Linear Supply and Demand in Heterogeneous Markets*

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

November 15, 2024

Abstract

We modify the classic linear supply and demand system to allow for the coefficients on price to be unobservable random variables that vary across heterogeneous markets. Known conditions for point identification place strong requirements on the available instruments. We show how to construct and estimate bounds on scalar target parameters that are valid for any type of instrument, or even with no instrument at all. Numerical simulations calibrated to a well-known data set show that the model is not point identified. However, the bounds can be remarkably informative even under limited instrument variation. We apply our approach to study the welfare effects of sales tax.

1 Introduction

Consider the classic linear supply and demand system

$$\begin{aligned} Q_i &= b_1^d + b_p^d P_i + Z_i' b_z^d + U_i^d \qquad \text{(demand)} \\ Q_i &= b_1^s + b_p^s P_i + Z_i' b_z^s + U_i^s \qquad \text{(supply)}, \end{aligned} \tag{1}$$

where i indexes a market, Q_i is quantity, P_i is price, and Z_i are a set of covariates and/or shifters, some of which are usually assumed to enter with zero coefficients in either supply or demand as an exclusion restriction. This model treats the slopes of supply and demand, b_p^s and b_p^d, as constant (deterministic) parameters that are homogeneous across markets. In this paper, we analyze a random coefficient generalization of this model in which the slopes b_p^s and b_p^d are replaced by random variables $B_{p,i}^s$ and $B_{p,i}^d$

*Gaarder: Harris School of Public Policy, University of Chicago. Henry de Frahan, Torgovitsky, and Volpe: Kenneth C. Griffin Department of Economics, University of Chicago. Mogstad: Kenneth C. Griffin Department of Economics, University of Chicago, NBER, and Statistics Norway. Margaret Chen and Arnstein Vestre provided outstanding research assistance.

that are heterogeneous across markets due to factors unobservable to the researcher. Heterogeneity of this sort can arise naturally from spatial or temporal variation in the prices of alternative goods, tax and benefit structures, and unobserved consumer background characteristics, among many other reasons.

Masten (2018) studied identification and estimation in a class of simultaneous equations models that includes the random coefficient supply and demand model we analyze. He showed that the marginal distributions of the random coefficients are point identified if there is a continuous instrument with large support, or a continuous instrument with bounded support under some additional tail restrictions on the distribution of random coefficients. He also provided several negative non-identification results which show that continuous instrument variation is necessary for point identification, and that the joint distribution of random coefficients is fundamentally not point identified, even with large support instruments. See also Hoderlein et al. (2017) for similar negative results in a triangular random coefficient model. These results raise the question of how much can be learned in other settings, such as with discrete instruments, and how much can be learned about target parameters, such as welfare measures, which depend on the joint distribution of supply and demand coefficients.

We answer these questions by allowing for partial identification. We develop an approach for estimating sharp bounds on scalar functions of the joint distribution of random coefficients. The approach uses a linear basis expansion for the density of the underlying random coefficients. Such an expansion 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. Computing the estimator involves solving a sequence of convex linearly-constrained quadratic programming problems, so is feasible in high dimensions and with typical sorts of covariate specifications.

We first use this approach to study the empirical content of the random coefficients model in a numerical simulation calibrated to the well-known Fulton fish market data used by Angrist et al. (2000). The two stage least squares estimand is dramatically smaller than the average of the random coefficient on price in the demand equation. It also provides no guidance about interesting policy counterfactuals, such as the impact on consumer surplus of a marginal tax increase. By contrast, we show that bounds on the average of the price coefficient in the demand equation, as well as bounds on many other interesting target parameters, are informative and often quite tight, even when using an extremely flexible basis for the density of random coefficients. The implication is that the model and data together contain considerable information about the types of objects that a researcher would be interested in, but that one needs to consider

partial identification in order to extract this information.

Our paper builds on a classic econometric literature about the simultaneity problem. Analysis of the linear, constant coefficient model was heavily developed in the classical Cowles foundation work of the 1940s and 1950s (e.g. Koopmans et al., 1953). There is a large literature on linear random coefficient models with exogenous variables (e.g. Beran and Hall, 1992; Hoderlein et al., 2010; Lewbel and Pendakur, 2017; Gaillac and Gautier, 2022; Hermann and Holzmann, 2024). There is also a sizable literature with endogenous variables but triangular systems (e.g. Heckman and Vytlacil, 1998; Masten and Torgovitsky, 2016; Hoderlein et al., 2017), but triangular systems have been shown to be incompatible with simultaneity (Blundell and Matzkin, 2014). However, there are only four papers that allow for both random coefficients and simultaneity: Hurwicz (1950), Kelejian (1974), Hahn (2001), and Masten (2018). As Masten (2018) discusses, 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.

Our work is most related to Masten (2018), who provided several novel nonparametric point identification results with continuous instruments, as well as several novel non-identification results which show that continuity in the instrument is essential for point identification. Our contribution is to consider partial identification, although our framework nests point identification as a special case. The approach we consider allows for any type of instruments and both parametric and nonparametric specifications of the random coefficient distribution.

The linear random coefficient model is an example of a nonseparable model—a model in which the unobservables do not enter in an additively separable fashion. Matzkin (2008, 2015) and Berry and Haile (2018) provided identification results for nonseparable, nonparametric models that generalize the constant coefficient model in two ways: by relaxing linearity in price, and by allowing for nonseparability. The class of models they consider have one unobservable per endogenous variable, while the random coefficients model has at least two unobservables per endogenous variable. The two classes of models are not nested. We find that the additional dimensions of heterogeneity in the random coefficients model are useful for motivating microfoundations from traditional functional forms (see Section 2.2). They can also help accommodate empirical challenges, such as potential measurement error, or omitted but exogenous variables, which can be difficult to rationalize through a single unobservable.

Leamer (1981) and Manski (1995, 1997) also considered partial identification in simultaneous equations models. Leamer (1981) focused on a constant coefficients framework with no instrument, whereas we allow for random coefficients and instruments.

Manski (1995, 1997) considered the identifying content of assuming that demand slopes downward while allowing for completely general forms of nonlinearity and heterogeneity. Manski (1995, 1997) did not consider instruments, but they could be exploited using the results of Manski and Pepper (2000), who did not consider the simultaneity problem explicitly. Our focus on a linear random coefficients model entails considerably stronger functional form assumptions, which allows us to scale up our approach to a typically-sized empirical application and produces much tighter bounds.

In the next section, we define and microfound the random coefficients model. In Section 3, we revisit some results from Angrist et al. (2000) on interpreting linear IV estimators when the constant coefficient model is misspecified and the actual model has random coefficients. We extend several of their results, using the discussion to motivate the need for introducing our new approach. In Section 4, we discuss the identification problem, abstracting from statistical uncertainty in the distribution of observables, and report the results of numerical simulations that show how partial identification arises with limited instrument variation. We then propose a computationally-tractable estimator in Section 5. In Section 6, we apply our methodology to study the consumer surplus incidence of sales taxes. Section 7 contains a brief conclusion.

2 Linear Supply and Demand with Random Coefficients

In this section, we formally define the model, notation, and assumptions that we use throughout the paper. We provide economic microfoundations that motivate the consideration of random coefficients. We discuss natural target parameters and their relationship to the random coefficient distribution.

2.1 Random coefficients model

We write the random coefficients generalization of (1) as

$$\begin{aligned} Q &= Z' B_z^d - h^d(P)' B_p^d \qquad \text{(demand)} \\ Q &= Z' B_z^s + h^s(P)' B_p^s \qquad \text{(supply)}, \end{aligned} \tag{2}$$

where $B \equiv (B_z^d, B_p^d, B_z^s, B_p^s)$ is an unobserved vector of random coefficients, h^d and h^s are known, vector-valued functions, P and Q are observed equilibrium price and quantity, and Z is a vector of observed exogenous covariates and/or instruments that includes a constant term. Quantity and/or price could be measured in levels or logs, but we keep this implicit in the notation until it becomes relevant. The additive residual and constant terms in the classical model (1) have been combined into the component

of the random coefficients B_z^d and B_z^s that corresponds to the constant term of Z. We suppress the market subscript i until considering estimation in Section 5.

We assume throughout that $p \mapsto h^d(p)'B_p^d$ and $p \mapsto h^s(p)'B_p^s$ are always increasing functions of p, so that demand slopes down and supply slopes up. This implies that at most one (P, Q) pair can satisfy (2). We further assume that the demand and supply curves always intersect, so that an equilibrium exists and there is always exactly one (P, Q) pair that satisfies (2). For any realization of (B, Z), let $\epsilon^{\mathrm{P}}(B, Z)$ and $\epsilon^{\mathrm{Q}}(B, Z)$ denote the equilibrium price and quantity, so that the reduced form equations are:

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

In the leading case that $h^d(p) = p$ and $h^s(p) = p$ are linear functions of p, the reduced form equations have a simple closed form:

$$P = Z'\left(\frac{B_z^d - B_z^s}{B_p^d + B_p^s}\right) \quad \text{and} \quad Q = Z'\left(\frac{B_p^s B_z^d + B_p^d B_z^s}{B_p^d + B_p^s}\right). \tag{4}$$

In more general cases, ϵ^{P} and ϵ^{Q} may need to be computed numerically. This does not create any additional conceptual problems for our partial identification analysis in Section 4, although it does add to the computational challenge.

Masten (2018) observed that the moments of the reduced form equations might not exist when the coefficients on price are random. In the supply and demand setting with linear functions of price and reduced form (4), this could happen if B_p^d and/or B_p^s have densities that put a large amount of mass near zero. Most of our analysis is based on distribution functions, so does not depend on the existence of reduced form moments. In cases where it does, we assume without further mention that the appropriate moments exist.

We maintain the exogeneity condition that Z is independent of B. Exclusion restrictions will be imposed by assuming that certain components of B_z^d or B_z^s are constant and equal to zero. For example, assuming that the third component of B_z^d is zero would imply that Z_3 is a supply shifter, an assumption which would traditionally be used to estimate a constant demand slope b_p^d in the classical model. 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.

2.2 Microfoundations

In this section, show how the random coefficients system (2) can be motivated from the choices of consumers and firms. We focus on a linear example, though other known functional forms of deterministic functions h^d and h^s in (2) can be similarly micro-founded.

The first proposition provides conditions on consumer preferences that lead to a linear market demand with heterogeneous coefficients. The second proposition provides conditions on a firm's production function that lead to linear market supply with heterogeneous coefficients. Proofs for both propositions are in Appendix A.

Proposition 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*

$$U_j\left(q_{0,j}, q_j\right) = \begin{cases} q_{0,j} + \xi_j q_j^{\chi} - \gamma_j & \text{if } q_j > 0 \\ q_{0,j} & \text{if } q_j = 0 \end{cases}, \tag{5}$$

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 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. Then

$$\log(Q) = B_z^d - B_p^d \log(P)$$
$$\text{where} \quad B_z^d \equiv \log\left(\chi^{\frac{1+\psi\chi}{1-\chi}}\left(\frac{1-\chi}{\overline{\gamma}}\right)^{\psi}\int \xi^{\frac{1+\psi}{1-\chi}} N\left(\xi\right) d\xi\right) \quad \text{and} \quad B_p^d \equiv \frac{1+\psi\chi}{1-\chi}.$$

Proposition 1 gives one example of how a market demand equation that is linear in the logarithms of price and quantity can be derived from individual consumer behavior. The parametric assumptions are strong and stylized. The point of the proposition is not that one should necessarily embrace these assumptions. Instead, it is that *even if* one maintains these assumptions, the intercept and slope of the resulting demand

curve will depend on preference parameters that are likely to vary across markets. In Proposition 1, the slope coefficient depends on utility parameters χ and ψ that characterize how preferences differ across individuals. 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.

Proposition 2 (**Supply**). *Suppose that each market is characterized by a representative firm that produces output of the focal good in Proposition 1 using labor L and capital K through a Cobb-Douglas production function:*

$$F(K, L) \equiv AK^{\mathfrak{a}\mathfrak{g}}L^{(1-\mathfrak{a})\mathfrak{g}},$$

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

$$C(K, L) \equiv w_K^0 K^{1+\mathfrak{b}_K} + w_L^0 L^{1+\mathfrak{b}_L},$$

where $\mathfrak{b}_K, \mathfrak{b}_L \geq 0$ 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. Then

$$\begin{aligned} \log(Q) &= B_z^s + B_p^s \log(P) \\ \textit{where} \quad B_z^s &\equiv \iota^s(\mathfrak{a}, \mathfrak{g}, \mathfrak{b}_K, \mathfrak{b}_L, A, w_K^0) \\ \textit{and} \quad B_p^s &\equiv \frac{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_K) + (1-\mathfrak{a})(1+\mathfrak{b}_L)\right]}{(1+\mathfrak{b}_K)(1+\mathfrak{b}_L) - \mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_K) + (1-\mathfrak{a})(1+\mathfrak{b}_L)\right]}. \end{aligned}$$

where ι^s is a complicated function whose expression is given in the appendix.

Proposition 2 gives a supply-side counterpart to Proposition 1. The parametric assumptions are again strong and stylized, but the point is that even under these assumptions, the coefficient on price will depend on parameters that characterize both the production function and the input market. If the structure of the production function or input market differ across product markets, then the supply equation will have random coefficients. Differences in the structure of the input market elasticities in particular could be caused by differences in monopsony power or taxes.

2.3 Counterfactuals and target parameters

The marginal distributions of B_p^d and B_p^s control how supply and demand would change if there were exogenous shocks to one side of the market. If P and Q are specified in levels and (2) is assumed to be linear ($h^d(P) = h^s(P) = P$), then the average effect over markets of a supply-driven shock that increases price by a known fixed amount ΔP in all markets is $\mathbb{E}[\Delta Q] = -\mathbb{E}[B_p^d]\Delta P$. The marginal distributions of B_p^d and B_p^s can also be used to determine the excess supply created by imposing a binding price floor. Masten (2018) provided positive point identification results for marginal distributions that could be used to identify target parameters that summarize these types of counterfactuals, although the conditions require a continuous instrument.

For most counterfactuals, the relevant target parameters depend on the joint distribution of random coefficients, not just the marginals. In Section 6, we model Q and P in logs, and our primary counterfactual of interest is a change in the ad valorem sales tax, either by a marginal or discrete amount. For a marginal increase, the corresponding change in average log quantity and average log price are determined by the relative supply and demand elasticities:

$$\mathbb{E}\left[\Delta Q\right] = -\,\mathbb{E}\left[\frac{B_p^d B_p^s}{B_p^d + B_p^s}\right] \quad \text{and} \quad \mathbb{E}\left[\Delta P\right] = -\,\mathbb{E}\left[\frac{B_p^d}{B_p^d + B_p^s}\right]. \tag{6}$$

These averages depend on the joint distribution of B_p^d and B_p^s. Masten (2018) shows that this joint distribution is fundamentally not point identified without additional parametric assumptions. In fact, even in the classical, constant coefficient case, the dependence of these quantity and price responses on both the supply and demand slopes means that both supply and demand shifters would be necessary for point identification. These observations motivate the partial identification approach developed in Section 4, which gets around both Masten's negative result and the classical requirement of having shifters for both sides of the market.

In our application, we also consider standard welfare counterfactuals related to either a marginal or discrete sales tax change. Assuming there are no income effects, these counterfactuals can often be summarized through target parameters that depend only on the market level supply and demand curves (e.g. Harberger, 1964; Chetty, 2009). However, they are complicated functions of the joint distribution of random coefficients. For example, in Appendix B, we show that the average relative consumer

surplus impact (the incidence) of a marginal tax increase is given by

$$\text{INC} = \mathbb{E}\left[\frac{(1+t)B_p^s}{(1+t)B_p^s(1+t)+B_p^d}\right], \tag{7}$$

where t is the existing tax rate, assuming (for simplicity), that it is the same across markets. As another example, we also show in Appendix B that the average ratio of deadweight loss to revenue increased from a marginal tax increase is given by

$$\text{DWL} = \mathbb{E}\left[\frac{tB_p^dB_p^s}{B_p^d+B_p^s+t\left(1-B_p^d\right)B_p^s}\right]. \tag{8}$$

Expressions for the consumer surplus and deadweight loss impacts of a discrete tax increase, which are considerably more complicated, are also derived in Appendix B. The partial identification approach that we develop is computational, which makes it applicable to these and other target parameters, even despite their complicated dependence on the distribution of random coefficients.

3 Interpreting Linear IV Estimators

The traditional identification result for the classical linear model with constant coefficients is that the coefficients in the demand equation are identified if there is an excluded, exogenous, and relevant supply shifter that can be used as an instrument for price (see, for example Manski, 1995, Chapter 6). Angrist et al. (2000, Corollary 2) considered the interpretation of the IV estimand when the data is generated by a model with random coefficients. They analyzed the special case of a binary excluded supply shifter with no covariates. In this section, we extend their analysis to allow for a vector of any type of discrete or continuous instruments, as well as a vector of covariates.

Let β_{tsls} 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

$$\beta_{\text{tsls}} = \frac{\mathbb{E}[Q(\dot{P} - \mathbb{L}[\dot{P}|Z_2])]}{\mathbb{E}[P(\dot{P} - \mathbb{L}[\dot{P}|Z_2])]}, \tag{9}$$

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 first proposition shows that β_{tsls} will be a weighted

average of the demand slope if Z_1 are supply shifters that are excluded from the demand equation.

Proposition 3. *Suppose that $h^d(P) = P$ and $h^s(P) = P$. Divide $Z = (Z_1, Z_2)$ and assume that $B_z^d = (0, B_{z2}^d)$, so that Z_1 correspond to excluded supply shifters. Let $\tilde{Z}_1 \equiv Z_1 - \mathbb{L}[Z_1|Z_2]$ denote the population residuals from a linear regression of each component of Z_1 onto Z_2. Then*

$$-\beta_{tsls} = \mathbb{E}\left[B_p^d W\right]$$
$$\text{where} \quad W \equiv \mathbb{E}\left[\frac{B_{z1}^s}{B_p^d + B_p^s}\middle| B_p^d\right]' \frac{\mathbb{E}[\tilde{Z}_1 \tilde{Z}_1' \delta_1]}{\mathbb{E}[(\tilde{Z}_1'\delta_1)^2]}, \quad \text{with} \quad \delta_1 \equiv \mathbb{E}\left[\frac{B_{z1}^s}{B_p^d + B_p^s}\right].$$

The weights W satisfy $\mathbb{E}[W] = 1$.

Note that the interpretation in Proposition 3 is for $-\beta_{\text{tsls}}$ rather than β_{tsls} simply because of our normalization in (2) that B_p^d is non-negative.

Blandhol et al. (2022) show that linear IV estimands for specifications that include covariates will not in general be equal to a weighted average of causal effects unless $\mathbb{L}[Z_1|Z_2] = \mathbb{E}[Z_1|Z_2]$. Proposition 3 shows that the random coefficients structure breaks this necessary condition, ensuring that β_{tsls} is a weighted average of B_p^d with weights that sum to one. However, the weights can be negative. The next proposition provides two sufficient conditions for the weights to be non-negative.

Proposition 4. *Suppose that the conditions of Proposition 3 are satisfied. Then $\mathbb{P}[W \geq 0] = 1$ if either*

(a) Z_1 is scalar and either $\mathbb{P}[B_{z1}^s \geq 0] = 1$ or $\mathbb{P}[B_{z1}^s \leq 0] = 1$.

(b) B_{z1}^s is independent of (B_p^d, B_p^s).

The first condition in Proposition 4 is the same as in Angrist et al. (2000, Corollary 2). The assumption that B_{z1}^s takes a single sign is what those authors refer to as the monotonicity condition. Requiring Z_1 to be scalar is important for the monotonicity condition to reflect a sensible ordering (Mogstad et al., 2021). The second condition in Proposition 4 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.

Weighting results like Proposition 4 are commonly found in reverse engineering discussions of linear IV with heterogeneous treatment effects (e.g. Angrist and Pischke, 2009; Mogstad and Torgovitsky, 2024). Their attraction is in ensuring that if the underlying causal effect has the same sign for all units, then the IV estimand also has that

same sign. In the supply and demand context, the underlying "causal effect" of price on demand is already assumed to be non-negative, so it is unclear what can be learned by this type of weighting result. Knowing that the IV estimand is a non-negative weighted average in only helpful insomuch as it implies that obtaining a negative IV estimand is due to some deficiency of the underlying assumptions rather than some unattractive quirk of the IV estimand itself. It is not helpful for making constructive statements about magnitude.

One way to see this point is to sign the IV estimand relative to an interpretable feature of demand, such as the average slope, $\mathbb{E}[B_p^d]$. 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 β_{tsls} is smaller than the average slope of demand, $\mathbb{E}[B_p^d]$. The next proposition shows that this conclusion holds in considerably more generality.

Proposition 5. *Suppose that the conditions of Proposition 3 are satisfied. Then*

$$-\beta_{tsls} = \mathbb{E}[B_p^d] + \mathbb{C}[B_p^d, W]. \tag{10}$$

Suppose further that condition (b) of Proposition 4 is satisfied. Then $-\beta_{tsls} \leq \mathbb{E}[B_p^d]$ *if and only if*

$$\mathbb{C}\left[B_p^d, \frac{1}{B_p^d + B_p^s}\right] \leq 0. \tag{11}$$

In particular, (11) *is satisfied under either of the following two conditions:*

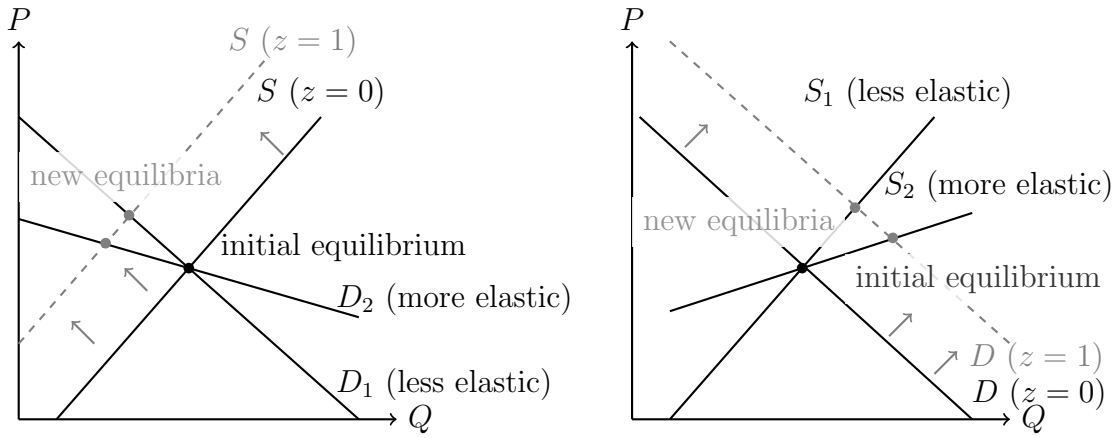
(a) $\mathbb{E}[B_p^d \mid B_p^s] = \mathbb{E}[B_p^d]$.

(b) $\mathbb{E}[(B_p^d + B_p^s)^{-1} \mid B_p^d = b_p^d]$ *is a weakly decreasing function of* b_p^d.

The intuition behind Proposition 5 can be seen with the aid of a standard Marshallian cross. The left side of Figure 1 shows the impact of an 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 change more in the less elastic market because quantity demanded in that market declines less rapidly as price increases. This means that the instrument (supply shifter) has a larger impact on the endogenous variable (price) in markets with more inelastic demand, leading these markets contribute more to the statistical weighting used by the IV estimator.

This reasoning depends on the slope of supply as well. The right side of Figure 1 shows the same additive supply shift in two markets with the same demand curve but

Figure 1: *Supply shifts lead to a larger price change when demand is more inelastic and vice versa.*



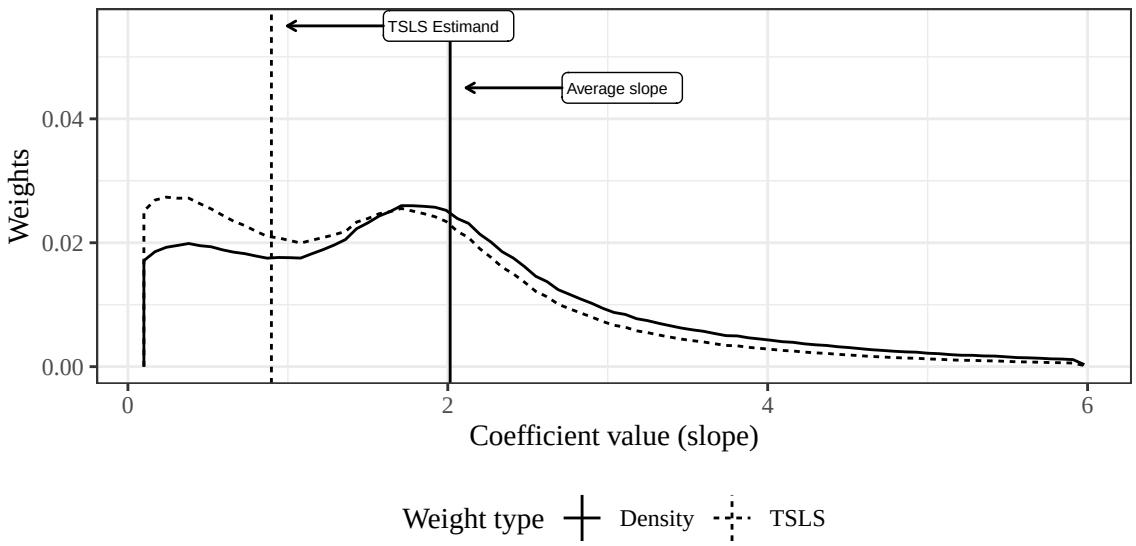
different supply curves. Prices again change more in the market with less elastic supply because prices need to increase more to restore equilibrium quantity. Together, Figure 1 shows that markets that are more inelastic—both in supply and demand—receive larger weight in the IV estimand.

If markets that have highly inelastic demand also tend to have highly *elastic* supply, so that B_p^d and B_p^s are strongly negative dependent, then it's possible for the IV estimand to overstate the average demand slope. The condition given in Proposition 5 quantifies how strong this negative dependence needs to be.

All of this reasoning implicitly depends on the size of the supply shifts B_{z1}^s being independent of the price elasticities in different markets, or at least, larger in more inelastic markets. If the supply shocks have a systematically larger impact in more elastic markets, then the relationship of the IV estimand to the average slope of demand becomes ambiguous.

Proposition 5 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 (2), with an equation like $P = Z'\tilde{B}$ for random coefficients $\tilde{B}$. Heckman and Vytlacil (1998) observed that if the component of $\tilde{B}$ 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. Proposition 5 shows that this type of reasoning no longer applies when there is simultaneity. The reason is that 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 language of Heckman et al.

Figure 2: *The TSLS estimand is smaller than the average demand slope.*



(2006).

To get a sense of the magnitudes involved in Proposition 5, we fit a data generating process (DGP) to the Fulton fish market data used by Angrist et al. (2000). We discuss the details of how we constructed the DGP in Section 4.4, where we also use it to compute bounds on some of the target parameters discussed in Section 2.3. Figure 2 shows the linear IV (TSLS) estimand and the average slope of demand in this DGP together with the underlying weights on the heterogeneity demand slope that determine these quantities. The linear IV estimand is less than half the size of the average coefficient on demand, consistent with the intuition supported by Proposition 5. This is due to the weights for the IV estimand overweighting markets that are highly inelastic. These results suggest that the linear IV estimand is not a particularly useful measure of the average slope of demand.

4 Partial Identification

In this section, we develop an approach for computing identified sets of a general class of target parameters. For the identification analysis we assume that we know the population joint distribution of (P, Q) conditional on Z. This analysis forms the basis of the estimation procedure in the next section, in which we observe only a finite sample from the population distribution.

4.1 The identified set

The random coefficients model is fully parameterized by the distribution of B. Let F denote the cumulative distribution function of B. Let $\mathcal{F}$ denote the set of distribution functions F that satisfy the researcher's prior assumptions.

Each distribution F implies a conditional distribution of (P, Q) given Z via the reduced form relationship (4):

$$g(p, q, z; F) \equiv \int \mathbb{1}[\epsilon^{\mathrm{P}}(b, z) \leq p, \epsilon^{\mathrm{Q}}(p, z) \leq q] \, dF(b), \tag{12}$$

where the integration is of the Lebesgue-Stieltjes type. Denote the actual conditional distribution of price and quantity as

$$g(p, q, z) \equiv \mathbb{P}[P \leq p, Q \leq q | Z = z].$$

The identified set for F is then defined as

$$\mathcal{F}^\star \equiv \{F \in \mathcal{F} : g(p, q, z; F) = g(p, q, z) \text{ for all } (p, q) \text{ and almost every (a.e.) } z\},$$

that is, as the set of all distributions F that both satisfy the researcher's assumptions and reproduce the distribution of the observed data.

The researcher's target parameter is a scalar-valued functional $\tau : \mathcal{F} \to \mathbb{R}$ that maps each F into a single summary quantity of interest. Any of the target parameters discussed in Section 2.3 can be chosen as τ. The identified set for the target parameter is defined as the image of $\mathcal{F}^\star$ under τ:

$$\mathcal{T}^\star \equiv \tau(\mathcal{F}^\star) = \{t \in \mathbb{R} : t = \tau(F) \text{ for some } F \in \mathcal{F}^\star\}. \tag{13}$$

The smallest and largest elements of $\mathcal{T}^\star$ can be expressed as the solutions to two optimization problems

$$\tau^\star_{\mathrm{lb}} \, (\tau^\star_{\mathrm{ub}}) \equiv \min_{F \in \mathcal{F}} (\max) \, \tau(F) \quad \text{s.t.} \quad g(p, q, z; F) = g(p, q, z) \text{ for all } (p, q), \text{ a.e. } z. \tag{14}$$

The optimal values are sharp bounds in the sense that $[\tau^\star_{\mathrm{lb}}, \tau^\star_{\mathrm{ub}}]$ is the smallest closed interval that contains $\mathcal{T}^\star$.

Masten (2018) provided several novel point identification results for simultaneous equations models with random coefficients. All of his positive results require the instrument to have jointly continuous support, and some of them require the instruments to have positive support on the entire real line ("large support"). Masten (2018) also

provided several negative non-identification results which show that this type of continuous variation is *necessary* for point identification. For example, using a continuous variable and its square as an instrument leads to a failure of point identification even for the reduced form coefficients (Masten, 2018, Corollary 1). The joint distributions of the random coefficients are also not point identified even when the instruments have large support (Masten, 2018, Theorem 4). This suggests that target parameters which depend on both supply and demand, such as welfare measures, will also fail to be point identified even under extremely strong conditions on the instruments. Taken together, Masten's findings suggest that point identification of interesting target parameters is rarely obtained with random coefficients.

Failure of point identification does not mean that the model does not admit useful bounds on interesting target parameters. Our characterization of the sharp bounds in (14) recognizes this possibility, allowing for the instrument to have any type of support while still permitting point identification as a special case. Instruments with richer supports create more equality constraints in (14) and thus necessarily reduce the width of the identified set. Greater instrument variation is rewarded in the form of tighter bounds, but is not required to proceed with the empirical analysis.

4.2 Computing identified sets

To make computing (14) tractable, we assume that the set of random coefficient distributions considered by the researcher can be parameterized as

$$\mathcal{F} = \left\{ F : F(b) = \sum_{k=1}^{d_\phi} \phi_k F_k(b;\alpha) \quad \text{for some } (\phi,\alpha) \in \Phi\mathcal{A} \right\}, \tag{15}$$

where $\{F_k(b;\alpha)\}_{k=1}^{d_\phi}$ are a collection of known basis functions. There are two parameters in this specification: a linear parameter, ϕ, which takes values in a known set $\Phi \subseteq \mathbb{R}^{d_\phi}$, and a nonlinear parameter, α, which takes values in a known set $\mathcal{A} \subseteq \mathbb{R}^{d_\alpha}$. The distinction between these two parameters is important for computation. The set $\Phi\mathcal{A} \subseteq \mathbb{R}^{d_\phi + d_\alpha}$ is possibly a proper subset of $\Phi \times \mathcal{A}$ to allow for the feasible set of ϕ to depend on α, a generality we will need in Section 4.3.

The specifications we consider will be such that each $F_k(\cdot;\alpha)$ is a distribution function for each fixed α. This means that for fixed α, F is a mixture distribution over F_k with mixing weights ϕ_k. We will often take d_ϕ to be large, so that these mixtures are quite flexible. This allows us to consider bounds generated from rich parameter spaces that approximate nonparametric specifications, like a sieve (e.g. Chen, 2007). However,

(15) also allows for more tightly parameterized specifications that will produce more informative bounds and possibly even point identification of certain target parameters.

We primarily use the nonlinear parameter α to allow for some of the basis functions to be degenerate point masses. For example, we will often restrict some of the components of B_z^d and B_z^s that correspond to covariates to be constant, while still allowing for randomness in the more salient components of B, such as the intercept and the coefficient on price. The constant components are treated as constant coefficients that are contained in α. Changing these constant coefficients changes F through the definition of one or more of the basis components $F_k(\cdot;\alpha)$ rather than through the mixing weights.

Assuming that $\mathcal{F}$ has the structure given in (15) implies that

$$g(p,q,z;F) = \sum_{k=1}^{d_\phi} \phi_k \int \mathbb{1}[\epsilon^{\mathrm{P}}(b,z) \leq p, \epsilon^{\mathrm{Q}}(b,z) \leq q]\, dF_k(b;\alpha) \equiv \sum_{k=1}^{d_\phi} \phi_k \bar{g}_k(p,q,z;\alpha), \tag{16}$$

which is now a linear function of ϕ with coefficients $\bar{g}_k(p,q,z;\alpha)$ that can be calculated numerically for any given α. We focus our attention on target parameters that have a similar form,

$$\tau(F) = \int t(b)\, dF(b) \tag{17}$$

for some known function t that could depend on the distribution of observables. All of the target parameters discussed in Section 2.3 can be written in this form. If $\mathcal{F}$ is specified as (15) then

$$\tau(F) = \sum_{k=1}^{d_\phi} \phi_k \int t(b)\, dF_k(b;\alpha) \equiv \sum_{k=1}^{d_\phi} \phi_k \bar{\tau}_k(\alpha) \tag{18}$$

is also a linear function of ϕ with coefficients $\bar{\tau}_k(\alpha)$ that can be calculated given α.

We exploit this linearity by profiling α out of the optimization problem (14) that characterizes the sharp bounds on $\mathcal{T}^\star$. For the lower bound problem,

$$\tau_{\mathrm{lb}}^\star = \min_{\alpha \in \mathcal{A}} \left[\min_{\phi \in \Phi(\alpha)} \sum_{k=1}^{d_\phi} \phi_k \bar{\tau}_k(\alpha) \quad \text{s.t.} \quad \sum_{k=1}^{d_\phi} \phi_k \bar{g}_k(p,q,z;\alpha) = g(p,q,z) \text{ for all } p,q,z \right], \tag{19}$$

where $\Phi(\alpha) \equiv \{\phi : (\phi,\alpha) \in \Phi\mathcal{A}\}$. (We follow the usual convention that the minimum

of an infeasible program is positive infinity in case α is such that the inner problem is infeasible.) The inner problem of (19) is a linear program as long as $\Phi(\alpha)$ is polyhedral, which it will be in the leading case when ϕ are mixing weights and $\Phi(\alpha) = \Phi$ is the simplex for any α. The upper bound problem maximizes over both α and ϕ, and we profile it in the same way as an outer maximization over α and an inner maximization of ϕ given α.

Since P and Q are potentially continuously distributed, the constraints in the inner problem cannot be imposed for all p and q in their support. A simple approach is to impose the constraints for a finite (but large) grid of (p, q) and all z, although this can lead the resulting bounds to be potentially non-sharp. We make the grid as rich as possible while keeping computation manageable. As a further defense against numerical non-sharpness, we also evaluate the candidate solutions on a much larger grid, add points of large constraint violations to the program, and then re-solve the optimization problems, iterating a few times until the constraints are satisfied on the larger grid as well. These computational considerations are only an issue for our numerical simulations that use the population distribution; the estimator we develop in Section 5 asymptotically incorporates all of the information in the distribution of (P, Q).

4.3 Constant coefficients

The outer problem of (19) can be solved with an unstructured optimization over $\mathcal{A}$, such as a grid search. In general, this is only reliable if α is a low dimensional parameter. However, for our leading case in which α is used to characterize constant coefficients, the linear structure can be used to reduce dimension of α down to two in general, or one with an exclusion restriction.

To see this, separate $Z \equiv (Z_1, Z_2)$ into two subvectors, where the coefficients on Z_1 are random and the coefficients on Z_2 are specified as constant in both the supply and demand equations. Write the coefficient vectors as $B_z^d \equiv (B_{z1}^d, b_{z2}^d)$ and $B_z^s \equiv (B_{z1}^s, b_{z2}^s)$. Substituting this notation into the reduced form equation (4) and taking expectation (assuming existence of moments) produces

$$\mathbb{E}[P|Z] = Z_1' \, \mathbb{E}\left[\frac{B_{z1}^d - B_{z1}^s}{B_p^d + B_p^s}\right] + Z_2' \left(b_{z2}^d - b_{z2}^s\right) \mathbb{E}\left[\frac{1}{B_p^d + B_p^s}\right],$$

$$\mathbb{E}[Q|Z] = Z_1' \, \mathbb{E}\left[\frac{B_{z1}^d B_p^s + B_{z1}^s B_p^d}{B_p^d + B_p^s}\right] + Z_2' \left(b_{z2}^d \, \mathbb{E}\left[\frac{B_p^s}{B_p^d + B_p^s}\right] + b_{z2}^s \, \mathbb{E}\left[\frac{B_p^d}{B_p^d + B_p^s}\right]\right).$$

The coefficients on Z_2 in linear regressions of P and Q onto Z are thus

$$\rho_2^p \equiv \beta^c \left(b_{z2}^d - b_{z2}^s \right), \qquad \text{where} \quad \beta^c \equiv \mathbb{E}\left[\frac{1}{B_p^d + B_p^s} \right],$$
$$\text{and} \quad \rho_2^q \equiv (1-\beta^d) b_{z2}^d - \beta^d b_{z2}^s, \qquad \text{where} \quad \beta^d \equiv \mathbb{E}\left[\frac{B_p^d}{B_p^d + B_p^s} \right].$$

Solving this system of equations for b_{z2}^d and b_{z2}^s gives

$$b_{z2}^d = \rho_2^q + \frac{\beta^d}{\beta^c} \rho_2^p \quad \text{and} \quad b_{z2}^s = \rho_2^q + \left(\frac{\beta^d - 1}{\beta^c} \right) \rho_2^p. \tag{20}$$

Equation (20) shows that the constant coefficient vectors b_{z2}^d and b_{z2}^s are fully determined by the scalars β^d and β^c together with the reduced form regression coefficients ρ_2^p and ρ_2^q. Let $\alpha = (\beta^d, \beta^c, b_{z2}^d, b_{z2}^s)$. Then (20) shows that α is two dimensional given knowledge of ρ_2^p and ρ_2^q, regardless of the dimension of Z_2. The grid search over $\mathcal{A}$ in the outer problem of (19) has dimension two, which is computationally manageable. In practice, we fix a point of (β^d, β^c) on the grid, solve for b_{z2}^d and b_{z2}^s, then solve the inner problem of (19) while adding the deterministic constraints

$$\beta^c = \sum_{k=1}^{d_\phi} \phi_k \int \left(\frac{1}{b_p^d + b_p^s} \right) f_k(b; \alpha) \, d\mu(b),$$
$$\beta^d = \sum_{k=1}^{d_\phi} \phi_k \int \left(\frac{b_p^d}{b_p^d + b_p^s} \right) f_k(b; \alpha) \, d\mu(b)$$

to the definition of $\Phi(\alpha)$. We make use of this strategy when including covariates in our numerical simulations and application.

If there is an exclusion restriction, then the dimension of α can be further reduced from two to one. Suppose that the coefficient on the first component of Z_2 in the demand equation is known to be zero, so that this component is an excluded supply shifter. Then (20) implies that

$$0 = \rho_{21}^q + \frac{\beta^d}{\beta^c} \rho_{21}^p \quad \text{or} \quad \beta^c = -\frac{\rho_{21}^q}{\rho_{21}^p} \beta^d, \tag{21}$$

so that β^c is fully determined by β^d and the reduced form regression coefficients.

If both the first and second demand coefficients of Z_2 are known to be zero, so that there are two exclusion restrictions in the demand equation, then (21) for both components produces the testable implication that $\rho_{21}^q / \rho_{21}^p = \rho_{22}^q / \rho_{22}^p$. (This requires

the mild assumption that $\beta^d \neq 0$, which could only happen if $B_p^d = 0$ deterministically.) This represents a partial preservation of the familiar overidentification test from the constant coefficient case to the random coefficient case. The difference is that the overidentified quantity is not the slope on price in the demand equation—which is now non-random—but rather the ratio β^d/β^c, which is a weighted average of the slope of price in the demand equation.

A third case is when one coefficient on Z_2 in the demand equation is zero, $b_{z21}^d = 0$, and a coefficient for a different component on Z_2 is zero in the supply equation, $b_{z22}^s = 0$. Then (21) still holds, but the second equation of (20) additionally implies that

$$0 = \rho_{22}^q + \left(\frac{\beta^d - 1}{\beta^c}\right)\rho_{22}^p. \tag{22}$$

Combining (21) and (22) shows that β^c and β^d are point identified from the reduced form regression coefficients:

$$\beta^d = \left(1 - \frac{\rho_{21}^q \rho_{22}^q}{\rho_{21}^p \rho_{22}^p}\right)^{-1} \quad \text{and} \quad \beta^c = -\frac{\rho_{21}^q}{\rho_{21}^p}\left(1 - \frac{\rho_{21}^q \rho_{22}^q}{\rho_{21}^p \rho_{22}^p}\right)^{-1}.$$

The outer problem of (19) disappears entirely in this case.

4.4 Numerical simulations

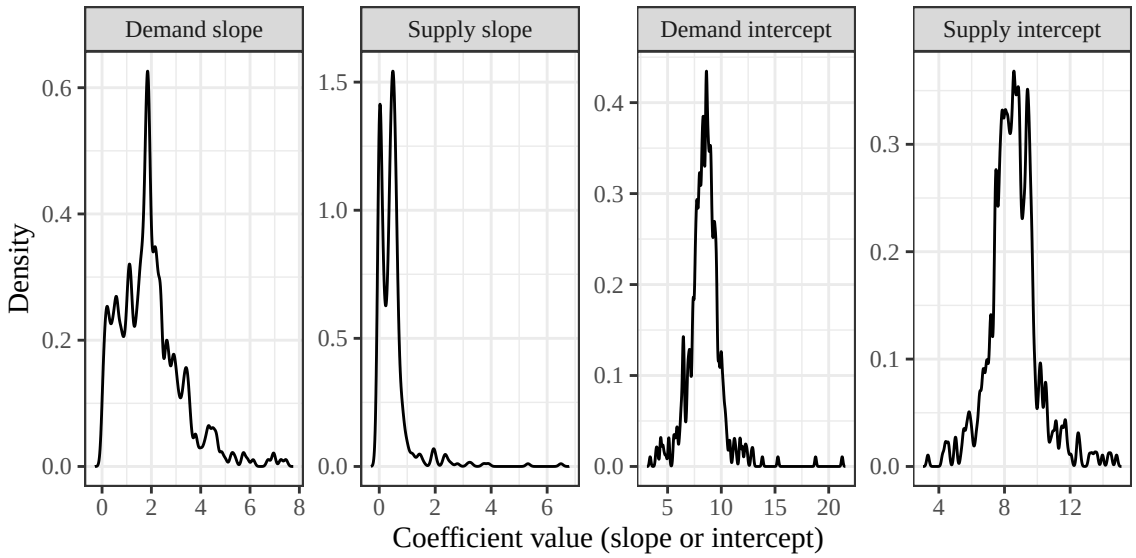
In this section, we report the results of numerical simulations designed to shed some light on the width of bounds for the random coefficient model.

We constructed the data generating process (DGP) by fitting a random coefficients model to the Fulton fish market data used by Angrist et al. (2000) (see also Graddy, 1995). The data consists of daily aggregated prices and quantity for whiting (a type of fish) on each of 111 weekdays, which are viewed as the market. We use log quantity and price for Q and P. There are two mutually exclusive binary instruments indicating stormy or mixed weather at sea, which are used as excluded supply shifters. The covariates are a constant (intercept) and four binary day-of-the-week indicators, and two binary indicators for good weather and rain on the shore. See Angrist et al. (2000, Section 5) for more detail.

We fit the random coefficients model by minimizing a sample criterion function, the population version of which is zero if and only if the random coefficient density is in the population identified set. The construction of this criterion function is discussed in the next section.

We use a linear basis for the four-dimensional random coefficients reflecting the

Figure 3: *Distribution of random coefficients in the fish market DGPs*

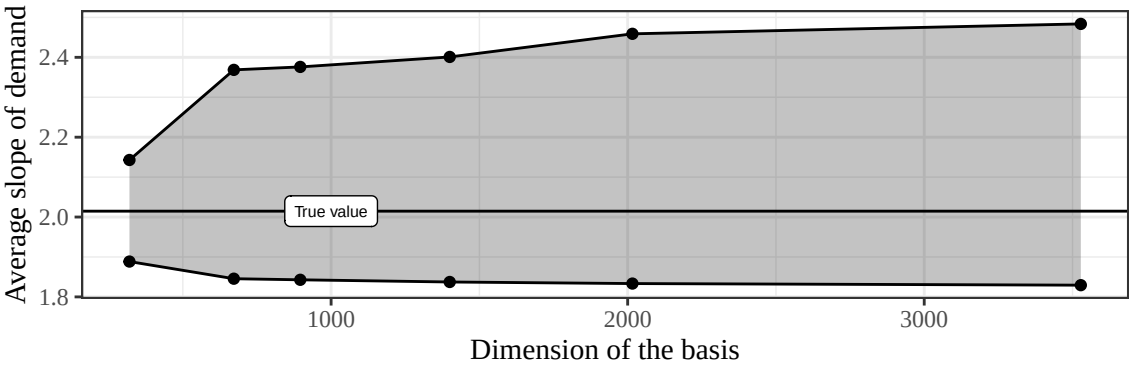


intercept and price. We allow for the coefficients on the intercept and price to be random in both equations, while we restrict all other coefficients to be constant. We set the basis to be a tensor product of a mixture of Erlang (gamma with integer shape) distributions with four or five terms, leading to $d_\phi = 320$. Figure 3 plots the densities of the random coefficients in the DGP.

Figure 4 plots bounds on the average slope of demand as a function of the number of terms in the basis, keeping the DGP fixed. We compute these bounds by solving (19) on a grid of 100 price and quantity pairs crossed with all $2^8 = 256$ values that the eight binary variables in Z can take. We found that making the grid larger than this did not appreciably reduce the bounds for the subset of our results that we tested. The bounds are remarkably tight, ranging from about 1.8 to 2.5 for the richest basis. Notably, the lower bound of both set of bounds is more than twice as large as the linear IV estimand shown in Figure 2. The bounds widen as the basis becomes richer, allowing for more flexible distributions.

There are three important takeaways from Figure 4. First, a natural target parameter is not point identified with binary instruments, consistent with Masten's (2018) results. Second, while not point identified, the identified sets are still remarkably informative. Third, the bounds remain wide even under an extremely flexible basis for the random coefficient distribution, suggesting that the bounds reflect the nonparametric structure of the model, rather than any particular functional form for the distribution

Figure 4: *Bounds on the average slope of demand*

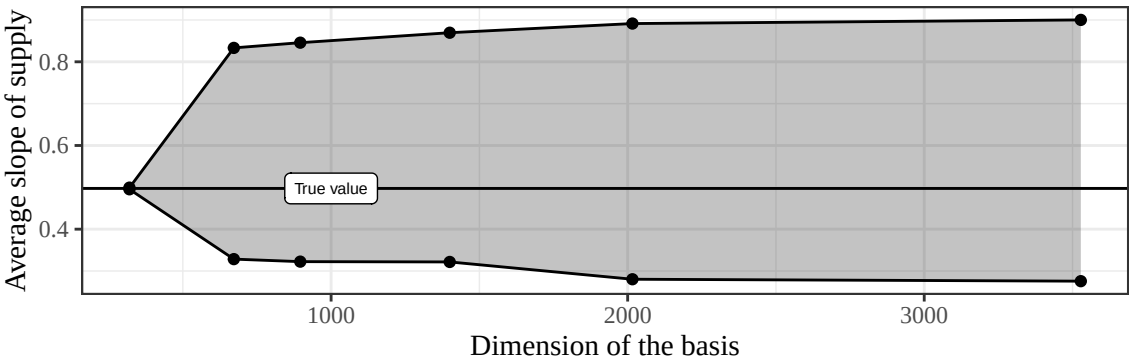


of random coefficients.

Figure 5 reports analogous bounds for the average supply slope. Perhaps surprisingly, the bounds are somewhat informative, even though there is no demand shifter. As in Figure 4, the bounds increase with the flexibility of the model but then stabilize, suggesting that the information on the average slope of supply is not produced by the functional form of the distribution of random coefficients. In Figure 6, we report bounds on the averages of the demand and supply coefficients while changing the strength of the excluded supply shifters. Both bounds widen considerably as the shifters becomes weaker, but only the bounds on the average demand coefficient continue to shrink as the shifters become stronger. Taken together, Figures 5 and 6 suggest that the excluded supply shifter does in fact carry some information about the supply curve. However, consistent with classical reasoning, even a strong supply shifter is insufficient to point identify a natural feature the slope of supply.

Figure 7 reports bounds on the deadweight loss generated by a counterfactual sales tax increase that scales demand side prices by $(1+t)$. (As mentioned in Section 2.3, the deadweight loss interpretation is premised on the assumption that there are no income effects, so that the demand curve is Marshallian (uncompensated).) Low tax changes cannot result in equilibria much different than the ones in the observed data, so result in narrow bounds. For larger tax changes, the randomness in the supply and demand slopes leads to greater uncertainty about the new equilibrium and the resulting size of the Harberger triangle. Overall, the relative deadweight loss is small in this market because supply is quite inelastic and demand is not overly inelastic. An interesting aspect about Figure 7 is that we are able to obtain quite informative bounds even though these calculations require some knowledge of the supply curve and we have no

Figure 5: *Bounds on the average slope of supply*



demand shifters. Intuitively, this is because deadweight loss is the product of both demand and supply. This leads to tight bounds in our situation because the demand side bounds are quite informative, while the supply side bounds are not overly wide.

5 Estimation

In this section, we use the identification analysis in the previous section to develop an estimator. We base estimation on a criterion function that measures the extent to which the distribution of price and quantity implied by a given distribution F matches the observed distribution. Let $Y(p,q) \equiv \mathbb{1}[P \leq p, Q \leq q]$, so that $g(p,q,z) \equiv \mathbb{E}[Y(p,q)|Z = z]$, and let

$$c(p,q;F) \equiv \mathbb{E}\left[\left(Y(p,q) - g(p,q,Z;F)\right)^2\right], \tag{23}$$

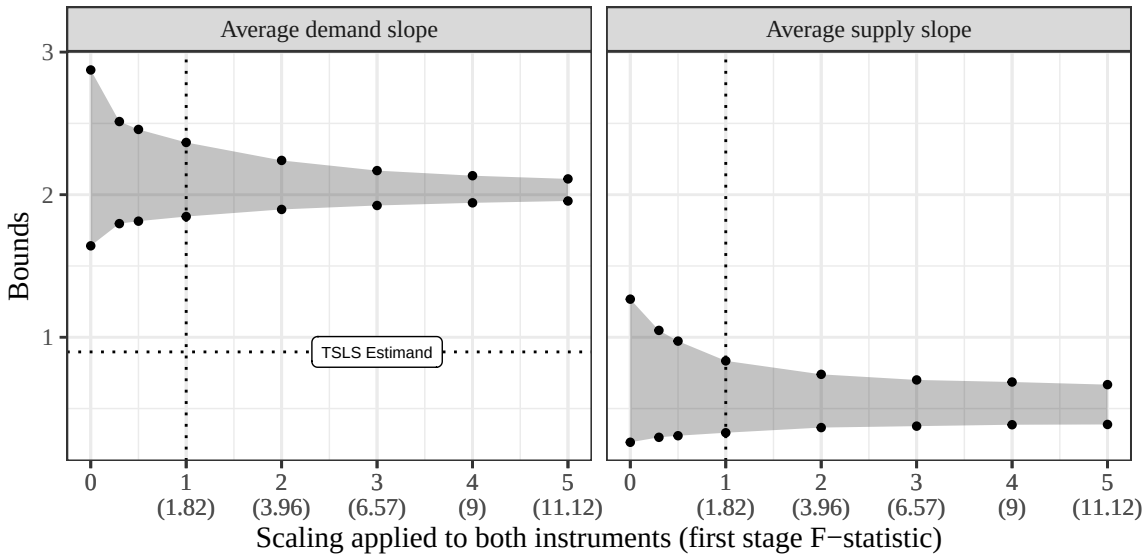
where the expectation is taken over the joint distribution of (P,Q,Z) for a fixed (p,q) evaluation pair. Then

$$c(p,q;F) \geq \mathbb{E}\left[\left(Y(p,q) - g(p,q,Z)\right)^2\right] \text{ for any } F, \tag{24}$$

by standard least squares arguments, with equality obtained if and only if $g(p,q,z;F) = g(p,q,z)$ for almost every z. The population criterion function is an aggregation over (p,q) pairs:

$$c(f) \equiv \int c(p,q;F)\, dG(p,q),$$

Figure 6: *Impact of instrument strength on bounds of average slopes*



where G is the unconditional population distribution of (P, Q). Then F minimizes c if and only if $g(p, q, z; F) = g(p, q, z)$ for every (p, q) and almost every z. The criterion function therefore provides an alternative characterization of the identified set:

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

For computation, we continue to restrict the set of distributions $\mathcal{F}$ to have the form (15), so that F is parameterized by (α, ϕ), and the criterion function can be written as

$$c(\alpha, \phi) = \int \mathbb{E}\left[(Y(p, q) - \phi' \bar{g}(p, q, Z; \alpha))^2\right] \, dG(p, q),$$

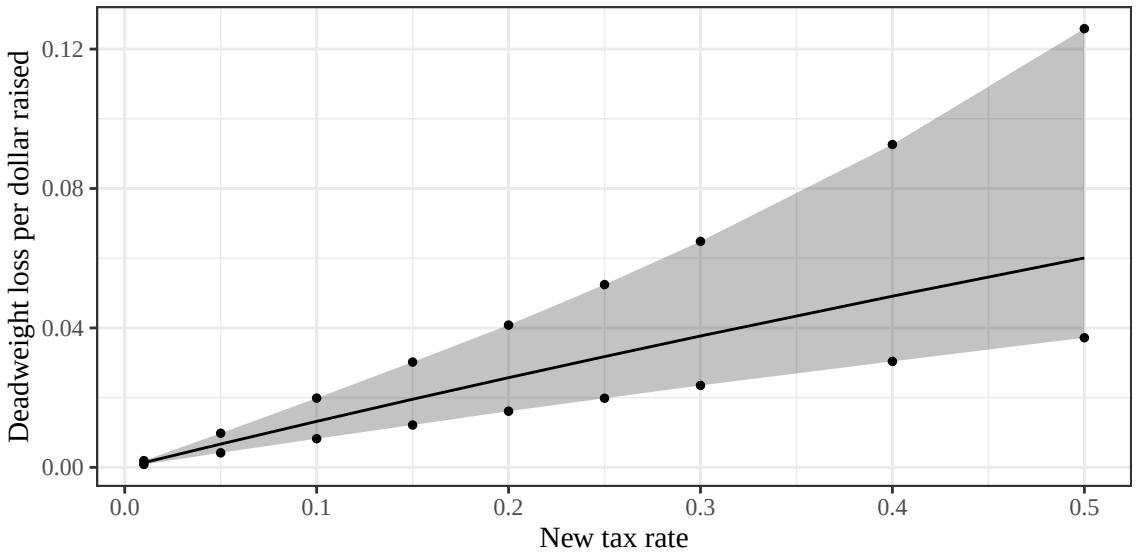
where ϕ and $\bar{g}(p, q, z; \alpha)$ are d_ϕ-dimensional vectors containing ϕ_k and $\bar{g}_k(p, q, z; \alpha)$ for $k = 1, \ldots, d_\phi$. Given data $\{P_i, Q_i, Z_i\}_{i=1}^n$ from a sample of n markets, we estimate c using sample analogs that replace G by the empirical distribution of (P, Q) and the inner expectation in the definition of c by the joint distribution of (P, Q, Z):

$$c_n(\alpha, \phi) \equiv \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left(Y_j(p_i, q_i) - \phi' \bar{g}(p_i, q_i, z_j; \alpha)\right)^2, \tag{25}$$

where $Y_j(p, q) \equiv \mathbb{1}[P_i \leq p, Q_i \leq q]$.

We use c_n to construct penalized estimators of the sharp bounds on the identified

Figure 7: *Bounds on deadweight loss from imposing a sales tax*



set $\mathcal{T}^\star$. The lower bound estimator is

$$\tau_{\text{lb},n} \equiv \min_{\alpha\in\mathcal{A},\phi\in\Phi} \sum_{k=1}^{d_\phi} \phi_k \bar{\tau}_{k,n} + \lambda_n \left(c_n(\alpha,\phi) - c_n^\star\right), \quad \text{where} \quad c_n^\star \equiv \min_{\alpha\in\mathcal{A},\phi\in\Phi} c_n(\alpha,\phi), \quad (26)$$

$\bar{\tau}_{k,n}$ are estimators of $\bar{\tau}_k$ (if it needs to be estimated), and λ_n is a tuning parameter that diverges with the sample size. The upper bound estimator is the optimal value of a maximization problem that penalizes in the opposite direction:

$$\tau_{\text{ub},n} \equiv \max_{\alpha\in\mathcal{A},\phi\in\Phi} \sum_{k=1}^{d_\phi} \phi_k \bar{\tau}_{k,n} - \lambda_n \left(c_n(\alpha,\phi) - c_n^\star\right). \quad (27)$$

6 Application

In this section, we use our approach to estimate the welfare impacts of sales taxes.

6.1 Sales Taxes in the United States

In the United States, ad valorem sales taxes are imposed on most goods and some services in 45 states and the District of Columbia. They are the second largest source of tax revenue for state and local governments. In 2021, aggregate state revenue from sales taxes was around $370 billion with $107 additional billion raised by local governments

according to figures from the Annual Survey of State and Local Government Finances.

There is considerable spatial variation in sales tax rates. Multiple levels of jurisdiction impose sales taxes: states, counties, cities, and even smaller local authorities such as special districts and metropolitan transit authorities. The resulting population-weighted average total tax rate is 7.2pc (Gaarder and Henry de Frahan, 2024b). State-level sales taxes account for most of this total rate. The median state (weighted by population) imposes a 6pc sales tax, while the median city does not impose a sales tax.

There is also considerable spatial variation in the definition of the tax base. The share of consumption subject to sales taxes varies between 0.20 in California and 0.59 in Hawaii. The median share across states is 0.33 (Gaarder and Henry de Frahan, 2024a). Historically, sales taxes applied to goods and exempted services. However, the definition of the tax base has expanded in most states to include at least some services. Within goods, an approximate rule of thumb is that food products are tax-exempt while non-food products are taxable. However, there are many exceptions to this rule in state and local jurisdictions.

6.2 Data

We use Nielsen scanner data to obtain measures of quantities and before-tax prices. The scanner data includes weekly sales, price, and volume for millions of uniquely defined food and non-food products at the UPC level for a large number of stores spanning 48 U.S. states. We obtain data on sales tax rates by county from the Thomson Reuters OneSource Sales Tax database (Gaarder and Henry de Frahan, 2024b). From this database we are able to measure the average statutory tax rate in each county over the period 2008 to 2014, along with the taxability status of each product, as collected by (Gaarder and Henry de Frahan, 2024b). Combined, the data provides spatial variation on both sales tax rates and taxability status by product.

To reduce the dimension of the estimation sample, we aggregate UPCs into larger categories called product modules. We restrict our attention to product modules that are in the fourth quartile of the distribution of total yearly sales in 2008 for both food and non-food products. A price and quantity index are calculated for each module-store combination in six months time periods. The estimation sample is balanced and includes only pairs of stores and modules that appear in the data at least once every period. The resulting data is a panel of 263 modules in 22,626 stores across 2,180 counties observed in every six month period between January 2008 and December 2014. See Gaarder and Henry de Frahan (2024b) for more details on data construction, price index calculations, and sample definition.

For estimation, we perform an additional adjustment to control for aggregate time effects. First, for each module-store combination, we calculate the two-year difference in the log of quantity and price, and tax rate. Then, we residualize by substracting the mean two-year change within each combination of a module, Census region, and time period.

6.3 Summary statistics

The results are pending review by Thompson Reuters, so are not included in this draft.

6.4 Estimates

The results are pending review by Thompson Reuters, so are not included in this draft.

7 Conclusion

We considered a classical linear supply and demand system modified to allow for heterogeneous supply and demand slopes. This modification is natural in the context of the modern literature on instrumental variables with heterogeneous treatment effects, and arises naturally from microfounded models that lead to linear market demand and supply, but raises substantial complications for identification. 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.

A Proofs

Proof of Proposition 1. Plugging the budget constraint $y_j = q_{0,j} + Pq_j$ into the preferences (5), we obtain the following utility maximization problem for the consumer:

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

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 first that the consumer makes 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

$$\xi_j^{\frac{1}{1-\chi}} \chi^{\frac{\chi}{1-\chi}} (1-\chi) P^{-\frac{\chi}{1-\chi}} \geq \gamma_j \tag{28}$$

The share of consumers of type ξ that consume a positive amount of the good is given by the proportion for which (28) is true:

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

where F is the distribution of γ_j, which is assumed to be independent of ξ. The aggregate market demand of the focal good is then determined by the combination of the intensive and extensive margins of demand integrated over ξ:

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

Taking logs yields the claimed expression. *Q.E.D.*

Proof of Proposition 2. The representative firm's cost of producing q of the focal good is given by

$$C(q) \equiv \left\{ \min_{K,L} w_K^0 K^{1+\mathfrak{b}_K} + w_L^0 L^{1+\mathfrak{b}_L} \quad \text{s.t.} \quad q = AK^{\mathfrak{a}\mathfrak{g}} L^{(1-\mathfrak{a})\mathfrak{g}} \right\}$$

The first-order conditions are:

$$\begin{aligned} (1+\mathfrak{b}_K) w_K^0 K^{\mathfrak{b}_K} &= \nu \mathfrak{a}\mathfrak{g} A K^{\mathfrak{a}\mathfrak{g}-1} L^{(1-\mathfrak{a})\mathfrak{g}} \\ \text{and} \quad (1+\mathfrak{b}_L) w_L^0 K^{\mathfrak{b}_L} &= \nu (1-\mathfrak{a})\mathfrak{g} A K^{\mathfrak{a}\mathfrak{g}} L^{(1-\mathfrak{a})\mathfrak{g}-1}, \end{aligned}$$

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

$$K = \left[\frac{\mathfrak{a}}{1-\mathfrak{a}} \frac{(1+\mathfrak{b}_L) w_L^0}{(1+\mathfrak{b}_K) w_K^0} \right]^{\frac{1}{1+\mathfrak{b}_K}} L^{\frac{1+\mathfrak{b}_L}{1+\mathfrak{b}_K}}. \tag{29}$$

Substituting (29) into the production function yields labor demand conditional on q:

$$L = \left[\frac{1-\mathfrak{a}}{\mathfrak{a}} \frac{(1+\mathfrak{b}_K) w_K^0}{(1+\mathfrak{b}_L) w_L^0} \right]^{\frac{\mathfrak{a}}{\mathfrak{a}(1+\mathfrak{b}_L)+(1-\mathfrak{a})(1+\mathfrak{b}_K)}} \left(\frac{q}{A} \right)^{\frac{1+\mathfrak{b}_K}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_L)+(1-\mathfrak{a})(1+\mathfrak{b}_K)\right]}}. \tag{30}$$

Substituting (30) into (29) and re-arranging then yields capital demand conditional on q:

$$K = \left[\frac{\mathfrak{a}}{1-\mathfrak{a}} \frac{(1+\mathfrak{b}_L) w_L^0}{(1+\mathfrak{b}_K) w_K^0} \right]^{\frac{1-\mathfrak{a}}{\mathfrak{a}(1+\mathfrak{b}_L)+(1-\mathfrak{a})(1+\mathfrak{b}_K)}} \left(\frac{q}{A} \right)^{\frac{1+\mathfrak{b}_L}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_L)+(1-\mathfrak{a})(1+\mathfrak{b}_K)\right]}}. \tag{31}$$

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

$$\begin{aligned} C(q) &= \tilde{C}(\mathfrak{a}, \mathfrak{g}, \mathfrak{b}_K, \mathfrak{b}_L, w_K^0, w_L^0) \times \left(\frac{q}{A} \right)^{\frac{(1+\mathfrak{b}_K)(1+\mathfrak{b}_L)}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_L)+(1-\mathfrak{a})(1+\mathfrak{b}_K)\right]}} \\ \text{where} \quad \tilde{C}(\mathfrak{a}, \mathfrak{g}, \mathfrak{b}_K, \mathfrak{b}_L, w_K^0, w_L^0) &\equiv \left((1-\mathfrak{a})(1+\mathfrak{b}_L) w_L^0\right)^{-\frac{(1-\mathfrak{a})(1+\mathfrak{b}_K)}{\mathfrak{a}(1+\mathfrak{b}_L)+(1-\mathfrak{a})(1+\mathfrak{b}_K)}} \\ &\quad \times \left((\mathfrak{a}(1+\mathfrak{b}_K) w_K^0\right)^{-\frac{\mathfrak{a}(1+\mathfrak{b}_L)}{\mathfrak{a}(1+\mathfrak{b}_L)+(1-\mathfrak{a})(1+\mathfrak{b}_K)}} \\ &\quad \times \left[\mathfrak{a}(1+\mathfrak{b}_K) w_K^0 + (1-\mathfrak{a})(1+\mathfrak{b}_L) w_L^0\right]. \end{aligned}$$

To maximize profit taking output price P as given, the firm chooses quantity Q so solve

$$Q = \arg\max_q Pq - C(q)$$

The first-order condition is:

$$P = \frac{(1+\mathfrak{b}_K)(1+\mathfrak{b}_L)}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_L)+(1-\mathfrak{a})(1+\mathfrak{b}_K)\right]}\tilde{C}(\mathfrak{a},\mathfrak{g},\mathfrak{b}_K,\mathfrak{b}_\mathfrak{L},w_K^0,w_L^0)A^{-1}\left(\frac{Q}{A}\right)^{\frac{(1+\mathfrak{b}_K)(1+\mathfrak{b}_L)}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_L)+(1-\mathfrak{a})(1+\mathfrak{b}_K)\right]}-1}.$$

Rearranging and taking logs yields the claimed expression with

$$\begin{aligned}&\iota^s(\mathfrak{a},\mathfrak{g},\mathfrak{b}_K,\mathfrak{b}_L,A,w_K^0)\\&\quad\equiv\log\left[\frac{(1+\mathfrak{b}_L)(1+\mathfrak{b}_K)}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_K)+(1-\mathfrak{a})(1+\mathfrak{b}_L)\right]}\tilde{C}(\mathfrak{a},\mathfrak{g},\mathfrak{b}_K,\mathfrak{b}_\mathfrak{L},w_K^0,w_L^0)A^{\frac{-(1+\mathfrak{b}_L)(1+\mathfrak{b}_K)}{\mathfrak{g}\left[\mathfrak{a}(1+\mathfrak{b}_K)+(1-\mathfrak{a})(1+\mathfrak{b}_L)\right]}}\right].\end{aligned}$$

Q.E.D.

Proof of Proposition 3. Since Z and B are independent and $B_{z1}^d = 0$, we get from (4) that

$$\mathbb{E}[P|Z] = Z_1'\,\mathbb{E}\left[\frac{-B_{z1}^s}{B_p^d+B_p^s}\right] + Z_2'\,\mathbb{E}\left[\frac{B_{z2}^d-B_{z2}^s}{B_p^d+B_p^s}\right] \equiv -Z_1'\delta_1 + Z_2'\delta_2.$$

Because this is linear in Z, 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 (9),

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

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

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

For the numerator, first note that from (4), the independence of B and Z, and $B_{z1}^d = 0$,

$$\mathbb{E}[Q|Z] = Z_1'\,\mathbb{E}\left[\frac{B_p^dB_{z1}^s}{B_p^d+B_p^s}\right] + Z_2'\,\mathbb{E}\left[\frac{B_p^sB_{z2}^d+B_p^dB_{z2}^s}{B_p^d+B_p^s}\right] \equiv Z_1'\eta_1 + Z_2'\eta_2,$$

where

$$\eta_1 \equiv \mathbb{E}\left[B_p^d\,\mathbb{E}\left[\frac{B_{z1}^s}{B_p^d+B_p^s}\middle|B_p^d\right]\right].$$

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

$$\mathbb{E}[(Z_1'\eta_1 + Z_2'\eta_2)\tilde{Z}_1'\delta_1] = \eta_1' \, \mathbb{E}[Z_1\tilde{Z}_1'\delta_1] = \mathbb{E}\left[B_p^d \, \mathbb{E}\left[\frac{B_{z1}^s}{B_p^d + B_p^s}\middle| B_p^d\right]' \mathbb{E}[\tilde{Z}_1\tilde{Z}_1'\delta_1]\right].$$

Combining numerator and denominator, we arrive at

$$-\beta_{\text{tsls}} = \mathbb{E}\left[B_p^d \, \mathbb{E}\left[\frac{B_{z1}^s}{B_p^d + B_p^s}\middle| B_p^d\right]' \frac{\mathbb{E}[\tilde{Z}_1\tilde{Z}_1'\delta_1]}{\mathbb{E}[(\tilde{Z}_1'\delta_1)^2]}\right] \equiv \mathbb{E}[B_p^d W],$$

which is the claimed expression. To see that $\mathbb{E}[W] = 1$, notice that

$$\mathbb{E}\left[\mathbb{E}\left[\frac{B_{z1}^s}{B_p^d + B_p^s}\middle| B_p^d\right]\right] = \mathbb{E}\left[\frac{B_{z1}^s}{B_p^d + B_p^s}\right] \equiv \delta_1.$$

Q.E.D.

Proof of Proposition 4. (a) If Z_1 is scalar, then δ_1 is also scalar, so that

$$\frac{\mathbb{E}[\tilde{Z}_1\tilde{Z}_1'\delta_1]}{\mathbb{E}[(\tilde{Z}_1'\delta_1)^2]} = \delta_1^{-1} \equiv \frac{1}{\mathbb{E}[B_{z1}^s/(B_p^d + B_p^d)]}.$$

It follows that

$$W = \frac{\mathbb{E}\left[B_{z1}^s/(B_p^d + B_p^s)|B_p^d\right]}{\mathbb{E}[B_{z1}^s/(B_p^d + B_p^d)]}.$$

Because $B_p^d + B_p^s$ is always non-negative, W is also non-negative if B_{z1}^s only takes one sign, which was the claim.

(b) If B_{z1}^s is independent of (B_p^d, B_p^s), then $\delta_1 = \mathbb{E}[B_{z1}^s]\,\mathbb{E}[1/(B_p^d + B_p^s)]$, so that

$$\mathbb{E}\left[\frac{B_{z1}^s}{B_p^d + B_p^s}\middle| B_p^d\right] = \mathbb{E}[B_{z1}^s]\,\mathbb{E}\left[\frac{1}{B_p^d + B_p^s}\middle| B_p^d\right] = \delta_1 \frac{\mathbb{E}\left[1/(B_p^d + B_p^s)|B_p^d\right]}{\mathbb{E}\left[1/(B_p^d + B_p^s)\right]}$$

It follows that

$$W = \mathbb{E}\left[\frac{B_{z1}^s}{B_p^d + B_p^s}\middle| B_p^d\right]' \frac{\mathbb{E}[\tilde{Z}_1\tilde{Z}_1'\delta_1]}{\mathbb{E}[(\tilde{Z}_1'\delta_1)^2]} = \frac{\mathbb{E}[1/(B_p^d + B_p^s)|B_p^d]}{\mathbb{E}[1/(B_p^d + B_p^s)]}, \tag{32}$$

which is always positive because $B_p^d + B_p^s$ is always positive. *Q.E.D.*

Proof of Proposition 5. Equation (10) follows under the assumptions of Proposi-

tion 3 because

$$\mathbb{E}[B_p^d W] = \mathbb{E}[B_p^d]\,\mathbb{E}[W] + \mathbb{C}[B_p^d, W] = \mathbb{E}[B_p^d] + \mathbb{C}[B_p^d, W],$$

noting that $\mathbb{E}[W] = 1$.

Now suppose that condition (b) of Proposition 4 is satisfied. Then (32) in the proof of Proposition 4 is satisfied. So

$$\mathbb{C}[B_p^d, W] = \mathbb{C}\left[B_p^d, \mathbb{E}\left[\frac{1}{B_p^d + B_p^s}\middle| B_p^d\right]\right] \mathbb{E}\left[\frac{1}{B_p^d + B_p^s}\right]^{-1}.$$

Because B_p^d, B_p^s are assumed to be non-negative, the sign of this term is determined by the covariance, which simplifies into

$$\begin{aligned}
\mathbb{C}\left[B_p^d, \mathbb{E}\left[\frac{1}{B_p^d + B_p^s}\middle| B_p^d\right]\right] &= \mathbb{E}\left[\left(B_p^d - \mathbb{E}[B_p^d]\right)\mathbb{E}\left[\frac{1}{B_p^d + B_p^s}\middle| B_p^d\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(B_p^d - \mathbb{E}[B_p^d]\right)\frac{1}{B_p^d + B_p^s}\middle| B_p^d\right]\right] \\
&= \mathbb{E}\left[\left(B_p^d - \mathbb{E}[B_p^d]\right)\frac{1}{B_p^d + B_p^s}\right] = \mathbb{C}\left[B_p^d, \frac{1}{B_p^d + B_p^s}\right].
\end{aligned}$$

Together with (10), this shows that (11) is sufficient and necessary for $-\beta_{\text{tsls}} \leq \mathbb{E}[B_p^d]$.

If $\mathbb{E}[(B_p^d + B_p^s)^{-1} \mid B_p^d = b_p^d]$ is a weakly decreasing function of b_p^d, then

$$\mathbb{C}\left[B_p^d, \frac{1}{B_p^d + B_p^s}\right] = -\,\mathbb{C}\left[B_p^d, -\,\mathbb{E}\left[\frac{1}{B_p^d + B_p^s}\middle| B_p^d\right]\right] \leq 0,$$

because the covariance of two weakly increasing functions of B_p^d is non-negative (e.g. Thorisson, 1995, Section 2). Alternatively, if B_p^d is mean independent of B_p^s, then

$$\begin{aligned}
\mathbb{C}\left[B_p^d, \frac{1}{B_p^d + B_p^s}\right] &= \mathbb{E}\left[(B_p^d - \mathbb{E}[B_p^d])\frac{1}{B_p^d + B_p^s}\right] \\
&= \mathbb{E}\left[(B_p^d - \mathbb{E}[B_p^d])\left(\frac{1}{B_p^d + B_p^s} - \frac{1}{\mathbb{E}[B_p^d] + B_p^s}\right)\right], \qquad (33)
\end{aligned}$$

where the second equality follows because

$$\mathbb{E}\left[(B_p^d - \mathbb{E}[B_p^d])\frac{1}{\mathbb{E}[B_p^d] + B_p^s}\right] = \mathbb{E}\left[(\mathbb{E}[B_p^d|B_p^s] - \mathbb{E}[B_p^d])\frac{1}{\mathbb{E}[B_p^d] + B_p^s}\right] = 0.$$

Simplifying (33) then shows that

$$\mathbb{C}\left[B_p^d, \frac{1}{B_p^d + B_p^s}\right] = \mathbb{E}\left[(B_p^d - \mathbb{E}[B_p^d])\left(\frac{\mathbb{E}[B_p^d] - B_p^d}{(B_p^d + B_p^s)(\mathbb{E}[B_p^d + B_p^s])}\right)\right] \leq 0.$$

Q.E.D.

B Derivation of Welfare Target Parameters

We focus on the case where Q and P are specified in logs and $h^d(p) = h^s(p) = p$ are linear, because this is what we use in our application. Similar expressions can be derived when Q and P are specified in levels and when h^d and/or h^s are nonlinear. For notation, we let $q \equiv \exp(Q)$ and $p \equiv \exp(P)$ be equilibrium quantity and price in levels.

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

$$p(\theta) = \exp\left(Z'\left(\frac{B_z^d + B_z^s}{B_p^d + B_p^s}\right) - \frac{B_p^d}{B_p^d + B_p^s}\theta\right) \tag{34}$$

$$q(\theta) = \exp\left(Z'\left(\frac{B_p^s B_z^d + B_p^d B_z^s}{B_p^d + B_p^s}\right) - \frac{B_p^d B_p^s}{B_p^d + B_p^s}\theta\right) \tag{35}$$

Equilibrium sales as a function of θ are therefore:

$$\text{SAL}(\theta) \equiv p(\theta)\, q(\theta) = \exp\left(Z'\left(\frac{B_z^d\left(1 + B_p^s\right) - B_z^s\left(1 - B_p^d\right)}{B_p^d + B_p^s}\right) - \frac{B_p^d\left(1 + B_p^s\right)}{B_p^d + B_p^s}\theta\right). \tag{36}$$

And government revenue is:

$$\text{REV}(\theta) = t \times \text{SAL}(\theta) \equiv (\exp(\theta) - 1)\text{SAL}(\theta). \tag{37}$$

The difference in consumer and producer surplus from a tax of t relative to a state with no taxes is given by

$$\text{CS}(\theta) \equiv \int_{(1+t)p(\theta)}^{p(0)} \exp(Z' B_z^d - \log(x)' B_p^d)\, dx \tag{38}$$

$$\text{PS}(\theta) \equiv \int_{p(0)}^{p(\theta)} \exp(Z' B_z^s + \log(x)' B_p^s)\, dx. \tag{39}$$

Then deadweight loss from the tax is given by

$$\textsc{dwl}(\theta) = -\left(\textsc{cs}(\theta) + \textsc{ps}(\theta) + \textsc{rev}(\theta)\right). \tag{40}$$

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

$$\begin{aligned}
\textsc{cs}'(\theta) &\equiv -(1+t)\exp(\textsc{sal}(\theta))\frac{B_p^s}{B_p^d + B_p^s} \\
\textsc{ps}'(\theta) &\equiv -\exp(\textsc{sal}(\theta))\frac{B_p^d}{B_p^d + B_p^s} \\
\textsc{rev}'(\theta) &\equiv \exp(\textsc{sal}(\theta))\left(1 + t\frac{B_p^s\left(1 - B_p^d\right)}{B_p^d + B_p^s}\right).
\end{aligned}$$

The corresponding change in deadweight loss is then

$$\textsc{dwl}'(\theta) \equiv -\left(\textsc{cs}'(\theta) + \textsc{ps}'(\theta) + \textsc{rev}'(\theta)\right) = t\exp(\textsc{sal}(\theta))\frac{B_p^d B_p^s}{B_p^d + B_p^s}.$$

Usually, we normalize the change in deadweight loss by the corresponding change in revenue and consider the quantity

$$\overline{\textsc{dwl}}'(\theta) \equiv \frac{\textsc{dwl}'(\theta)}{\textsc{rev}'(\theta)} = \frac{tB_p^d B_p^s}{B_p^d + B_p^s + t\left(1 - B_p^d\right)B_p^s}. \tag{41}$$

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

$$\textsc{inc}(\theta) \equiv \frac{\textsc{cs}'(\theta)}{\textsc{cs}'(\theta) + \textsc{ps}'(\theta)} = \frac{(1+t)B_p^s}{B_p^d + (1+t)B_p^s}. \tag{42}$$

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

$$\textsc{dwl}(\theta_0 \to \theta_1) \equiv \textsc{dwl}(\theta_1) - \textsc{dwl}(\theta_0) = \int_{\theta_0}^{\theta_1} \textsc{dwl}'(\theta)\, d\theta. \tag{43}$$

Normalizing this against the change in government revenue gives

$$\overline{\textsc{dwl}}(\theta_0 \to \theta_1) \equiv \frac{\textsc{dwl}(\theta_1) - \textsc{dwl}(\theta_0)}{\textsc{rev}(\theta_1) - \textsc{rev}(\theta_0)} = \frac{\int_{\theta_0}^{\theta_1} \textsc{dwl}'(\theta)\, d\theta}{\int_{\theta_0}^{\theta_1} \textsc{rev}'(\theta)\, d\theta}, \tag{44}$$

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.

References

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. 2, 4, 9, 10, 11, 13, 19

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press. 10

Beran, R. and P. Hall (1992): "Estimating Coefficient Distributions in Random Coefficient Regressions," *The Annals of Statistics*, 20, 1970–1984. 3

Berry, S. T. and P. A. Haile (2018): "Identification of Nonparametric Simultaneous Equations Models With a Residual Index Structure," *Econometrica*, 86, 289–315. 3

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022): "When Is TSLS Actually LATE?" Tech. Rep. w29709, National Bureau of Economic Research, Cambridge, MA. 10

Blundell, R. and R. L. Matzkin (2014): "Control Functions in Nonseparable Simultaneous Equations Models," *Quantitative Economics*, 5, 271–295. 3

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. 15

Chetty, R. (2009): "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Review of Economics*, 1, 451–488. 8

Dubé, J.-P. (2019): "Microeconometric Models of Consumer Demand," in *Handbook of the Economics of Marketing*, Elsevier, vol. 1, 1–68. 27

Gaarder, I. and L. Henry de Frahan (2024a): "The Distributional Incidence of Sales Taxes in the US," Working Paper, University of Chicago. 25

——— (2024b): "The Welfare Effect of Non-Marginal Changes in Sales Taxes," Working Paper, University of Chicago. 25

Gaillac, C. and E. Gautier (2022): "Adaptive Estimation in the Linear Random Coefficients Model When Regressors Have Limited Variation," *Bernoulli*, 28, 504–524. 3

Graddy, K. (1995): "Testing for Imperfect Competition at the Fulton Fish Market," *The RAND Journal of Economics*, 26, 75–92. 19

Hahn, J. (2001): "Consistent Estimation of the Random Structural Coefficient Distribution from the Linear Simultaneous Equations System," *Economics Letters*, 73, 227–231. 3

Harberger, A. C. (1964): "Front Matter," *The American Economic Review*, 54. 8

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. 3, 12

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. 12

Hermann, P. and H. Holzmann (2024): "BOUNDED SUPPORT IN LINEAR RANDOM COEFFICIENT MODELS: IDENTIFICATION AND VARIABLE SELECTION," *Econometric Theory*, 1–30. 3

Hoderlein, S., H. Holzmann, and A. Meister (2017): "The Triangular Model with Random Coefficients," *Journal of Econometrics*, 201, 144–169. 2, 3

Hoderlein, S., J. KlemelÃ¤, and E. Mammen (2010): "ANALYZING THE RANDOM COEFFICIENT MODEL NONPARAMETRICALLY," *Econometric Theory*, 26, 804–837. 3

Hurwicz, L. (1950): "Systems with Nonadditive Disturbances," in *Cowles 10*, ed. by T. Koopmans, no. 10 in Cowles Commission Monographs, 410–418. 3

Kelejian, H. H. (1974): "Random Parameters in a Simultaneous Equation Framework: Identification and Estimation," *Econometrica*, 42, 517–527. 3

Koopmans, T. C., W. C. Hood, et al. (1953): "The Estimation of Simultaneous Linear Economic Relationships," *Studies in Econometric Method, Cowles Commission Monograph*, 14, 112–199. 3

Leamer, E. E. (1981): "Sets of Estimates of Location," *Econometrica*, 49, 193–204. 3

Lewbel, A. and K. Pendakur (2017): "Unobserved Preference Heterogeneity in Demand Using Generalized Random Coefficients," *Journal of Political Economy*, 125, 1100–1148. 3

Manski, C. F. (1995): *Identification Problems in the Social Sciences*, Harvard University Press. 3, 4, 9

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334. 3, 4

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. 4

Masten, M. A. (2018): "Random Coefficients on Endogenous Variables in Simultaneous Equations Models," *The Review of Economic Studies*, 85, 1193–1250. 2, 3, 5, 8, 14, 15, 20, 26

Masten, M. A. and A. Torgovitsky (2016): "Identification of Instrumental Variable Correlated Random Coefficients Models," *Review of Economics and Statistics*, 98, 1001–1005. 3

Matzkin, R. L. (2008): "Identification in Nonparametric Simultaneous Equations Models," *Econometrica*, 76, 945–978. 3

——— (2015): "Estimation of Nonparametric Models With Simultaneity," *Econometrica*, 83, 1–66. 3

Mogstad, M. and A. Torgovitsky (2024): "Instrumental Variables with Unobserved Heterogeneity in Treatment Effects," Tech. Rep. w32927, National Bureau of Economic Research, Cambridge, MA. 10

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. 10

Thorisson, H. (1995): "Coupling Methods in Probability Theory," *Scandinavian Journal of Statistics*, 22, 159–182. 31