar{\tau}(\alpha) \text{ s.t. } \pi'\bar{g}(y,z;\alpha) = \mathbb{P}_{F^{\star}_n}[Y \leq y | Z = z] \text{ for all } y, z \right], \tag{77}$$

denote the pseudo-population upper bound, where "pseudo" refers to the fact that the distribution of observables being matched here is that which would be produced by $F^{\star}_n$, rather than that of the true (but unknown) population distribution. Let $\tau^{\star}_{\text{lb},n}$ be the analogous pseudo-population lower bound. Alternatively, when using the limited information criterion, we let $\rho^{\star}_n$ denote the vector of reduced form moments implied by $F^{\star}_n$ and then define

$$\tau^{\star}_{\text{ub},n} \equiv \max_{\alpha\in\mathcal{A}} \left[\max_{\pi\in\Pi(\alpha)} \pi'\bar{\tau}(\alpha) \text{ s.t. } L(\alpha)\pi = \rho^{\star}_n \right],$$

and similarly for the lower bound.

In the second step, we use $F^{\star}_n$ to conduct a Monte Carlo simulation with m samples of size n. The exogenous variables, Z_i, are drawn from the empirical distribution. The random coefficients B_i are drawn from $F^{\star}_n$ and then combined with Z_i to produce a draw of (Y_i, Z_i). Let $\tau^{\ell}_{\text{lb},n}(\kappa)$ and $\tau^{\ell}_{\text{ub},n}(\kappa)$ denote estimators that use the ℓth sample and tuning parameter value κ, where the dependence on κ is now made explicit in the notation. We find the κ that minimizes the average root mean-squared error of $\tau^{\ell}_{\text{lb},n}(\kappa)$ and $\tau^{\ell}_{\text{ub},n}(\kappa)$ as estimators for the pseudo-population bounds $\tau^{\star}_{\text{lb},n}$ and $\tau^{\star}_{\text{ub},n}$:

$$\kappa^{\star}_n \equiv \arg\min_{\kappa\geq 0} \left(\frac{1}{m}\sum_{\ell=1}^{m} \left(\tau^{\ell}_{\text{lb},n}(\kappa) - \tau^{\star}_{\text{lb},n} \right)^2 \right)^{1/2} + \left(\frac{1}{m}\sum_{\ell=1}^{m} \left(\tau^{\ell}_{\text{ub},n}(\kappa) - \tau^{\star}_{\text{ub},n} \right)^2 \right)^{1/2}.$$

We take $\kappa^{\star}_n$ to be the tuning parameter choice and use it to construct $\tau_{\text{lb},n}$ and $\tau_{\text{ub},n}$ with data from the original sample by solving (74) and (72).

SA.5.5 Statistical inference

In this section, we describe two methods of conducting inference for the limited information approach, which is what we use in the sales tax application in Section 5. As in that application, we assume that there is no nonlinear parameter, α.

One method is to apply the testing procedure developed in Fang et al. (2023) for systems of linear equations with known coefficients. Augment the linear system of equations (68) with an additional linear equation reflecting the target parameter:

$$L\pi = \rho \quad \text{and} \quad \bar{\tau}'\pi = t, \tag{78}$$

where t is some hypothesized value of the target parameter. Note that L is a deterministic matrix of coefficients, so if $\bar{\tau}$ is known, then testing the null hypothesis that there exists

a $\pi \in \Pi$ satisfying (78) is exactly the problem analyzed by Fang et al. (2023). Confidence intervals can be constructed by inverting a series of tests for different t. All of the target parameters we consider in Section 5, such as $\mathbb{E}[B_{\mathrm{P}}^{\mathrm{D}}]$, are such that $\bar{\tau}$ is known.

A simpler method is to first construct a confidence region for ρ and then optimize π over this confidence region. Let $\mathcal{R}$ denote a confidence region for ρ. Then the lower bound of the confidence interval is the optimal value of the program

$$\min_{\pi \in \Pi} \ \pi' \bar{\tau} \quad \text{s.t.} \quad L\pi \in \mathcal{R}, \tag{79}$$

while the upper bound is the optimal value of the corresponding maximization problem. We construct $\mathcal{R}$ from the usual Euclidean norm using a robust asymptotic variance estimator Σ_n for ρ_n, viewed as a system of OLS estimators:

$$\mathcal{R} \equiv \left\{ \rho : (\rho_n - \rho)' \Sigma_n^{-1} (\rho_n - \rho) \leq \chi^2(1-a) \right\},$$

where $\chi^2(1-a)$ is the $1-a$ quantile of a chi-squared distribution with degrees of freedom equal to the dimension of ρ. This confidence region makes (79) and its maximization counterpart a (convex) second-order cone program.

SA.6 Algebra for the Limited Information Approach

In this appendix, we derive the regression specification for $P^j Q^k$ given in (33). We start by rewriting the reduced form using more compact notation:

$$p^{\mathrm{E}}(Z,B) = Z' \left(\frac{B_{\mathrm{Z}}^{\mathrm{D}} - B_{\mathrm{Z}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}} \right) \equiv Z' A_{\mathrm{P}} \quad \text{and} \quad q^{\mathrm{E}}(Z,B) = Z' \left(\frac{B_{\mathrm{P}}^{\mathrm{S}} B_{\mathrm{Z}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{D}} B_{\mathrm{Z}}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}} \right) \equiv Z' A_{\mathrm{Q}}.$$

Using the multinomial theorem, we have

$$P^j \equiv \sum_{\boldsymbol{j}:|\boldsymbol{j}|=j} \binom{j}{\boldsymbol{j}} A_{\mathrm{P}}^{\boldsymbol{j}} Z^{\boldsymbol{j}} \quad \text{and} \quad Q^k \equiv \sum_{\boldsymbol{k}:|\boldsymbol{k}|=k} \binom{k}{\boldsymbol{k}} A_{\mathrm{Q}}^{\boldsymbol{k}} Z^{\boldsymbol{k}},$$

where $\boldsymbol{j} \equiv (j_1, \ldots, j_{d_z})$ and $\boldsymbol{k} \equiv (k_1, \ldots, k_{d_z})$ are multi-indices with the usual notation:

$$|\boldsymbol{j}| \equiv \sum_{i=1}^{d_z} j_i, \quad \binom{j}{\boldsymbol{j}} \equiv \frac{j!}{j_1! \cdots j_{d_z}!} \equiv \frac{j!}{\boldsymbol{j}!}, \quad \text{and} \quad Z^{\boldsymbol{j}} \equiv \prod_{i=1}^{d_z} Z_i^{j_i}.$$

Multiplying these two monomial expansions and applying a change of variables $\boldsymbol{\ell} = \boldsymbol{j} + \boldsymbol{k}$ gives

$$\begin{aligned} P^j Q^k &= \sum_{|\boldsymbol{j}|=j} \sum_{|\boldsymbol{k}|=k} \frac{j!k!}{\boldsymbol{j}!\boldsymbol{k}!} A_{\mathrm{P}}^{\boldsymbol{j}} A_{\mathrm{Q}}^{\boldsymbol{k}} Z^{\boldsymbol{j}+\boldsymbol{k}} \\ &= \sum_{|\boldsymbol{\ell}|=j+k} Z^{\boldsymbol{\ell}} \left(\sum_{\substack{|\boldsymbol{j}|=j \\ \boldsymbol{j}+\boldsymbol{k}=\boldsymbol{\ell}}} \frac{j!}{\boldsymbol{j}!} \frac{k!}{\boldsymbol{k}!} A_{\mathrm{P}}^{\boldsymbol{j}} A_{\mathrm{Q}}^{\boldsymbol{k}} \right) \equiv \sum_{|\boldsymbol{\ell}|=j+k} Z^{\boldsymbol{\ell}} A_{j,k}(\boldsymbol{\ell}). \end{aligned}$$

Substituting the definitions of A_{P} and A_{Q}, we get that

$$A_{j,k}(\boldsymbol{\ell}) = \sum_{\substack{|\boldsymbol{j}|=j \\ \boldsymbol{j}+\boldsymbol{k}=\boldsymbol{\ell}}} \frac{j!}{\boldsymbol{j}!} \frac{k!}{\boldsymbol{k}!} \prod_{i=1}^{d_z} \left(\frac{B_{\mathrm{Z},i}^{\mathrm{D}} - B_{\mathrm{Z},i}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}} \right)^{j_i} \left(\frac{B_{\mathrm{P}}^{\mathrm{S}} B_{\mathrm{Z},i}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{D}} B_{\mathrm{Z},i}^{\mathrm{S}}}{B_{\mathrm{P}}^{\mathrm{D}} + B_{\mathrm{P}}^{\mathrm{S}}} \right)^{k_i},$$

which is the expression given in (33).

SA.7 Microfounding a Linear Random Coefficients Model

In this section, we provide models of consumer and firm behavior that produce a market-level demand system that has random coefficients.

SA.7.1 Demand

Consider a set $\mathcal{J}$ of consumers indexed by j in a single market. Each consumer has preferences over quantities of a numeraire $q_{0,j}$ and a focal good q_j given by

$$U_j\left(q_{0,j}, q_j\right) = \begin{cases} q_{0,j} + \xi_j q_j^{\chi} - \gamma_j & \text{if } q_j > 0 \\ q_{0,j} & \text{if } q_j = 0, \end{cases} \tag{80}$$

where $0 < \chi < 1$ measures the concavity of the sub-utility over the focal good, $\xi_j > 0$ is the relative weight of the focal good, and γ_j is a disutility shock from consuming any positive amount of the product, for example due to a fixed transaction cost. Let $N(\xi)$ denote the mass of consumers of type $\xi_j = \xi$, and suppose that γ_j is distributed independently of ξ_j according to the Pareto distribution:

$$F(\gamma) = \begin{cases} \left(\frac{\gamma}{\overline{\gamma}}\right)^{\psi} & \text{if } \gamma \leq \overline{\gamma} \\ 0 & \text{if } \gamma > \overline{\gamma}. \end{cases}$$

Each consumer chooses q_j to maximize utility subject to the budget constraint $q_{0,j} + pq_j = y_j$ where y_j is individual income, p is the price of the focal good, and the price of the numeraire is normalized to one. Substituting the budget constraint into preferences (80),

we obtain the following utility maximization problem for the consumer:

$$\max_{q_j} \begin{cases} y_j + \xi_j q_j^{\chi} - p q_j - \gamma_j & \text{if } q_j > 0 \\ y_j & \text{if } q_j = 0 \end{cases}.$$

The fixed transaction cost γ_j creates the possibility of corner solutions where some consumers prefer not to consume any amount of the focal product (e.g., Dubé, 2019). Suppose that the consumer is at an interior solution. Solving the first-order conditions at an interior solution yields individual-level demand

$$q_j = \left(\frac{p}{\chi \xi_j}\right)^{-\frac{1}{1-\chi}}.$$

The associated indirect utility conditional on consuming a positive amount is

$$y_j + \xi_j^{\frac{1}{1-\chi}} \chi^{\frac{\chi}{1-\chi}} (1-\chi)\, p^{-\frac{\chi}{1-\chi}} - \gamma_j.$$

This indirect utility is greater than the utility from purchasing none of the focal good if and only if

$$\xi_j^{\frac{1}{1-\chi}} \chi^{\frac{\chi}{1-\chi}} (1-\chi)\, p^{-\frac{\chi}{1-\chi}} \geq \gamma_j. \tag{81}$$

The share of consumers of type ξ that consume a positive amount of the good is given by the proportion for which (81) is true:

$$F\left(\xi^{\frac{1}{1-\chi}} \chi^{\frac{\chi}{1-\chi}} (1-\chi)\, p^{-\frac{\chi}{1-\chi}}\right) = \left(\frac{\xi^{\frac{1}{1-\chi}} \chi^{\frac{\chi}{1-\chi}} (1-\chi)\, p^{-\frac{\chi}{1-\chi}}}{\overline{\gamma}}\right)^{\psi},$$

where F is the distribution of γ_j, which is assumed to be independent of ξ.

The aggregate market demand for the focal good is then determined by the combination of the intensive and extensive margins of demand integrated over ξ:

$$\begin{aligned} Q^{\mathrm{D}}(p) &\equiv \int \left(\frac{p}{\chi \xi}\right)^{-\frac{1}{1-\chi}} F\left(\xi^{\frac{1}{1-\chi}} \chi^{\frac{\chi}{1-\chi}} (1-\chi)\, p^{-\frac{\chi}{1-\chi}}\right) N(\xi)\, d\xi \\ &= \left[\chi^{\frac{1+\psi\chi}{1-\chi}} \left(\frac{1-\chi}{\overline{\gamma}}\right)^{\psi} \int \xi^{\frac{1+\psi}{1-\chi}} N(\xi)\, d\xi\right] p^{-\frac{1+\psi\chi}{1-\chi}}. \end{aligned}$$

Taking logs yields the following expression for market-level demand:

$$\begin{aligned} &\log(Q^{\mathrm{D}}(p)) = B_1^{\mathrm{D}} - B_{\mathrm{P}}^{\mathrm{D}} \log(p) \\ &\text{where} \quad B_1^{\mathrm{D}} \equiv \log\left(\chi^{\frac{1+\psi\chi}{1-\chi}} \left(\frac{1-\chi}{\overline{\gamma}}\right)^{\psi} \int \xi^{\frac{1+\psi}{1-\chi}} N(\xi)\, d\xi\right) \quad \text{and} \quad B_{\mathrm{P}}^{\mathrm{D}} \equiv \frac{1+\psi\chi}{1-\chi}. \end{aligned} \tag{82}$$

Equation (82) has the form (RC) with $q^{\mathrm{D}}(p, B, Z) \equiv \log(Q^{\mathrm{D}}(p))$ being linear in $\log(p)$.

The slope coefficient in (82) depends on utility parameters χ and ψ that characterize preferences. While consumer theory provides a characterization of how individual consumers with these preferences will behave, it provides no restriction on how these preference parameters vary across markets composed of different consumers. The assumption that $B_{\mathrm{P}}^{\mathrm{D}}$ is constant would effectively require imposing this type of homogeneity restriction.

SA.7.2 Supply

Suppose that each market is characterized by a representative firm that produces output of the focal good in the consumer demand model of the previous section using labor ℓ and capital k through a Cobb-Douglas production function:

$$Ak^{\mathfrak{a}\mathfrak{g}}\ell^{(1-\mathfrak{a})\mathfrak{g}},$$

where A is total factor productivity, $0 < \mathfrak{a} < 1$ is capital intensity, and $0 < \mathfrak{g} \leq 1$ measures returns to scale. Total cost is given by

$$c\left(k,\ell\right) \equiv w_k^0 k^{1+\mathfrak{b}_k} + w_\ell^0 \ell^{1+\mathfrak{b}_\ell},$$

where $\mathfrak{b}_k, \mathfrak{b}_\ell \geq 0$ allow input costs to increase with the use of each input. Suppose that the firm takes both input and output prices as given and chooses inputs to maximize profit.

The representative firm's cost of producing q of the focal good is given by

$$c(q) \equiv \min_{k,\ell} \; w_k^0 k^{1+\mathfrak{b}_k} + w_\ell^0 \ell^{1+\mathfrak{b}_\ell} \quad \text{s.t.} \quad q = Ak^{\mathfrak{a}\mathfrak{g}}\ell^{(1-\mathfrak{a})\mathfrak{g}}.$$

The first-order conditions are

$$\begin{aligned} (1+\mathfrak{b}_k) w_k^0 k^{\mathfrak{b}_k} &= \nu \mathfrak{a}\mathfrak{g} A k^{\mathfrak{a}\mathfrak{g}-1}\ell^{(1-\mathfrak{a})\mathfrak{g}} \\ \text{and} \quad (1+\mathfrak{b}_\ell) w_\ell^0 \ell^{\mathfrak{b}_\ell} &= \nu (1-\mathfrak{a})\mathfrak{g} A k^{\mathfrak{a}\mathfrak{g}}\ell^{(1-\mathfrak{a})\mathfrak{g}-1}, \end{aligned}$$

where ν is the Lagrange multiplier for the output constraint. Taking the ratio of first-order conditions and rearranging, we obtain:

$$k = \left[\frac{\mathfrak{a}}{1-\mathfrak{a}} \frac{(1+\mathfrak{b}_\ell) w_\ell^0}{(1+\mathfrak{b}_k) w_k^0}\right]^{\frac{1}{1+\mathfrak{b}_k}} \ell^{\frac{1+\mathfrak{b}_\ell}{1+\mathfrak{b}_k}}. \tag{83}$$

Substituting (83) into the production function yields labor demand conditional on q:

$$\ell = \left[\frac{1-\mathfrak{a}}{\mathfrak{a}} \frac{(1+\mathfrak{b}_k) w_k^0}{(1+\mathfrak{b}_\ell) w_\ell^0}\right]^{\frac{\mathfrak{a}}{\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)}} \left(\frac{q}{A}\right)^{\frac{1+\mathfrak{b}_k}{\mathfrak{g}[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)]}}. \tag{84}$$

Substituting (84) into (83) and rearranging yields capital demand conditional on q:

$$k = \left[\frac{\mathfrak{a}}{1-\mathfrak{a}}\frac{(1+\mathfrak{b}_\ell)w_\ell^0}{(1+\mathfrak{b}_k)w_k^0}\right]^{\frac{1-\mathfrak{a}}{\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)}}\left(\frac{q}{A}\right)^{\frac{1+\mathfrak{b}_\ell}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]}}. \tag{85}$$

These factor demands imply that the cost function for producing q is

$$c(q) = \tilde{c}\times\left(\frac{q}{A}\right)^{\frac{(1+\mathfrak{b}_k)(1+\mathfrak{b}_\ell)}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]}}$$

where

$$\begin{aligned}\tilde{c} \equiv{}& \left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]\\ &\times\left(\frac{w_k^0}{\mathfrak{a}(1+\mathfrak{b}_\ell)}\right)^{\frac{\mathfrak{a}(1+\mathfrak{b}_\ell)}{\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)}}\\ &\times\left(\frac{w_\ell^0}{(1-\mathfrak{a})(1+\mathfrak{b}_k)}\right)^{\frac{(1-\mathfrak{a})(1+\mathfrak{b}_k)}{\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)}}.\end{aligned}$$

To maximize profit taking output price p as given, the firm chooses quantity $Q^{\mathrm{S}}(p)$ to solve

$$Q^{\mathrm{S}}(p) = \arg\max_q pq - c(q).$$

The first-order condition is:

$$p = \frac{(1+\mathfrak{b}_k)(1+\mathfrak{b}_\ell)}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]}\tilde{c}A^{-1}\left(\frac{Q^{\mathrm{S}}(p)}{A}\right)^{\left(\frac{(1+\mathfrak{b}_k)(1+\mathfrak{b}_\ell)}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]}-1\right)}.$$

Rearranging and taking logs yields

$$\begin{aligned}&\log(Q^{\mathrm{S}}(p)) = B_1^{\mathrm{S}} + B_{\mathrm{P}}^{\mathrm{S}}\log(p)\\ &\text{where}\quad B_1^{\mathrm{S}} \equiv \log A + \frac{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]}{(1+\mathfrak{b}_k)(1+\mathfrak{b}_\ell)-\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]}\\ &\qquad\qquad\times\log\left[\frac{A\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]}{(1+\mathfrak{b}_k)(1+\mathfrak{b}_\ell)\tilde{c}}\right]\\ &\text{and}\quad B_{\mathrm{P}}^{\mathrm{S}} \equiv \frac{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]}{(1+\mathfrak{b}_k)(1+\mathfrak{b}_\ell)-\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_\ell)+(1-\mathfrak{a})(1+\mathfrak{b}_k)\right]}.\end{aligned} \tag{86}$$

Equation (86) has the form (RC) with $q^{\mathrm{S}}(p,B,Z) \equiv \log(Q^{\mathrm{S}}(p))$ being linear in $\log(p)$.

The slope coefficient in (86) depends on parameters that characterize both the production function and the input market. If the structure of the production function or input market differs across product markets, then the supply equation will have random coefficients. Differences in the structure of the input market elasticities in particular could

Table SA.3: Estimation sample and reduced form estimates

Panel A. Estimation sample	
Sample window	2008–2014
Modules	249
Counties	2,111
Biennial periods	10
Observations	3,994,604

Panel B. Reduced form estimates

Outcome	Estimate	Std. error
Consumer price (P)	0.90	0.01
Before-tax price ($P - Z$)	-0.10	0.01
Quantity (Q)	-0.56	0.05

Notes: See Section SA.8.1 for definitions and details.

be caused by differences in monopsony power or taxes.

SA.8 Application Details

This section provides more details on the data construction and empirical specification for the sales tax application in Section 5.

SA.8.1 Data construction

We provide more details on how we construct the data used in Section 5; see Gaarder and Henry de Frahan (2025) for a more extensive discussion.

The NielsenIQ scanner data provides store-level prices and quantities for a large set of universal product codes (UPCs). The Thomson Reuters OneSource Sales Tax database provides statutory tax rates and product-level taxability status. We merge these two datasets in the same way as described in Gaarder and Henry de Frahan (2025).

To construct the estimation sample, we begin by aggregating individual UPCs to NielsenIQ's product modules. We restrict attention to modules in the top quartile of the 2008 sales distribution. Next, we balance the sample at the store-by-product level, keeping only store-product pairs that are observed throughout the entire sample window of January 2008 to December 2014. We aggregate the balanced store-level data to the module-by-county level. We define a market to be a county-module pair.

We construct price and quantity indices for each market in each of the 14 semiannual periods following the same approach as Gaarder and Henry de Frahan (2025). The indices are constructed following a regression procedure proposed by Handbury and Weinstein (2015). Gaarder and Henry de Frahan (2025) show that using chained Laspeyres indices produces similar measures of price, both inclusive of sales tax (consumer prices) and net of

Table SA.4: Reduced form moments used for limited information estimation

Order	j	k	$\rho_{j,k,n}$	SE
Panel A. Baseline moments $(j+k \leq 1)$				
1	1	0	0.90	0.01
1	0	1	-0.56	0.05
Panel B. Additional moments $(j+k \leq 2)$				
2	2	0	0.78	0.06
2	1	1	-0.26	0.16

Notes: The coefficients $\rho_{j,k,n}$ are estimates of $\rho_{j,k}$, the reduced form moments matched by the limited information estimator. Panel A reports the first-order moments used in the baseline specification. Panel B reports the additional higher-order moments used when the information set is expanded to $(j+k \leq 2)$.

sales tax (before-tax prices). Then we use these 14 time periods to define rolling changes in log quantity and log prices for 10 two-year periods, in order to capture longer-run responses to tax changes. Finally, we residualize the rolling two-year changes in prices and quantity by regressing them on module-by-census region-by-time period fixed effects and retaining the residuals as P and Q. We define the instrument Z as being the two-year change in $\log(1+r)$, residualized in the same way, where r is the ad valorem tax rate.

Panel A of Table SA.3 shows the dimensions of the resulting estimation sample. Panel B shows reduced form estimates from regressions of P, Q, and $P - Z$ onto Z.

SA.8.2 Basis specification

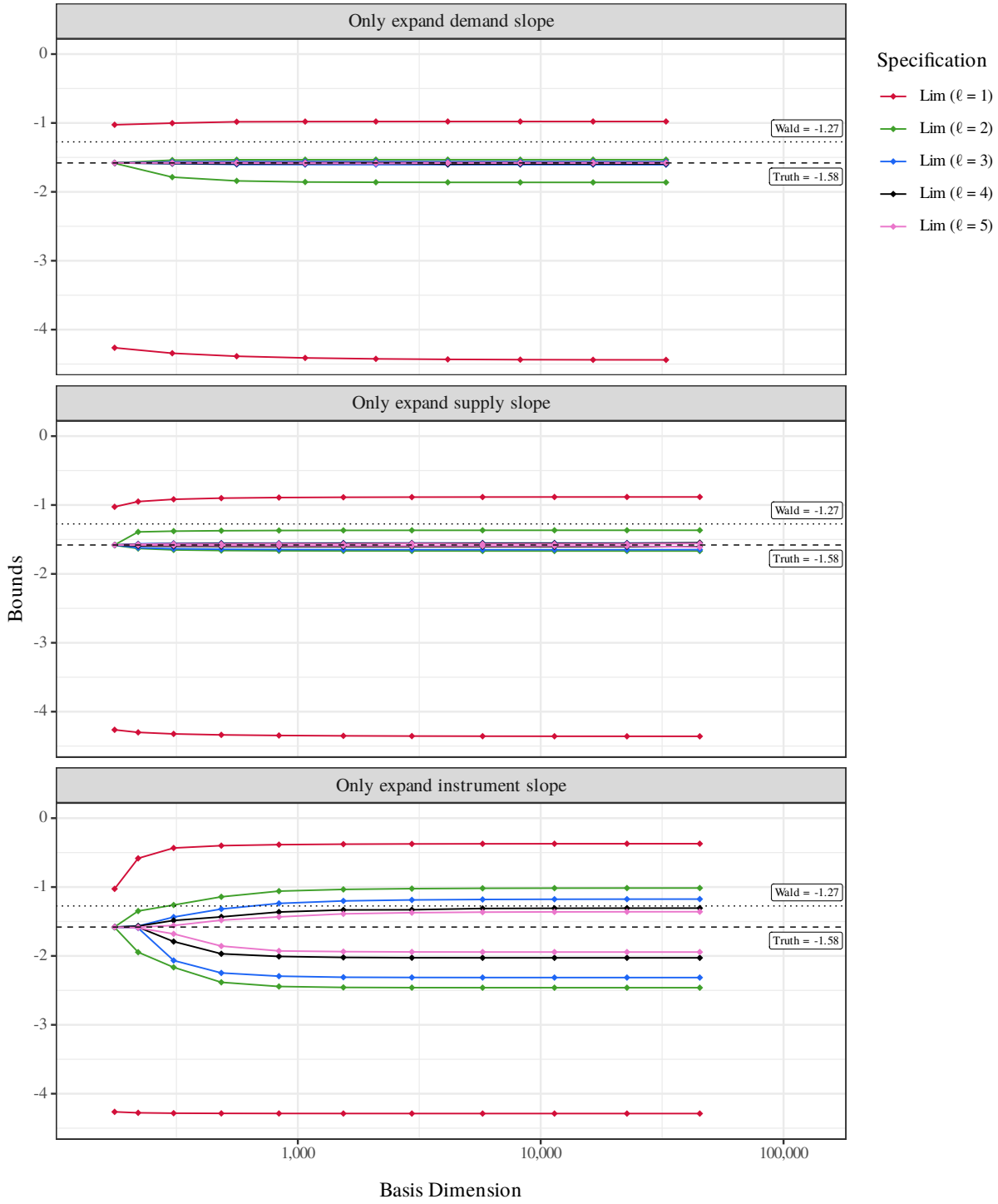
We model the distribution of $\tilde{B} \equiv (B_{\text{P}}^{\text{D}}, B_{\text{P}}^{\text{S}}, B_{\text{Z},1}^{\text{S}})$ with a tensor product basis of univariate cubic B-splines. Each component is restricted to be non-negative and normalized to integrate to one, so that the resulting convex mixture is always a density (see, for example, Matsuda and Iwafuchi, 2025). We specify each univariate basis with an open uniform knot vector: the first and last four knots are repeated, and the interior knots are equally spaced. We use seven knots each for B_{P}^{D} and B_{P}^{S}, and twelve knots for $B_{\text{Z},1}^{\text{S}}$. The tensor product basis has 1,134 terms. Imposing the supply-side Ramsey Exclusion Restriction $B_{\text{Z},1}^{\text{S}} = -B_{\text{P}}^{\text{S}}$ reduces the basis to 81 terms.

SA.8.3 Reduced form moments

Table SA.4 reports the estimates of $\rho_{j,k}$ used in the application. For $j+k \leq 1$, we use both moments $(j,k) = (1,0)$ and $(j,k) = (0,1)$. These moments also appear in Table SA.3; they are the usual reduced form moments used in a constant coefficients approach. For $j+k \leq 2$, we additionally include $(j,k) = (2,0)$ and $(j,k) = (1,1)$. We omit $(j,k) = (0,2)$ because this moment—which is for $P^0Q^2 = Q^2$—carries little price information, so it seems to add more noise than information.

SA.9 Additional Exhibits

Figure SA.15: Expanding the smooth bases one component at a time



Notes: *See notes for Figure 13. In each panel, we only expand the basis for one of the components of* $\tilde{B} \equiv (B_P^D, B_P^S, B_{Z,1}^S)$ *at a time.*