Prices or Quantities?*

Mauricio Varela
Department of Economics, University of Arizona
1130 E Helen St, Tucson, AZ 85721
E-mail: mvarela@email.arizona.edu

Madhu Viswanathan
Department of Marketing, University of Arizona
1130 E Helen St, Tucson, AZ 85721
E-mail: madhu@email.arizona.edu

January 15, 2019

Running Head: Prices or Quantities?

*We thank XXX and XXX for useful comments. All errors are our own.

Abstract

When empirically analyzing industries, it is often unknown if firms set prices and sell all demanded amounts-i.e. ‘a la Bertrand- or if they set quantities and prices are adjusted to clear the market -i.e. ‘a la Cournot’. We extend the BLP model of demand and supply estimation to also estimate the form of competition, in prices or in quantities. The extension is easily implementable and does not add significant computational burden to the BLP model. The paper then estimates the type of competition in a variety of industries, including commericial aviation, milk retail, and auto retail.

Keywords: random coefficients, strategic complements, strategic substitutes

JEL Classification: D??

1. Introduction

In a partial equilibrium setting, evaluating policies requires understanding both consumers and suppliers. Much work has been done at understanding consumers [cite], much less so at understanding suppliers. In their seminal paper, Berry et al. (1995) (henceforth BLP) propose a method to evaluate suppliers incentives in determining prices based on the premise that prices are set in a Nash-Bertrand game. This method has become the basis of much applied work, used in studying merger effects, tax changes, and firm product repositioning, among others. [[insert citations]] However, the assumption of Nash-Bertrand is rarely justified in empirical work and may be unrealistic in many applications. In this paper we build on BLP and allow for prices to arise according to either a Nash-Bertrand equilibrium or a differentiated Cournot equilibrium, and estimate which of the two is most appropriate for the industry at hand. Our method imposes a minimal additional burden, in terms of available data, research design, and computing power, over that required by the BLP method, and therefore should be easy for applied researchers to implement.

The key difference between Nash-Bertrand equilibria and differentiated Cournot equilibria is that firms' strategies are strategic complementes in the former, and strategic substitutes (cf. Bulow et al. (1985)) in the latter. Therefore, a given shock (e.g cost increase, tax increase, pricing-pressure decrease, etc.) affecting firms results in a much larger price shock under Nash-Bertrand than under differentiated-Cournot when firms compete in monopolistic competition. For example, suppose firm A's cost suddenly increases. In a Nash-Bertrand game, firm A reacts by increasing price. This increase in price incentivizes rival firms' to increase their own prices, which incentivizes firm A to increase price further. In contrast, the same cost shock in a differentiated Cournot game entices firm A to decrease output, therefore increasing its price. Rivals' react by increasing their own output, resulting in a decrease in firm A's price and mitigating firm A's initial price increase. Hence, in analyzing a policy effects, it is critically important to correctly infer how prices are being generated.

In short, our empirical strategy replicates the demand estimation framework in BLP and uses estimated demand functions to calculate mark-ups under both a Nash-Betrand equilibrium and under a differentiated-Cournot equilibrium. Then, in the spirit of the conduct parameter literature (cf. Bresnahan (1989); Corts (1999)), we insert both mark-up terms into a single supply equation with each mark-up term interacted with a model selection parameter. Thus, we are nesting both models, i.e. Nash-Bertrand pricing and differentiated-Cournot, into a single estimating equation. Finally, we propose to estimate such equation using a GMM framework, using rivals' cost shifters as an instrument for the model selection parameter, and stacking the supply moments with the demand estimation moments for efficiency. Therefore, the only incremental burden, over and beyond that in BLP, is in calculating mark-ups under differentiated-Cournot; mark-ups which are as easily calculated as those from a Nash-Bertrand game.

To see why rivals' cost-shifers are an effective instrument at identifying the form of competition, note how firms' own cost-passthroughs identify firms' costs. Therefore, conditional on knowing firms' costs, rivals' cost-passthrough identifies the type of competition. If the rivals' cost passthrough is large, firms ought to be competiing in strategic complementes (i.e. Nash-Bertrand); if it is moderate, firms ought to be competing in strategic substitutes (i.e. differentiated-Cournot). [[NEED TO DEVELOP BETTER]]

The paper has three sections. In the first section we introduce competition with differentiated-Cournot under the BLP framework and formalize the emprical strategy discussed above. Afterwards, in the second section, we use Monte Carlo simulations to illustrate the importance of correctly specifying the type of competition and to illustrate how rivals' cost-shifters indeed serve as an effective instrument in identifying the form of competition. In a third section we apply our empirical framework to various industries in which their is no consensus on the type of competition, i.e. in prices or in quantities. Specifically, we study the US domestic airline industry, the US diary retail industry, and the US beer industry. We find XXX and YYY.

Nash-Bertrand has often associated with short-run pricing strategies, [[citations]] while Cournot competition with long-run strategies. [[citations]]. To the extent that policies are usually being evaluated on their medium to long-run effects, having a model that allows for prices to arise under

Cournot competition seems critical to these policy evaluations. It is important to caveat that both forms of competition abstract from important drivers of pricing, including product repositioning[1] and dynamic considerations[2]. Therefore, in allowing for differentiated-Cournot, we do not expect to have the ideal model of price formation, only one that improves on the existing benchmark model (i.e. BLP) without adding a significant burden to the applied researcher.

[[Lit review?]]

2. *Competition in Prices vs Quantities: Why it Matters*

In many empirical applications, firms' costs are inferred by assuming firms price optimally and backing-out costs from necessary conditions for equilibrium pricing, e.g. [insert literature that uses BLP demand & supply, innequality based approaches, and dynamic games]. Here we extend this approach and allow for an alternative method for inferring costs: assume firms simulteanously choose output allocations -i.e. quantities-, given these choices prices clear the market, and costs are inferred from firms' optimaliy conditions of their equilibrium choices. In what follows we formalize these two methods for inferring costs and show how the inferred costs differ across the two methods. Section 3 then proposes a method to estimate which of the two methods is a more accurate representation of the industry and Section 4 illustrates the effectiveness of the estimating method using simulated data.

Let us briefly introduce relevant notation. We assume there are I firms in a market, indexed $i = 1..I$. Each firm offers J_i products, with j indexing one such product and $\mathscr{J}$ being the total number of products across all firms. Let $J(j)$ be the set of products offered by the firm who also offers product j. The demand for product j is given by $Q_j(p)$, where $p \equiv \left(p_1, \ldots, p_{\mathscr{J}}\right)$. We assume the demand vector, $Q(p) \equiv (Q_1(p), \ldots, Q_J(p))$, is twice continuously differentiable in prices and has a well defined inverse $P(q)$.[3] Finally, we assume the costs of supplying product j are characterized by

[1] Cf. Sweeting et al. (2018), Seim et al. (ming), and Gandhi et al. (2008)

[2] Cf. Ericson and Pakes (1995) and subsequent works

[3] A sufficient condition for such inverse to exist is that the Jacobian of $Q(\cdot)$ have a dominant diagonal for all relevant prices: $\left|\frac{\partial Q_j}{\partial p_j}\right| > \sum_{j' \neq j} \left|\frac{\partial Q_j}{\partial p_{j'}}\right| \quad \forall j = 1..N$.

constant returns to scale with a per-unit cost given by c_j.

We contrast two different supply models. The first model, labelled 'Price', assumes firms simulteanously choose prices. The second model, labelled 'Quantity', assumes firms simulteanously choose quantities.[4] For both models we only consider pure strategy equilibria, give sufficient conditions under which such equilibria exists, and characterize equilibrium outcomes using firms' first-order optimality conditions.

For each of the two models, firms' profits are given as

$$\text{Price Model} \pi_i(p) = \sum_{j \in J_i} \left(p_j - c_j\right) Q_j(p) \tag{1}$$

$$\text{Quantity Model} \pi_i(q) = \sum_{j \in J_i} \left(P_j(q) - c_j\right) q_j \tag{2}$$

And firms' FOCs are

$$\text{Price Model} 0 = \sum_{j \in J(l)} \left(p_j - c_j\right) \cdot \frac{\partial Q_j(p)}{\partial p_l} + Q_l(p) \qquad \forall l = 1..\mathscr{J} \tag{3}$$

$$\text{Quantity Model} 0 = \sum_{j \in J(l)} \frac{\partial P_j(q)}{\partial q_l} \cdot q_j + P_l(q) - c_l \qquad \forall l = 1..\mathscr{J} \tag{4}$$

These equations can be expressed in more compact vector notation: let c, p, and q, denote the $\mathscr{J}$-by-1 vectors of variable costs, equilibrium prices, and equilibrium quantities. Also, let dQ (and dP) denote the $\mathscr{J}$-by-$\mathscr{J}$ symmetric[5] matrix of demand derivatives (inverse demand derivatives) evaluated at equilibrium prices (quantities) and let A be an $\mathscr{J}$-by-$\mathscr{J}$ 'ownership' matrix, a symmetric matrix of ones and zeros with $a_{lj} = 1$ if $j \in J(l)$. Finally, denote with '$\circ$' the Hadamard

[4]We do not contrast a third class of models in which a subset of firms choose prices while the remaining firms choose quantities. Although technically feasible, these games are out of the scope of the current paper.

[5]The matrices must be symmetric if $Q(\cdot)$ is derived from utility maximization as dQ is simply Slutsky's substitution matrix, a symmetric matrix. dP is symmetric as symmetry is preserved under matrix inversion.

product operator, i.e. elementwise multiplication. The firms' FOCs are

$$\text{Price Model} 0 = (A \circ dQ)(p - c) + Q \tag{5}$$

$$\text{Quantity Model} 0 = (A \circ dP) q + P - c \tag{6}$$

These equations indeed characterize pure strategy equilibria if such equilibria exists. Sufficient conditions for such equilibria to exist is that demand be downward sloping, invertible, and each products' demand be weakly-concave in the owning firms' own prices.[6] Equilibrium may also exist under weaker conditions, and in the application below we show existance for a particular Logit demand function that is not everywhere concave in prices.

Using equations 5 and 6, costs are inferred by isolating 'c' as a function of price and a mark-up term:[7]

$$\text{Price Model} c = p + (A \circ dQ)^{-1} Q \tag{7}$$

$$\text{Quantity Model} c = P + (A \circ dP) q \tag{8}$$

Clearly, the above two equations are identical if and only if the markup terms are identical, i.e. $(A \circ dQ)^{-1} = A \circ dP$. By the inverse function theorem, $dP = (dQ)^{-1}$, and therefore the two equations are identical iff A is such that $a_{lj} = 1$ whenever $\frac{dQ_l}{dp_j} \neq 0$. That is, when firms are monopolists in their niche markets.

Importantly, if $\hat{q}$ are observed sales and $d\hat{Q}$ are the observed (or estimated) demand slopes, the

[6]Invertability implies demand is continous in all firms' prices. Downward sloping together with invertability implies the Jacobian of demand is negative definite. Weak concavity of demand, along with a negative definite Jacobian, implies profit functions in the 'Price Model' are concave in firms' own prices, and therefore best-response functions in this model are continuous. Invertability of demand implies there is a one-to-one mapping between prices and quantities, and therefore best-response functrions in the 'Quantity Model' are also continuous. As the strategy set, i.e. the set of potential prices, can be restricted to an arbitrarily large subset of the positive real numbers, they are non-empty and convex. Thus, Brouwer's fixed point theorem can be applied to assert existance of equilibria in both models.

[7]The inverse $(A \circ dQ)^{-1}$ is well defined as the inverse of dQ is assumed to exist -i.e. it is dP-, and invertability of dQ implies all principal components of dQ are invertible: $A \circ dQ$ is one such principal component.

difference in inferred costs between the ‘Price Model’ and the ‘Quantity Model’ is

$$c^{\text{Prc}} - c^{\text{Qty}} = \left(\left(A \circ d\hat{Q} \right)^{-1} - A \circ \left(d\hat{Q} \right)^{-1} \right) \hat{q} = -\Delta \hat{q} \tag{9}$$

where Δ is a block-diagonal negative-semidefinite matrix, with each block corresponding to each firms’ products (details in the Appendix). Therefore inferred costs under the ‘Price Model’ are likely to be higher than inferred costs under the ‘Quantity Model’. Importantly, the size of Δ is proportional to the cross-derivatives of demand, such that if cross-product substitution is small the difference in inferred costs between the two models is also small: firms are quasi-monopolists in their niche markets. It is when there is significant cross-product subsitition that imposing the ‘correct’ supply model matters.

3. Extending the BLP Framework to Estimate Supply Model

We extend the BLP estimation framework to allow for the two different supply models described above, the ‘Price Model’ and the ‘Quantity Model’, and to estimate which model is most consistent with data patterns. Model selection is done by nesting both models, akin to the conduct parameter approach literature. We also illustrate briefly how model selection based on goodness-of-fit (e.g. mean-squared-errors, likelihood ratio tests, etc.) are biased in favor of the ‘Price Model’ and therefore should not be used.

The overall strategy is to first build a parametric model of demand and to obtain from this model the Jacobian of demand with respect to prices. This Jacobian is then used to obtain two different mark-up terms, one for each of the two supply models discussed above. These mark-up terms are then nested into a single supply equation with a parameter selecting the ‘right’ mark-up. The demand model and the supply equation are then stacked into a single system of unrelated equations and estimated via GMM.

In what follows, we assume there are M distinct markets, where each market has I_m firms offering a total of $\mathscr{J}_m$ products. Firm i offers J_{im} products, with $\mathscr{J}_m \equiv \cup_{i=1}^{I_m} J_{im}$. The researcher observes products' prices, shares, demand shifters, and cost shifters, denoted p_{mj}, s_{mj}, x^{D}_{mj}, and x^{C}_{mj}, respectively, with $x^{\mathrm{D}}_{mj} \in \Re^{K_{\mathrm{D}}}$ and $x^{\mathrm{C}}_{mj} \in \Re^{K_{\mathrm{C}}}$. For notational compactness, denote with p_m and s_m the vectors of all firms' prices and shares in market m.

3.1 Demand

Demand is given as the aggregation of many consumers' discrete choice of selecting one of the available products, with one such available option being the 'no-purchase' product. Specifically, consumer r, in market m, has utility for product j given as

$$u_{mjr} = \delta_{mj} + \mu^r_{mj} + \varepsilon_{mjr} \quad ; \quad \delta_{mj} = x^{\mathrm{D}}_{mj}\beta + p_{mj}\alpha + \xi_{mj} \quad ; \quad \mu^r_{mj} = x^{\mathrm{D}}_{mj}\beta_r + p_{mj}\alpha_r \tag{10}$$

and a utility for the 'no-purchase' option given by $u_{m0r} = \varepsilon_{m0r}$. Here, we assume $\left\{\varepsilon_{mjr}\right\}_{j=0}^{\mathscr{J}_m}$ are idiosyncratic taste shocks which are iid across products, consumers, and markets, and are distributed according to a Type 1 Extreme Value distribution with scale parameter equal to one. Therefore, the consumers' utility for a specific product is composed of three parts, a mean utility common to all consumers, δ_{mj}, a random utility that is correlated across products, μ^r_{mj}, and a random utility that is uncorrelated across products, consumers, and markets. Importantly, the mean utility contains an unobserved shock, ξ_{mj}, which will become the basis on which we form estimating moments. In addition, the correlated random utility, μ^r_{mj}, is formed by interacting consumer specific tastes, β_r, with observed product characteristics, x_{mj}. We assume these product specific tastes are distributed according to some parametrized CDF $F(\beta|\theta)$.

Consumers choose the product that gives them highest utility. Given the assumptions on $\left\{\varepsilon_{mjr}\right\}_{j=0}^{\mathscr{J}_m}$, market shares of each product can be expressed as a function of mean utilities and parameters θ with the following equation:

$$s_{mj}(\delta_m, \theta) = \int \exp\left[\delta_{mj} + \mu^r_{mj}\right] / \left(1 + \sum_{l=1}^{\mathscr{J}} \exp\left[\delta_{ml} + \mu^r_{ml}\right]\right) dF(\beta|\theta) \tag{11}$$

The above equation provides a model by which market shares arise as a function of demand primitives. This suggests equating model shares (i.e. eq. 11) to observed shares, $\hat{s}_{mj}$, and inverting such equation to obtain mean utilities as a function of observed shares and parameters θ. BLP(1994) show this inversion exists and is unique for most CDFs $F(\beta|\theta)$.[8] Denote with $\delta_{mj}(\theta)$ such inverse, and note how unobserved mean utility shocks, ξ, can be expressed as a function of these inverse values and observed characteristics:

$$\xi_{mj} = \delta_{mj}(\theta) - p_{mj}\alpha - x^{\mathrm{D}}_{mj}\beta \tag{12}$$

This equation provides the basis for estimating demand parameters through GMM. Before detailing the GMM procedure, however, we introduce a supply side equation to be used in estimating supply parameters and the model selection parameter.

3.2 Supply

In the prior section we described how to infer costs from conditions characterizing equilibrium play, i.e. eqs. 7 and 8. These equations depend crutially on the slope of demand with respect to price. Given the demand model above, such Jacobian for demand in market m can be calculated as

$$ds_m(\alpha,\theta) \equiv \begin{bmatrix} \frac{\partial s_{m1}}{\partial p_1} & \cdots & \frac{\partial s_{m1}}{\partial p_J} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_{mJ}}{\partial p_1} & \cdots & \frac{\partial s_{mJ}}{\partial p_J} \end{bmatrix} \quad ; \quad \begin{array}{l} \frac{\partial}{\partial p_l} s_{mj}(\alpha,\theta) = \int (\alpha+\alpha_r)\left(s^r_{mj}\cdot 1\{j=l\} - s^r_{mj}s^r_{ml}\right) dF(\beta|\theta) \\ s^r_{mj} = \exp\left[\delta_{mj}(\theta)+\mu^r_{mj}\right] / \left(1+\sum_{l=1}^{\mathscr{J}_m}\exp\left[\delta_{ml}(\theta)+\mu^r_{ml}\right]\right) \end{array}$$

Given this Jacobian, observed shares $\hat{s}_m$, and binary matrix A_m relating products owned by the same firm, the mark-up terms in eqs. 7 and 8 can be estimated as

$$h^{\mathrm{Prc}}_m(\alpha,\theta) \equiv -\left(A_m \circ ds_m\right)^{-1}\hat{s}_m \quad ; \quad h^{\mathrm{Qty}}_m(\alpha,\theta) \equiv -\left(A_m \circ (ds_m)^{-1}\right)\hat{s}_m$$

[8] [[INSERT HERE A FOOTNOTE ON WHAT FUNCTIONS ARE ALLOWED]]

If we assume costs can be estimated with a linear function of products' characteristics, i.e. $c_{mj} = x^{\mathrm{S}}_{mj}\gamma + \varepsilon_{mj}$, and we nest eqs. 7 and 8 into a single overarching model where τ serves as a model selection parameter, an estimating equation can be formed as

$$\varepsilon_{mj} = p_{mj} - \tau h^{\mathrm{Prc}}_{mj} - (1-\tau)\, h^{\mathrm{Qty}}_{mj} - x^{\mathrm{S}}_{mj}\gamma \tag{13}$$

Two things are worth noting of the above equation. First, an estimated τ value of '1' is indicative that data patterns are more consistent with a model in which firms choose prices than with a model in which they choose quantities. In contrast, an estimated value of '0' is indicative of an opposite relationship being supported by the data. Second, the mark-up terms h^{Prc} and h^{Qty} depend on the price coefficient, α, and on the parameters governing the random coefficients, θ. Therefore, this supply equation can be used to identify these demand-specific parameters. An estimating procedure that uses both demand and supply equations simulteanously, e.g. GMM with stacked moments, will use both demand and supply variation to identify these demand related parameters.

3.3 A GMM Estimation Procedure

We briefly describe a GMM estimation procedure in which moments generated from the demand equation (eq. 12) are stacked with moments generated from the supply equation (eq. 13). To ease notation, we index an observation with t, noting that an observation is a market-product pair, and we denote with T the total number of observations.

Let their be two sets of valid instruments, one pertaining the demand residuals and one perstaining the supply residuals, z^{D}_t and z^{S}_t, such that $\mathbb{E}\left[\xi z^{\mathrm{D}}\right] = 0$ and $\mathbb{E}\left[\varepsilon z^{\mathrm{S}}\right] = 0$. The instruments need not differ across the two sets, but must satisfy exogeneity and rank conditions. In particular, recall the demand and supply equations include K_{D} and K_{S} non-price characteristics, respectively. In addition, both the demand and the supply equation are dependant on (α, θ), a vector of K_{NL} non-linear parameters, and the supply equation also includes a model selection parameter, τ. Therefore, for the rank conditions to be satisfied, each set of instruments needs to include sufficient linearlyindependent (LI) instruments for identifying parameters corresponding to non-price characteristics, plus additional LI instruments such that, between the two sets at least K_{NL} additional instruments identify the non-linear parameters, and at least one additional LI instrument in z^{S} identifying the model selection parameter. For the latter, we suggest using rivals' cost shifters as instruments, since the key difference between the two mark-up values, h^{Prc} and h^{Qty}, is that the latter is a function of rivals' pass-throughs in a way that the former is not.

Importantly, the non-linear parameters need not be instrumented for using demand side instruments, i.e. z^{D}, as long as sufficient supply side instruments are incldued. For example, the demand instruments need not include an instrument for price (i.e. for identifying α), as long as the supply instruments include an additional instrument for mark-ups over-and-beyond those instruments included for the linear cost shifters and for model selection.

Stack the instruments and the non-price characteristics into two block-diagonal matrices, Z and X, such that

$$Z \equiv \begin{bmatrix} z_1^{\mathrm{D}} & 0 \\ \vdots & \vdots \\ z_T^{\mathrm{D}} & 0 \\ 0 & z_1^{\mathrm{S}} \\ \vdots & \vdots \\ 0 & z_T^{\mathrm{S}} \end{bmatrix} \quad ; \quad X \equiv \begin{bmatrix} x_1^{\mathrm{D}} & 0 \\ \vdots & \vdots \\ x_T^{\mathrm{D}} & 0 \\ 0 & x_1^{\mathrm{S}} \\ \vdots & \vdots \\ 0 & x_T^{\mathrm{S}} \end{bmatrix} \tag{14}$$

where Z is a $2T$-by-K_Z matrix, $K_Z \geq K_{\mathrm{D}} + K_{\mathrm{S}} + K_{\mathrm{NL}} + 1$, and X is a $2T$-by-$(K_D + K_S)$. Also, separate the parameters into two groups, linear parameters $\rho \equiv (\beta, \gamma)$ and non-linear parameters $\psi \equiv (\alpha, \theta, \tau)$, and denote with p, $\delta(\theta)$, h^{Prc}, and h^{Qty} the vectors of prices, inverted mean utilities, and markups across all observations. The demand and supply residuals are stacked into a single vector, $u(\rho,\psi)$, given as:

$$u(\rho,\psi)=\left[\begin{array}{c}\delta(\theta)-p\alpha\\ p-\tau h^{\mathrm{Prc}}-(1-\tau)\,h^{\mathrm{Qty}}\end{array}\right]-X\rho \tag{15}$$

Thus, the GMM estimator $(\hat{\rho},\hat{\psi})$ is the set of parameters that minimizes $f(\rho,\psi\mid\Omega)=u'Z\Omega Z'u$, for some K_Z-by-K_Z symmetric, positive semidefinite weighting matrix Ω. In order to obtain an efficient weighting matrix, we implement a two-step procedure. In the first step we set $\Omega_Z=(Z'Z)^{-1}$ and obtain estimates $(\tilde{\rho},\tilde{\psi})$ by minimizing $f(\rho,\psi\mid\Omega_Z)$. We then use these estimates to calculate an efficient weighting matrix as the inverse of the moments' variance. In particular, we calculate the moments corresponding to each observation, $\tilde{g}_t=\left(z_t^{\mathrm{D}}\tilde{\xi}_t,z_t^{\mathrm{S}}\tilde{\varepsilon}_t\right)'$, where $\tilde{\xi}_t$ and $\tilde{\varepsilon}_t$ are the estimated residuals given $(\tilde{\rho},\tilde{\psi})$. The efficient weighting matrix is therefore $\tilde{\Omega}=\mathrm{inv}\left[\frac{1}{T}\sum_t\tilde{g}_t\tilde{g}_t'\right]$. The efficient GMM estimator $(\hat{\rho},\hat{\psi})$ is given as the minimizer of $f(\rho,\psi\mid\tilde{\Omega})$.

The asymptotic variance of $(\hat{\rho},\hat{\psi})$ is obtained in the usual way: $\mathrm{Avar}(\hat{\rho},\hat{\psi})=\left(\hat{B}'\tilde{\Omega}\hat{B}\right)^{-1}/T$ where $\hat{B}$ is the expectation of the gradient of the moments, evaluated as the estimated parameters, i.e. $\hat{B}=\frac{1}{T}\sum_n\nabla_{(\rho,\psi)}\hat{g}_t$.

Minimizing the Estimating Function

In minimizing $f(\rho,\psi\mid\Omega)$, the researcher benefits from an exact analytical solution for $\hat{\rho}$. This subsection describes how to implement the minimization in two steps: hold ψ constant and minimize $f(\rho,\psi\mid\Omega)$ over ρ first, and then minimize $f(\hat{\rho}(\psi),\psi\mid\Omega)$ over ψ.

To ease on notation, let $y(\psi)$ correspond to the non-linear values of $u(\rho,\psi)$:

$$y(\psi)\equiv\left[\delta-p\alpha\;;\;p-\tau h^{\mathrm{Prc}}-(1-\tau)\,h^{\mathrm{Qty}}\right] \tag{16}$$

such that $u(\rho,\psi)=y(\psi)-X\rho$. The function to minimize is therefore

$$f(\rho,\psi\mid\Omega)=(y(\psi)-X\rho)'Z\Omega Z'(y(\psi)-X\rho) \tag{17}$$

and it's derivative with respecto to ρ is $-2\left(y(\psi)-X\rho\right)' Z\Omega Z'X$. By setting this derivative to zero and solving for ρ we obtain:

$$\rho = \left(X'Z\Omega Z'X\right)^{-1} X'Z\Omega Z' y(\psi) \tag{18}$$

which is simply the IV estimator for a linear system of unrelated equations.

To solve for the non-linear parameters, denote with Ξ the projection matrix in eq. 18, i.e. $\Xi \equiv \left(X'Z\Omega Z'X\right)^{-1} X'Z\Omega Z'$, and substitute eq. 18 into eq. 17. Simplifying we obtain:

$$\begin{aligned} f(\hat{\rho}(\psi), \psi \mid \Omega) &= (y(\psi) - X\Xi y)' Z\Omega Z' (y - X\Xi y(\psi)) \\ &= y(\psi)' (I - X\Xi)' Z\Omega Z' (I - X\Xi)' y(\psi) \end{aligned} \tag{19}$$

Note that $(I-X\Xi)' Z\Omega Z' (I-X\Xi)'$ is a symmetric, positive semi-definite matrix, and can therefore be thought of as a weighting matrix. Hence, we simplify further by denoting this matrix as W and re-write the estimating function as $f(\psi) = y(\psi)'Wy(\psi)$. The gradient of this function is therefore $\nabla_\psi f(\psi) = 2y(\psi)'W\nabla_\psi y(\psi)$.[9] Analytical forms for $\nabla_\psi y(\psi)$ are given in the appendix, as well as some additional suggestions for restricting parameters in a favorable way.

3.4 Failure of Model Selection in the Absence of an Exclusionary Instrument

In some circumstances the researcher may find it difficult to obtain an instrument that is effective at selecting the true model. Either such instrument is unavailable or performs poorly in identifying the true model, e.g. the estimated conduct parameter either rejects both models (i.e. is estimated tightly around 1/2) or accepts both models (i.e the estimated standard errors are large). In such circumstances it may appear appealing to apply a 'goodness-of-fit' criteria for model selection, either in the form of comparing R-sqs, mean-squared-errors, or a in conducting a Likelihood ratio

[9] The matrix W is a $2T$-by-$2T$ matrix, and hence may not be feasibly calculated when T is large. A feasible alternative is to calculate the intermediate matrix $G \equiv Z - X\Xi'Z$, an T-by-K matrix, and to compute $f(\psi)$ and its derivative in the following order: $f(\psi) = \left((G'y(\psi))'\Omega\right)(G'y(\psi))$ and $\nabla_\psi f(\psi) = 2\left((G'y)'\Omega\right)\left(G'\nabla_\psi y(\psi)\right)$.

test (cf.Vuong (1989)). In this section we briefly show how all three of the above methods fail in selecting the true model without bring to bear additional identifying variation.

To illustrate how these methods fail, consider a setting in which demand has already been estimated and the researcher estimates separately the two supply models:

$$\text{Price Model} \qquad p_{im} - \hat{h}_{im}^{\text{Prc}} = x_{im}\gamma^{\text{Prc}} + \varepsilon_{im}^{\text{Prc}} \tag{20}$$

$$\text{Quantity Model} \qquad p_{im} - \hat{h}_{im}^{\text{Qty}} = x_{im}\gamma^{\text{Qty}} + \varepsilon_{im}^{\text{Qty}} \tag{21}$$

where p_{im} is the observed price of product i in market m, x_{im} is a vector of cost shifters, and $\hat{h}_{im}^{\text{Price}}$ and $\hat{h}_{im}^{\text{Qty}}$ are estimates of mark-ups derived from the previously estimated demand function $\hat{q}(p)$. These mark-ups are computed as described previously: for observed sales, prices, demand function, and ownership matrix in market m, denoted q_m, p_m, $\hat{q}_m(p)$ and A_m, mark-ups are

$$\hat{h}_m^{\text{Prc}} \equiv -\left(A_m \circ \nabla_p \hat{q}_m(p_m)\right)^{-1} q_m \quad ; \quad \hat{h}_m^{\text{Qty}} \equiv -\left(A_m \circ \nabla_p \hat{q}_m(p_m)^{-1}\right) q_m \tag{22}$$

The researcher has access to exogenous instruments z_{im} which identify the cost parameters, γ, but not the model choice. We assume the researcher estimates each model using a standard linear-IV estimator, although our conclusions are supported under alternative estimators, e.g. GMM, Maximum Likelihood with normally distributed residuals, etc. Let p, $\hat{h}^{\text{Prc}}$, $\hat{h}^{\text{Qty}}$ be T-by-1 vectors of prices and mark-ups across all observations, X be an T-by-L matrix of cost shifters, and Z an T-by-K matrix of exogenous instruments. The estimated residuals from each model, $\hat{\varepsilon}^{\text{Price}}$ and $\hat{\varepsilon}^{\text{Qty}}$, are given by $\hat{\varepsilon}^{\omega} = H\left(p - \hat{h}^{\omega}\right)$, where $\omega \in \{\text{Prc}, \text{Qty}\}$ and H is the idempotent residual projection matrix

$$H \equiv I - X'\left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1}X'Z\left(Z'Z\right)^{-1}Z' \tag{23}$$

We next calculate the difference in the two models' residuals' variance, as this is a commonly used measure for goodness-of-fit. For compactness, define $\hat{\lambda}$ to be the difference in the mark-up terms:

$\hat{\lambda} \equiv \hat{h}^{\text{Qty}} - \hat{h}^{\text{Prc}}$. The difference in the models' residuals' variances is, therefore:

$$\begin{aligned}\hat{\sigma}^2_{\text{Prc}} - \hat{\sigma}^2_{\text{Qty}} &= \frac{1}{T}\left(p - \hat{h}^{\text{Prc}}\right)' H \left(p - \hat{h}^{\text{Prc}}\right) - \frac{1}{T}\left(p - \hat{h}^{\text{Qty}}\right)' H \left(p - \hat{h}^{\text{Qty}}\right) \\ &= \frac{2}{T}\left(p - \hat{h}^{\text{Qty}}\right)' H\hat{\lambda} + \frac{1}{N}\hat{\lambda}' H \hat{\lambda}' \qquad (24)\end{aligned}$$

Equation 24 is very revealling in understanding how the difference in the models' residuals' variances is not informative of the true data generating process. Specifically, if firms truly choose quantities, then $p - \hat{h}^{\text{Qty}}$ should be equal to costs, which we denote with c. In contrast, if firms truly choose prices, then $p - \hat{h}^{\text{Qty}}$ should be equal to $c - \hat{\lambda}$, and therefore, under each of these two different data generating process, the difference in variances is:

$$\text{True Model: Price} \qquad \hat{\sigma}^2_{\text{Prc}} - \hat{\sigma}^2_{\text{Qty}} = \frac{2}{T} c' H \hat{\lambda} - \frac{1}{T}\hat{\lambda}' H \hat{\lambda}' \qquad (25)$$

$$\text{True Model: Quantity} \qquad \hat{\sigma}^2_{\text{Prc}} - \hat{\sigma}^2_{\text{Qty}} = \frac{2}{T} c' H \hat{\lambda} + \frac{1}{T}\hat{\lambda}' H \hat{\lambda}' \qquad (26)$$

Only when true costs are zero or are not correlated with the difference in markups can the true data generating process be selected by choosing the model with the smallest variance. Unfortunately, costs are not uncorrelated with the difference in mark-ups, and it can be shown that $c' H \hat{\lambda}$ is a negative number whenever costs are positive.[10] Therefore, model selection based on the model with the lowest estimated residual variance will favor the model in which firms choose prices.

Alternative model selection criteria that are dependant on the models' residual variances will similarly be biased. For example, model ω's R-square is defined as $R^2 = 1 - \hat{\sigma}^2_{\omega} / (y_{\omega} H_0 y_{\omega})$, where H_0 is the idempotent demeaning matrix, $H - 1/T$. Therefore, the difference in R-squares across the

[10] $\hat{\lambda}$ can be expressed as $\hat{\lambda} = \hat{\Delta} q$ where q are observed sales and $\hat{\Delta}$ is a positive-definite block diagonal matrix, each block corresponding to a market, with block matrix $\hat{\Delta}_m$ given as in eq. 9. Costs, prices, and sales are assumed to be related by the equation $p(q) = h^{\omega}(p(q), q) + c$ where $h^{\omega}(p(q), q)$ is the mark-up term. Assuming the demand curve is twice continously differentiable, the implicit function theorem states firms' FOC equations can be inverted to express costs as a function of sales. Let this expression be $C(\cdot)$. In addition, the intermediate value theorem states that for given sales q, costs c, and intermediate sales vector $\overline{q} \in [0, q]$, costs can be written as $c = \nabla_{\vartheta} C(\overline{q}) \cdot q$. Under XX assumptions, $\nabla_{\vartheta} C(\overline{q})$ is a negative semi-definite matrix. Therefore, the term $2c' T \hat{\lambda}$ can be re-written as $2q' \left(\nabla_{\theta} C(\overline{q})\right)' H \hat{\Delta} q$, where $\nabla_{\theta} C(\overline{q})$ is negative semi-definite and H and Δ are positive semi-definite. As these matrices are being multiplied on both sides by the same vector, the scalar outcome must be non-positive.

two models is calculate as the difference in the models' SSR weighted by a model specific scalar. It is straightforward to show these weights do not account for the bias in the difference in variances, i.e. $R^2_{\text{Prc}} - R^2_{\text{Qty}} \not\propto \hat{\sigma}^2_{\text{Price}} - \hat{\sigma}^2_{\text{Qty}} - \frac{2}{T} c' H \hat{\lambda}$.

A likelihood ratio test based on Normally distributed error terms suffers from a similar problem. The sample log-likelihood of a given model, assuming errors are Normally distributed with variance σ^2, is simply $-\frac{1}{2} T \ln \sigma^2 - \frac{1}{2} \varepsilon' \varepsilon / \sigma^2$. This expression simplifies to $-\frac{T}{2}\left(1 + \ln \hat{\sigma}^2\right)$ if σ^2 is estimated as $\hat{\varepsilon}' \hat{\varepsilon} / T$. Therefore, the log-likelihood of model ω is larger than that of model ω' iff $\hat{\sigma}^2_{\omega'} \geq \hat{\sigma}^2_{\omega}$. Thus, a likelihood ratio test that assumes errors are distributed normally would also favor the 'firms choose prices' model over the 'firms choose quantities' model, even when the true data generating process is one in which firms choose quantities.

4. An Exercise Illustrating the Importance of Correct Model Choice

To demonstrate how imposing the incorrect form of competition can bias estimation, we simulate demand preferences and firm costs, solve for equilibrium pricing and sales under both types of competition, and then attempt to estimate the demand and cost values assuming both the correct and the incorrect type of competition.

For simplicity we assume there are no random coefficients, and therefore demand is given according to a Logit specification. We also assume there are only two products in every market, each characterized by their price and a single random mean-preference, such that the utility to consumer i of purchasing product j is given by $u_{ij} = -\alpha p_j + \xi_j + \varepsilon_{ij}$, where ε_{ij} is the T1EV idionscratic shock, iid across consumers, products, and markets, and ξ_j is a random Normal$(\overline{\delta}, \sigma^2_{\xi})$ shock, iid across products and markets. Products' costs have two sources of costs, first a cost shift $\eta \cdot x$, where η measures the effect of x on costs, x is the cost shifter observed to the researcher, an iid shock across products and markets distributed exponential with scale parameter $1/2$. Secondly, an additive cost shock ε, unobservable to the researcher but known to the firms, a Normally distributed cost shock, iid across products and markets, and with mean $\overline{c}$ and variance σ_{ε}^2. In summary, demand and variable costs for product i in market m are given by

$$s_{im}(p_m) = \frac{e^{\alpha p_{im}+\xi_{im}}}{1+\sum_{j=1}^{2} e^{\alpha p_{jm}+\xi_{jm}}} \qquad c_{im} = \eta x_{im} + \varepsilon_{im} \qquad i \in \{1,2\} \tag{27}$$

We simulate a thousand markets. For each market we calculate equilibrium prices and sales assuming first that firms compete in prices, and second that firms compete in quantities.[11] Details are given in the Appendix. These two simulations produce two sets of equilibrium prices and quantities, to which we append the cost shifter x. Then, for each of the two data sets (each set containing prices, quantities, and cost shifters across all markets), we estimate the original model parameters, $(\alpha, \overline{\delta}, \sigma_{\xi}^2, \eta, \overline{c}, \sigma_{\varepsilon}^2)$, under three different models. The first model, 'Price Model', uses the estimation procedure described above but imposes $\tau = 1$ and therefore assumes mark-ups in the supply equations are calculated as if firms competed in prices. The second model, 'Quantity Model', imposes $\tau = 0$ and therefore assumes mark-ups are given as if firms competed in quantities. The third model, 'Selection Model', does not pre-impose a value for τ but estimates it along with the model parameters. In all three estimations we use as demand instruments a constant and the observed cost shifter x, and as supply instruments a constant, the observed cost shifter x, and the rivals' products' cost shifter. The cost shifter in the demand equation identifies the price coefficient, while that in the supply equation identifies η. The rivals' chost shifter value in the supply equation identifies the model selection parameter τ.

Table 1 contains the true parameter values as well as the estimated parameters under the three different models, for each of the two data sets, i.e. when firms truly compete in prices and when they truly compete in quantities. As is apparent in the table, when parameters are estimated assuming

[11] Vives (2001) shows the equilibrium exists and is unique when firms compete in prices in the above setting. To show existance and uniqueness when firms compete in quantities, note how firms' own prices have a one-to-one mapping to firms' own sales, and therefore any solution to firms' problem when choosing prices is a solution to firms' problem when choosing quantities. This shows that best responses exist and are unique when firms choose quantities. In addition, profit functions are continuous, and therefore, by Brouwer's fixed point theorem, an equilibrium exists when firms choose quantities. Uniqueness follows from the fact that best-reponses are a contraction, as $\partial^2\pi/\partial^2 q_i + \left|\partial^2\pi/\partial q_i \partial q_j\right| = \partial p_i/\partial q_i < 0$; where the equality follows from calculating derivatives directly.

Table 1: Monte Carlo Simulations

	Firms truly choose prices				Firms truly choose quantities			
Estimated Model:	True Value	Price Model	Quantity Model	Selection Model	True Value	Price Model	Quantity Model	Selection Model
Price Coeff. (α)	-2.5	-2.56	-2.56	-2.56	-2.5	-2.56	-2.56	-2.56
		(0.05)	(0.05)	(0.05)		(0.06)	(0.06)	(0.06)
Mean Util. ($\overline{\delta}$)	4.0	4.08	4.08	4.08	4.0	4.10	4.10	4.10
		(0.09)	(0.09)	(0.09)		(0.11)	(0.11)	(0.11)
Mean Cost ($\overline{c}$)	0.5	0.51	-0.18	0.51	0.5	0.75	0.52	0.52
		(0.02)	(0.06)	(0.04)		(0.02)	(0.03)	(0.05)
Cost Shifter (η)	1.0	1.02	1.45	1.02	1.0	0.89	1.02	1.02
		(0.02)	(0.05)	(0.03)		(0.02)	(0.02)	(0.03)
Conduct Parameter	1			0.998	0			6.8E-7
				(0.05)				(0.20)
Demand Shock StdDev (ξ)	1.0	1.01	1.01	1.01	1.0	1.02	1.02	1.02
Supply Shock StdDev (ε)	0.5	0.24	1.93	0.24	0.5	0.18	0.24	0.24
Price Increase from Merger	15%	14%	3.0%	14%	3.4%	8.2%	3.3%	3.3%

Estimated parameters using simulated data from two alternative data generating processes, one in which firms choose prices and one in which firms choose quantities. Each simulated data sets simulates a thousand markets with two firms active in each market. Firms' demand and costs in each market are given as in Eq. 27, with parameter values for α and η reported under the 'True Value' columns. Demand and cost shocks, ξ and ε, are simulated from Normal distributions with means ($\overline{\delta}$ and $\overline{c}$) and standard deviations as reported under the 'True Value' columns. Cost shifters, x, distributed from an Exponential distriubtion with scale parameter $1/2$. Using this simualted cost shifter, along with equilibrium prices and quantities, the model parameters are estimated under three different models. 'Price Model' assumes equilibrium outcomes arise by firms choosing prices, 'Quantity Model' assumes firms choose quantities, and 'Selection Model' nests both prior models and estimates a model selection parameter, bounded between 0 and 1, which selects one of the two prior models (or a mixture of both). Standard errors are shown in parenthesis and standard deviations of estimated residuals are reported. In addition, estimated parameters and residuals are used to predict the average market price increase from a merger between the two firms active in each market. Market price is measured as the share weighted average of each firms' price. 'Price Increase from Merger' reports the percentage market price increase, averaged across markets.

the incorrect form of competition the estimated values are significantly biased. Specifically, the linear cost $\bar{c}$ is underestimated when the model assumes firms compete in quantities when in truth they compete in prices, and overestimated when assuming firms compete in prices when in truth they compete in quantities. However, it is not only the constant cost that is affected, the estimate on the cost shifter, η, is also underestimated when the model assumes firms compete in prices when in truth they compete in quantities. The incosistency is sufficiently large so that the true value is not within a 95% confidence interval of the estimated parameter. In contrast, when the 'correct' form of competition is assumed or when it is estimated, i.e. 'Selection Model', the estimated 95% confidence intervals indeed cover the true parameters. Not surprisingly, estimated standard errors are slightly larger when the form of competition is estimated than when it is assumed correctly.

We would also like to note how the estimated supply shock variance, σ_{ε}^{2}, is always smaller under 'Price Model' than under 'Quantity Model', regardless of the true form of competition. This corraborates the earlier statements on how model selection cannot be based on this statistic. Importantly, both 'Price Model' and 'Quantity Model' include overidentifying instruments that are effective at selecting the 'right' model, i.e. the rivals' costs. Therefore, even with these identifying instruments, the difference in models' 'goodness-of-fit' favor the 'Price Model' over the 'Quantity Model', regardless of the true data generating process. This goes to stress the importance of nesting both models in order to appropriately do model selection, and not simply select models based off of fit.

At the end of the day, demand and cost estimation are themselves an input into a bigger question. One may ask how important is it to correctly estimate the 'right' model when addressing such bigger question. Could it be that the biased estimates obtained from a misspecified model generate, under that misspecified model, similar counterfactuals than unbiased estimates would generate under a correctly specified model? Could one error fix another? Not really. For example, Table 1 includes a row specifying the actual price increases that would result from a merger between the two products in each market, as well as the inferred price increases using the estimated parameters and the assumed model. For the specific values chosen in this exercise, prices increase on average 15% if firms truly compete in prices. However, the infered price increase is of only 3% if it is assumed they compete in quantities. Similarly, the true average price increase is of 3.4% if firms compete in quantities, but it is inferred to be 8.2% if one assumes they compete in prices. Clearly, one error does not fix another: assuming firms compete in prices when in truth they compete in quantities results in overstating price hikes following a merger, and assuming firms compete in quantities when in truth they compete in prices results in understating price increases from a merger.

5. An Application to Various Industries [pending work]

We now apply the above estimation framework to three different industries, all of which have been described at times as industries in which firms set prices and at other times in which firms choose quantities. For each industry we gather information on demand and costs, in which we define a unit of observation a product-market-time triplet. We then implement the estimation procedure described above in which we use cost shifters to identify price elasticities and rivals' cost shifters to identify model selection. Details for each industry are given below.

5.1 The US Domestic Commerical Airline Industry

The US Airline industry has been extensively studied (cite papers) to analyze effects of entry, market power in networks, multi-market contact, merger effects, and more. Research modelling it in strategic substitutes. Research modelling it in strategic complements.

5.1.1 Data Sources

Our data is obtained from four different sources. First, we use data from the Bureau of Transportation Statistics' (BTS) Airline Origin and Destination Survey (DB1B database), which contains a

10% sample of all itineraries sold for domestic flights in the US. The data is reported quarterly and we obtain data from the first quarter of 2010 up to the last quarter of 2017. We aggregate itineraries to the product-route-quarter level, where we define a route as a unidirectional city pair, a product as a carrier-service type pair, and the service type as either being non-stop or connecting service. As the connecting service can be through various hubs, e.g. American Airlines' offers flights from Tucson to Atlanta through both Dallas and Phoenix, we retain information from the modal hub. From this database we retain information on total passengers flown, average price paid, distance flown, and hub city used for connecting service.

Second, we use the Air Carrier Statistics database (T-100 Segment). This data contains, at the monthly level, the number of seats assigned and passengers flown by each carrier's aircraft class on each (directional) airport pair. We aggregate this data to the carrier-city pair-quarter level, retain the identity of the modal aircraft class used (e.g. Boeing 737-300, Airbus A321, etc.) and calculate average load factors by dividing total passengers over total seats assigned. We also match this data with that in the Origin and Destination Survey,[12] calculate the discrepancy in passengers flown between these two databases, and define this discrepancy as the number of passengers using domestic flights as part of their international travel. Finally, we merge this data onto that of the Origin and Destination Survey, such that for all non-stop service flights we retain information on load factors, percent of passengers flying international, and the modal aircraft class used; and for all connecting service we retain similar information for each of the two flight segments conforming such connecting service. For connecting service, we then calculate a single load-factor value, and a single percent-international value, by taking the maximum values across the two flight segments.

Third, we incorporate cost information using the BTS' Form 41 - Financial Data, Schedule P.5.2. This data reports, at the quarterly level, domestic operating costs and statistics for each carrier and

[12] For each itinerary, the Origin and Destination Survey reports two carriers: an *operating carrier*, responsible for operating the flight, and a *ticketing carrier,* responsible for sales of the ticket. In defining a product as a carrier-service type, we use the *ticketing carrier* as our construct of carrier. However, in joining the O&D Survey with the Air Carrier Statistics' database, we match on the *operating carrier,* as this is the carrier whose information is reported in the latter database. Therefore, when aggregating the O&D Survey data, we retain the identity of the modal *operating carrier* corresponding to each *ticketing carrier*-route-quarter observation for non-stop services, and the two modal *operating carriers* corresponding to each flight segment of the connecting service *ticketing carrier*-route-quarter observations.

for each aircraft class. From this data we retain operating costs per hour of airtime, by quarter, carrier, and aircraft class, where operating costs include costs of crews, fuel, maintenance, depreciation, and rental equipment.[13] We merge this data onto data from the Air Carrier Statistics database, discussed above, and using the latters' information on passengers, distance, and air time, we calculate a cost per passenger-mile corresponding to each carrier, quarter, and aircraft class. This cost measure is then merged onto the O&D Survey data, similarly to the load factor data. As connecting service products in the O&D Survery data are conformed of two flight segments, we obtain a single cost measure by averaging the two cost measures of the corresponding flight segments.

Fourth, we obtain cities' economic outlook from the Bureau of Economic Analysis. Specifically, we obtain information on population estimates, personal income, total GDP, and GDP from the leisure industry (i.e. NAICS 71 and 72). The data is reported yearly but transformed to a quarterly basis using linear interpolation. We then merge it onto the O&D Survey data and create socioeconomic indicators at a route-quarter level by geometrically averaging the corresponding values of the two end-point cities. Finally, we use the Bureau of Labor Statistics' CPI to deflate all currency values to 2012 real dollars.

In our finalized dataset a unit of observation is a route-quarter-carrier-service type quadruplet, e.g. one observation would be American Airlines' non-stop service between Tucson and Chicago, for the third quarter of 2015; and a second observation would be American Airlines' service between Tucson and Chicago, connecting in Dallas, for the third quarter of 2015. For each observation, we have information on domestic passengers flown, price paid, distance travelled, aircrafts' average load-factor, cost per passenger-mile, percent of international passengers jointly flying with the domestic passengers, and macroeconomic indicators.

We cleanse the data further by retaining information of only the top domestic carriers[14] and the largest 3000 routes, excluding routes within Hawaii or Alaska and routes to, from, or within US

[13] Item no. *70.989 - Total Aircraft Operating Expense (Direct Operating Expense),* as given in Bureau of Transportation Statistics (2019)

[14] AirTran, Alaska, Allegiant, American, Continental, Delta, Frontier, Jet Blue, Southwest, Spirit, United, US Airways, Virgin

territories. We also remove products with non-substantial service on a given route: those with less than 15% market share *and* less than 80 weekly passengers. In addition, we drop observations with outrageous average prices: below $25 or above $2,500, and consider as incorrect any operating cost values that imply gross margins below 0 or above 90%.[15] Finally, as the model selection parameter is identified only in oligopolistic markets, we remove any monopolistic or quasi-monopolistic routes: routes for whom's largest firm sells 95% or more of all tickets.

5.1.2 Market Size, Demand Shifters, and Cost Shifters

To construct market size, we define as the size of the potential market the geoemtric average of population at endpoint cities [insert citation for who else does this], and scale this number by a constant, common to all markets and time periods, such that the lowest observed market share for the outside option is 2%. This normalizing constant is co-linear with the constant in the demand model, does not affect identification, and allows for numerical stability when calculating mark-ups.

In our main specification we include as demand shifters the average price paid by consumers, as well as an indicator for connecting service, which proxies for the disutility from taking multiple flights. We also include the additional miles travelled relative to the shortest available option to control for the discomfort from longer travel times, as well as the number of destinations available at end-point cities and the number of destinations available at the connecting hub airport, both of which control for the opportunity value to diverting your trip in the last minute as well as potential congestion effects at airports. To control for airport and hub ammenities (i.e. hub dominance effects, as in [insert citation]) we include carriers' market share at endpoint cities as well as the fraction of a carrier's enplanements that occur at the endpoint cities.[16] We control for aggregate demand shifts with quarterly fixed effects as well as the various macroeconomic indicators mentioned

[15]Among the finalized dataset, less than 3% of the passenger-weighted observations have missing or innacurate operating cost values.

[16]The former measure is traditionally used in the airline literature as a measure of hub dominance. The latter measure accounts for hub ammenities from smaller, more focalized carriers. For example, although JFK International Airport is Jet Blue's largest hub, Jet Blue's market share in NYC is limitted due to American and Delta's large presence. In such situation, Jet Blue's ammenities at NYC may best be reflected by way of the fraction of JetBlue's enplanements at endpoint cities rather than Jet Blue's market share at endpoint cities.

earlier and for long-run differences in market size across routes with route fixed effects. Finally, we allow for the coefficient on price and on connecting service to be Normally distributed, where the variance and covariance elements are parameters to be estimated.

As for the cost shifters, we include the operating cost variable derived from firms' financial reports and described above. We also include a dummy for connecting service and the additional miles travelled relative to the shortest flight to account for the incremental costs from longer triangulations and multiple take-offs. Average load factors control for aircraft utilization and economies of scale, and the percent of international passengers accounts for the opportunity cost of selling an international ticket at the last minute relative to a domestic ticket. We also include the number of destinations a carrier flys to from the endpoint cities and from the connecting hub airport to control for economies of scale at the relevant airports, and carriers' market share at the endpoint cities to proxy for any monopsony power the carrier may exert to reduce input costs. Finally, quarterly fixed effects control for nationwide shifts in costs, e.g, varying fuel costs; route fixed effects control for long-run differences across routes, e.g. differences in flight time; and carrier by aircraft class fixed effects control for differences in the aircraft used to service routes, and how aircraft costs are specific to firms due to, for example, differences in in union contracts and in fuel contracts.

We instrument for price with the operating cost measure and with the percent of international passengers. The variance and covariance parameters of the random coefficients are instrumented for with the variance in the operating cost measure across products offered on the same route and quarter, with the number of non-stop products offered, and with the product of these two instruments. Finally, the model selection parameter is instrumented using rivals' operating cost measure and a dummy variable for monopolistic routes.

For a few specifications we additionally include one of two groups of fixed effects in both the demand and the supply equation: either route by product fixed effects or route by time fixed effects. The former imposes that all identification arise from changes in behaviour over time, while the latter that identification arise from differences in behaviour across firms on the same route. Therefore, under the former identification is likely to arise from short-run differences while under the latter

Table 2: US airline industry - main results

	Logit			Random Coefficiencies		
	(I)	(II)	(III)	(IV)	(V)	(VI)
Model Selection Parameter	18.2	0.096	-27.3			
	(7.25)	(0.03)	(6.30)			
Price Coefficient	-0.006	-0.007	-0.005			
	(14.9)	(21.7)	(61.4)			
Variance of Demand Residuals	0.487	0.275	0.426			
Variance of Supply Residuals	239.5	24.49	369.7			
Route-Carrier-Aircraft FE		Y			Y	
Route-Year-Quarter FE			Y			Y

* T-statistics in parenthesis. All regressions use 136,492 observations. See text for details. A model selection parameter of ‘0’ is indicative of competition in quantities; a value of ‘1’ is indicative of competition in prices.

from long-run differences, thus potentially affecting the estimate on the model selection parameter.

5.1.3 Results

Discuss tables

6. Conclusion

....

References

Berry, S., J. Levinsohn, and A. Pakes (1995, July). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.

Bresnahan, T. F. (1989). Chapter 17 empirical studies of industries with market power. Volume 2 of *Handbook of Industrial Organization*, pp. 1011 – 1057. Elsevier.

Table 3: Additional demand and supply estimates for the US domestic airline industry

		Logit			Random Coefficiencies		
		(I)	(II)	(III)	(IV)	(V)	(VI)
Demand	Connecting service {0,1}	0.308	1.18	-0.330			
		(2.89)	(5.20)	(2.64)			
	Incremental distance (%)	-0.009	-0.006	-0.009			
		(60.3)	(17.8)	(69.9)			
	Market share at endpoints (%)	-0.000	0.005	0.001			
		(0.18)	(5.60)	(5.06)			
	Share of enplanements at endpoints (%)	0.023	0.040	0.024			
		(25.4)	(25.3)	(29.3)			
	Destinations at endpoints (#)	0.010	0.016	0.009			
		(50.8)	(47.3)	(45.6)			
	Destinations at hub (#)	0.006	0.005	0.006			
		(36.8)	(19.9)	(50.9)			
	Variance on price						
	Variance on Connecting service						
	Covariance on price & connecting service						
Supply	Operating cost ($)	0.120	0.115	0.14			
		(15.5)	(10.9)	(18.8)			
	Connecting service {0,1}	-5.06	15.7	20.4			
		(0.41)	(1.11)	(8.46)			
	Incremental distance (%)	0.103	-0.185	-0.020			
		(6.96)	(6.30)	(8.46)			
	Market share at endpoints (%)	0.684	2.024	0.748			
		(5.89)	(5.75)	(24.6)			
	Share of enplanements at endpoints (%)	0.604	0.251	3.19			
		(1.64)	(0.80)	(68.7)			
	Destinations at endpoints (#)	0.088	-0.662	-0.500			
		(1.51)	(0.39)	(21.9)			
	Destinations at hub (#)	0.327	0.492	0.300			
		(32.2)	(14.3)	(28.7)			
	Load factor (%)	-51.7	-39.79	19.6			
		(7.28)	(6.18)	(2.52)			
	International pass. (%)	29.34	24.92	90.7			
		(4.97)	(4.52)	(11.1)			

* T-statistics in parenthesis. See text for details.

Bulow, J. I., J. D. Geanakoplos, and P. D. Klemperer (1985, January). Multimarket oligopoly: Strategic substitutes and complements. *Journal of Political Economy 93*(3), 488.

Bureau of Transportation Statistics (2019, January). Financial reporting categories (item list guide). https://www.bts.gov/explore-topics-and-geography/topics/financial-reporting-categories-item-list-guide.

Corts, K. S. (1999). Conduct parameters and the measuremet of market power. *Econometrica 88*, 227–250.

Ericson, R. and A. Pakes (1995, January). Markov-perfect industry dynamics: A framework for empirical work. *The Review of Economic Studies 62*(1), 53.

Gandhi, A., L. Froeb, S. Tschantz, and G. J. Werden (2008). Post-merger product repositioning. *Journal of Industrial Economics 56*, 49–67.

Seim, K., M. Mazzeo, and M. Varela (forthcoming). Welfare consequences of product repositioning. *Journal of Industrial Economics*.

Sweeting, A., S. Li, J. Mazuer, J. Roberts, Y. Park, and J. Zhang (2018). Endogenous and selective service choices after airline mergers. *working paper*.

Vuong, Q. H. (1989, March). Likelihood ration tests for model selection and non-nested hypotheses. *Econometrica 57*(2), 307–333.

A. *Inferred costs under the 'Price Model' are higher than under the 'Quantity Model'*

Recall the difference in inferred costs is given by

$$c^{\text{Prc}} - c^{\text{Qty}} = \left(\left(A \circ d\hat{Q}\right)^{-1} - A \circ \left(d\hat{Q}\right)^{-1}\right)\hat{q}$$

where A is an ownership matrix of zeros and ones. Ordering products according to firm ownership, the matrices $\left(A \circ d\hat{Q}\right)^{-1}$ and $A \circ \left(d\hat{Q}\right)^{-1}$ are block diagonal matrices with each block corresponding to a given firms' products (rows) and prices (columns). Importantly, the block's to the diagonal of $\left(A \circ d\hat{Q}\right)^{-1}$ are simply $\left\{d\hat{Q}_{ii}^{-1}\right\}_{i=1}^{I}$, the slopes of firm i's products with respect to firm i's own prices. Let $d\hat{Q}_{i\cdot}$, $d\hat{Q}_{\cdot i}$, and $d\hat{Q}_{\cdot\cdot}$ be the corresponding conforming submatrices such that

$$d\hat{Q} = \begin{bmatrix} d\hat{Q}_{ii} & d\hat{Q}_{i\cdot} \\ d\hat{Q}_{\cdot i} & d\hat{Q}_{\cdot\cdot} \end{bmatrix}$$

Using block-matrix inverse, for some matrices Δ_i, B_i, and C_i, the inverse of $d\hat{Q}$ is

$$\left(d\hat{Q}\right)^{-1} = \begin{bmatrix} d\hat{Q}_{ii}^{-1} + \Delta_i & B_i \\ B_i' & C_i \end{bmatrix} \tag{28}$$

$$\Delta_i \equiv d\hat{Q}_{ii}^{-1} \cdot d\hat{Q}_{i\cdot} \left(d\hat{Q}_{\cdot\cdot} - d\hat{Q}_{\cdot i} d\hat{Q}_{ii}^{-1} d\hat{Q}_{i\cdot}\right)^{-1} d\hat{Q}_{\cdot i} \cdot d\hat{Q}_{ii}^{-1} \tag{29}$$

Therefore, $A \circ \left(d\hat{Q}\right)^{-1}$ is a block diagonal matrix where each diagonal block is simply $\left\{d\hat{Q}_{ii}^{-1} + \Delta_i\right\}_{i=1}^{I}$, and $\left(\left(A \circ d\hat{Q}\right)^{-1} - A \circ \left(d\hat{Q}\right)^{-1}\right)$ is a block diagonal matrix with $-\Delta_i$ on the diagonals. Define Δ as this block-diagonal matrix and note that Δ_i are negative semi-definite (n.s.d.) matrices: Δ_i is formed by squaring $d\hat{Q}_{ii}^{-1} \cdot d\hat{Q}_{i\cdot}$, weighted by $\left(d\hat{Q}_{\cdot\cdot} - d\hat{Q}_{\cdot i} d\hat{Q}_{ii}^{-1} d\hat{Q}_{i\cdot}\right)^{-1}$, and therefore is n.s.d. if the weighting matrix is n.s.d. This weighting matrix must be n.s.d. as it is equal to matrix C_i (in eq.

29), and C_i is n.s.d. as it is a principal sub-matrix of $\left(d\hat{Q}\right)^{-1}$, a negative definite matrix.

B. Implementation Suggestions for the GMM Procedure

B.1 Numerical Integration for Inverting the Share Equation

The estimation procedure requires solving for the vectors of mean utilities, $\{\delta_m(\theta)\}_{m=1}^{M}$, as the inverse to the set of equations

$$s_{mj}(\delta_m,\theta) = \int \exp\left[\delta_{mj}+\mu_{mj}^{r}\right] / \left(1+\sum_{l=1}^{J}\exp\left[\delta_{ml}+\mu_{ml}^{r}\right]\right) dF(\beta|\theta) \qquad \forall j=1..J_m \quad \forall m=1..M \tag{30}$$

$$\mu_{mj}^{r} = x_{mj}^{\mathrm{D}}\beta_r + p_{mj}\alpha_r \tag{31}$$

The integral in the above equation may not have a known closed-form solution, and therefore it may be convenient to solve the integral using numerical approximations. Two such approximations are Monte Carlo integration (for general distributions $F(\beta)$) and Guass-Hermite quadrature (whenever $F(\beta)$ is a Normal CDF; other quadratures may be used for alternative CDFs).

Monte Carlo integration approximates eq. 30 by simulating R draws of $\tilde{\beta}_r$ from distribution $F(\beta|\theta)$ and calculating the integral as

$$\frac{1}{R}\sum_{r=1}^{R}\exp\left[\delta_{mj}+\tilde{\mu}_{mj}^{r}\right] / \left(1+\sum_{l=1}^{J}\exp\left[\delta_{ml}+\tilde{\mu}_{ml}^{r}\right]\right)$$

For the special case in which $F(\beta|\theta)$ is a Normal distribution with zero mean and variance matrix $V(\theta)$, the calculation of $\tilde{\mu}_{mj}^{r}$ can be executed in matrix form by defining first Σ to be the principal square root of $V(\theta)$ and by simulating R draws from the standard multi-variate Normal. Let ζ^r be one such draw and define X_{mj}^{D} to be the J_m-by-(K_D+1) matrix of product characteristics (including price) for the J_m products in the market. Compute the market-specific vector of random utilities as $\tilde{\mu}_{m}^{r} = X_{mj}^{\mathrm{D}}\Sigma\zeta^{r}$.

Alternatively, a Gauss-Hermite quadrature approximates the integral given R weight-abscissa pairs $\{w^r, \zeta^r\}_{r=1}^R$ as

$$\sum_{r=1}^{R} w^r \exp\left[\delta_{mj} + \bar{\mu}^r_{mj}\right] / \left(1 + \sum_{l=1}^{J} \exp\left[\delta_{ml} + \bar{\mu}^r_{ml}\right]\right) \qquad ; \qquad \bar{\mu}^r_m \equiv X^{\mathrm{D}}_{mj}\Sigma\zeta^r$$

Therefore, when the random coefficients are Normally distributed, Monte Carlo approximations and Gauss-Hermite quadratures share a similar structure, differing solely in the values of the weight-abscissa pairs. This notation will be convenient in calculating analytical derivatives described below.

B.2 Simplifying Random Coefficients

If the distribtion of random coefficients, $F(\beta|\theta)$, is a parametrized joint-Normal distribution with mean zero and variance matrix $V(\theta)$, as is in the applications in this paper, it may be easiest to parametrize Σ, a square root of the variance matrix. That is, define $\Sigma(\theta)$ as

$$\Sigma(\theta) \equiv \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_{K_\mathrm{D}} \\ & \theta_{K_\mathrm{D}+1} & \cdots & \theta_{2K_\mathrm{D}-1} \\ & & \ddots & \vdots \\ & & & \theta_{K_\mathrm{D}(K_\mathrm{D}+1)/2} \end{bmatrix} \tag{32}$$

and calculate the variance matrix $V(\theta)$ as $\Sigma'\Sigma$. This approach requires making no restrictions on the parameters θ (e.g. $V(\theta)$ must be a positive-semidefinite matrix; Σ need not be) and simplifies calculating demand and slopes of demand using the numerical approximations described above, which rely on Σ and not on V.

B.3 Analytical Derivatives

B.3.1 Overall

Recall $\psi \equiv (\tau, \alpha, \theta)$ and $y(\psi) \equiv \left[\delta - p\alpha \; ; \; p - \tau h^{\text{Prc}} - (1-\tau)\, h^{\text{Qty}}\right]$. In particular, δ is obtained by inverting the share equation and therefore is a function of random coefficient parameters θ. As for the mark-up terms, h^{Prc} and h^{Qty}, these are functions of mean utilities δ, price coefficient α, and random coefficient parameters θ. Thus, the gradient of $y(\psi)$ with respect to ψ is:

$$\nabla_{(\tau,\alpha,\theta)} y(\psi) = \begin{bmatrix} 0 & -p & \nabla_\theta \delta \\ h^{\text{Qty}} - h^{\text{Prc}} & -\tau \frac{\partial}{\partial \alpha} h^{\text{Price}} - (1-\tau)\, \frac{\partial}{\partial \alpha} h^{\text{Qty}} & -\tau \frac{d}{d\theta} h^{\text{Price}} - (1-\tau)\, \frac{d}{d\theta} h^{\text{Qty}} \end{bmatrix}$$

where $\nabla_x \equiv \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_K}\right)$ and $\frac{d}{d\theta} h^\omega \equiv \nabla_\theta h^\omega + \nabla_\delta h^\omega \nabla_\theta \delta$ for $\omega \in \{\text{Prc}, \text{Qty}\}$.

Mean utilities and mark-ups are calculated market-by-market, and therefore their gradients can also be calculated market-by-market. In what follows we show how to calculate these derivatives for a specific market. Importantly, we assume mean utilities are obtained by inverting the share equation, where the integral in such equation is approximated with R weight-abscissa pairs $\{w^r, \zeta^r\}_{r=1}^R$. Therefore, δ_m is defined as the solution to the fixed point:

$$0 = \hat{s}_m - \sum_r w^r s_m^r \quad ; \quad s_m^r \equiv \frac{\exp\left[\delta_m + \bar{\mu}_m^r\right]}{1 + j_m' \cdot \exp\left[\delta_m + \bar{\mu}_m^r\right]} \quad ; \quad \bar{\mu}_m^r = X_{mj}^{\text{D}} \Sigma \zeta^r$$

where j_m is a J_m-by-1 vector of ones.

B.3.2 Mean Utilities

We use the implicit function theorem to calculate $\nabla_\theta \delta_m$, i.e. given the fixed point equation $g(x,a) = 0$, the slope of x wrt a is $\nabla_a x = -(\nabla_x g)\nabla_a g$. Therefore,

$$\begin{aligned}\nabla_\theta \delta_m &= -\left(\nabla_\delta\left(\hat{s}_m - \sum_r w^r s^r\right)\right)^{-1} \nabla_\theta\left(\hat{s}_m - \sum_r w^r s^r\right) \\ &= -\left(\sum_r w^r \nabla_\delta s_m^r\right)^{-1}\left(\sum_r w^r \nabla_\theta s_m^r\right)\end{aligned}$$

where

$$\nabla_\delta s_m^r = \text{diag}\left[s_m^r\right] - s_m^r s_m^{r\prime} \qquad ; \qquad \nabla_\theta s_m^r = \nabla_\delta s_m^r \cdot \nabla_\theta \mu_m^r$$

Finally, as $\bar{\mu}_m^r = X_m^{\text{D}} \Sigma \zeta^r$, $\nabla_\theta \bar{\mu}_m^r$ can be calculated one column at a time: $\frac{\partial \bar{\mu}_m^r}{\partial \theta_k} = X_m^{\text{D}} \frac{\partial \Sigma}{\partial \theta_k} \zeta^r$, where $\frac{\partial \Sigma}{\partial \theta_k}$ is a matrix of ones and zeros.

B.3.3 Mark-Up Terms

Recall the mark-up terms are calculated as $h^{\text{Prc}} \equiv -(A_m \circ ds_m)^{-1}\hat{s}_m$ and $h^{\text{Qty}} \equiv -\left(A_m \circ (ds_m)^{-1}\right)\hat{s}_m$. As with the share equation, ds_m can too be calculated using numerical approximations:

$$ds_m = \sum_r w_r \left[\alpha + \alpha^r\right] \nabla_\delta s_m^r \tag{33}$$

where $\alpha^r = \Sigma_{(c,\cdot)} \zeta^r$, $\Sigma_{(c,\cdot)}$ is the c-th row of Σ, and c is the column of X_m^{D} containing products' prices.

The derivatives of h^{Prc} and h^{Qty} can be calculated using matrix derivatives, e.g. $\frac{d}{dx}(A \circ U(x)) =$

$A \circ \frac{d}{dx}U(x)$ and $\frac{d}{dx}U(x)^{-1} = -U(x)^{-1} \cdot \frac{d}{dx}U(x) \cdot U(x)^{-1}$. Therefore,

$$\frac{\partial}{\partial \psi_k} h_m^{\text{Prc}} = (A \circ ds_m)^{-1} \left(A \circ \frac{\partial ds_m}{\partial \psi_k} \right) (A \circ ds_m)^{-1} \hat{s}$$
$$\frac{\partial}{\partial \psi_k} h_m^{\text{Qty}} = \left(A_m \circ \left(ds_m^{-1} \frac{\partial ds_m}{\partial \psi_k} ds_m^{-1} \right) \right) \hat{s}$$

In calculating $\frac{\partial ds_m}{\partial \psi_k}$, recall ds_m is given in eq. 33 and $\nabla_\delta s_m^r = \text{diag}[s_m^r] - s_m^r s_m^{r\prime}$. Therefore,

$$\frac{\partial ds_m}{\partial \psi_k} = \sum_{r=1} w_r \left[\left(\frac{\partial \alpha}{\partial \psi_k} + \frac{\partial \alpha^r}{\partial \psi_k} \right) \nabla_\delta s_m^r + (\alpha + \alpha_r) \left(\text{diag}\left[\frac{\partial s_m^r}{\partial \psi_k} \right] - s_m^r \left(\frac{\partial s_m^r}{\partial \psi_k} \right)' - \frac{\partial s_m^r}{\partial \psi_k} s_m^{r\prime} \right) \right]$$

$\partial \alpha / \partial \psi_k$ and $\partial \alpha^r / \partial \psi_k$ are straightforward to compute: $\partial \alpha / \partial \psi_k = 1$ if ψ_k refers to α and 0 otherwise, and $\partial \alpha^r / \partial \psi_k = \frac{\partial \Sigma_{(c,\cdot)}}{\partial \psi_k} \zeta^r$. To calculate $\partial s_m^r / \partial \psi_k$, recall s_m^r is a function of the sum of mean utilities and random utilities, and therefore $\frac{\partial s_m^r}{\partial \psi_k} = \nabla_\delta s_m^r \left(\frac{\partial \delta_m}{\partial \psi_k} + \frac{\partial \bar{\mu}_m^r}{\partial \psi_k} \right)$. As neither δ_m nor $\bar{\mu}_m^r$ depend on τ and α, these partial derivatives are zero whenever ψ_k refers to τ or α. For the cases in which ψ_k refers to an element of θ, $\frac{\partial \delta_m}{\partial \psi_k}$ and $\frac{\partial \bar{\mu}_m^r}{\partial \psi_k}$ have been described in the prior subsection.

B.4 Restricting Parameters

The model selection parameter τ is meant to choose between one of the two models. Hence the researcher may wish to restrict the feasible values of τ to at least lie within the unit interval, or, furthermore, to be either 0 or 1. If the latter restriction is applied, the restriction can be implemented by minimizing $f(\rho, \psi)$ imposing first that $\tau = 0$, and secondly that $\tau = 1$, and selecting the τ value corresponding to the smaller of the two minimized function values. This restriction is easily implemented but does not generate a statistical inference around the selection parameter. Alternatively, the researcher can restrict τ to lie within the unit interval by replacing τ with a monotonic Kernel function $K(\cdot) : \Re \to [0,1]$, such that $\tau = K(\tilde{\tau})$, and estimate $\tilde{\tau}$. An example of an appropriate Kernel function is the logit function: $\Lambda(\tilde{\tau}) = 1/(1 + \exp(-\tilde{\tau}))$. Having estimated $\hat{\tilde{\tau}}$, the assymptotic variance of τ can be obtained using the delta method, i.e. $\text{Avar}[\tau] = (K'(\tilde{\tau}))^2 \, \text{Avar}[\tilde{\tau}]$.

It may also be convenient to restrict price coefficients to be negative, as the optimal pricing equations used to estimate costs are no longer sufficient conditions for equilibrium when demand is upward sloping. Such a restriction can be implemented be setting the mean price coefficient to zero and defining the random-coefficient on price as: $\alpha_r = -\exp(-\tilde{\alpha} - \tilde{\alpha}_r)$ where $\tilde{\alpha}_r$ is distributed, jointly with β_r, according to some mean-zero parametrized distribution $F(\theta)$. This transformation still allows for a non-zero average price coefficient across all consumers, $-e^{-\tilde{\alpha}}\mathbb{E}\left[e^{\tilde{\alpha}_r}\right]$. Importantly, the demand equation (eq. 12) no longer includes an additive mean price term. Instead, the fixed point equation need to be adjusted, such that the inverted mean utilities, $\tilde{\delta}_m$, solve $\hat{s}_m = \int \exp\left[\tilde{\delta}_m - pe^{-\tilde{\alpha}-\tilde{\alpha}_r} + \tilde{\mu}_r\right] / \left(1 + j' \cdot \exp\left[\tilde{\delta}_m - pe^{-\tilde{\alpha}-\tilde{\alpha}_r} + \tilde{\mu}_r\right]\right)$, where $\tilde{\mu}_r$ no longer includes a price term: $\tilde{\mu}_r \equiv X_m^{\mathrm{D}}\beta_r$. Therefore, these mean utilies become a function of the mean price parameter $\tilde{\alpha}$, as well as the other parameters governing the random coefficients, θ. Gradients need to be adjusted accordingly.

B.5 Scaling for Numerical Stability

To allow for numerical stability, let $\overline{p} = p \cdot \nu$ for scaling scalar $\nu \equiv 1/\sqrt{p'p}$, and let $\overline{X} = X \cdot \Delta$ where Δ is a diagonal matrix whose i-th diagonal element is $1/\sqrt{X'_{(\cdot,i)}X_{(\cdot,i)}}$ if such inverse exists, and one otherwise. The estimating equations are

$$\begin{bmatrix} \xi \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \delta - \overline{p}\alpha/\nu \\ p - \tau h^{\text{Prc}} - (1-\tau)h^{\text{Qty}} \end{bmatrix} - \overline{X}\Delta^{-1}\rho$$

Substitute α/ν with $\overline{\alpha}$ and $\Delta^{-1}\rho$ with $\overline{\rho}$, and estimates $\hat{\overline{\alpha}}$ and $\hat{\overline{\rho}}$ directly. Given these estimates, recover $\hat{\alpha}$ and $\hat{\rho}$ as $\hat{\alpha} = \hat{\overline{\alpha}}\nu$ and $\rho = \Delta\rho$. With these adjustments, mark-ups also need to be adjusted for $\overline{\alpha}$, with $ds_m = \sum_{r=1}^{r} w^r \left(\overline{\alpha}\nu + \Sigma_{(c,\cdot)}\zeta^r\right) \nabla_\delta s_m^r$.

Additional scaling can also be done when computing random utilities: calculate $\overline{X}_m^{\mathrm{D}} = X_m^{\mathrm{D}}\Delta^{\mathrm{D}}$ where Δ^{D} a diagonal matrix whose i-th diagonal element is $1/\sqrt{X_{(\cdot,i)}^{\mathrm{D}\prime}X_{(\cdot,i)}^{\mathrm{D}}}$ is such inverse exists, and zero otherwise. Random utility is therefore $\mu_m^r = \overline{X}_m^{\mathrm{D}}\left(\Delta^{\mathrm{D}}\right)^{-1}\Sigma\zeta^r$, and one can replace Σ with $\overline{\Sigma} = \left(\Delta^{\mathrm{D}}\right)^{-1}\Sigma$, estimate $\hat{\overline{\Sigma}}$, and recover $\hat{\Sigma} = \Delta^{\mathrm{D}}\hat{\overline{\Sigma}}$. Given these changes of variables, ds_m needs to be adjusted when calculating the random price coefficients: $\Sigma_{(c,\cdot)}\zeta^r = \Delta^{\mathrm{D}}_{(cc)}\overline{\Sigma}_{(c,\cdot)}\zeta^r$.

B.6 Preserving Smoothness when using an Innerloop Search Process

Most numerical procedures that invert eq. 11 to obtain $\delta(\theta)$ implement an iterative procedure that repeats itself until a convergence criteria is met. For example, BLP suggest initiating δ^0 at some value, and updating δ^{v+1} using δ^v with the following market-by-market updating rule:

$$\delta_m^{v+1} = \delta_m^v + \ln \hat{s}_m - \ln\left[\sum_r w^r s_m^r(\delta^v)\right] \quad ; \quad s_m^r(\delta_m^v) \equiv \frac{\exp\left[\delta_m^v + \bar{\mu}_m^r\right]}{1 + j_m' \cdot \exp\left[\delta_m^v + \bar{\mu}_m^r\right]}$$

An iterative procedure as that described above implies the inverted mean utilities, δ, are non-smooth in θ: the number of cycles required for convergence may differ for different values of θ, therefore generating a discontinuous change in δ for a continuous change in θ.

Smoothness can be re-introduced by altering the procedure under which $\delta(\theta)$ is obtained. In particular, instead of iterating until convergence, one can iterate for a fixed period of cycles. The gradient of $\delta(\theta)$ with respect to θ can then be calculated also through an iterative procedure: set $\frac{\partial \delta^0}{\partial \theta} = 0$ and iterate

$$\frac{\partial \delta_m^{v+1}}{\partial \theta} = \frac{\partial \delta_m^v}{\partial \theta} - \frac{1}{\sum_r w^r s_m^r(\delta_m^v)} \circ \left(\sum_r w^r \nabla_\delta s_m^r \cdot \left(\frac{\partial \delta_m^v}{\partial \theta} + \frac{\partial \bar{\mu}_m^r}{\partial \theta}\right)\right)$$

If the number of iterations is fixed, say at $\bar{v}$, the resulting $\delta_m^{\bar{v}}(\theta)$ is a smooth function of θ, and its derivative is exactly $\frac{\partial}{\partial \theta}\delta_m^{\bar{v}}$. To ensure that $\delta_m^{\bar{v}}$ is a sufficiently close approximation of $\delta(\theta)$, $\bar{v}$ should be chosen large enough so that $\sup_{\theta \in \Theta}\left|\delta_m^{\bar{v}}(\theta) - \delta(\theta)\right| < \varepsilon$ for a pre-specificed tolerance level ε. Hence, one can create a random sample of points $\{\theta^w\}_{w=1}^W$, $\theta^w \in \Theta$, and find the required iterations for convergence at each of these points: $v_w = \min \tilde{v} \quad s.t. \quad \left|\delta_m^{\tilde{v}}(\theta) - \delta(\theta)\right| < \varepsilon$. Given these iteration counts, define $\overline{v}$ as the maximum of these, i.e. $\overline{v} = \max_w v_w$.

As this method does not guarantee that $\left|\delta_m^{\bar{v}}(\theta) - \delta(\theta)\right| < \varepsilon$ at the parameters that minimize the objective function, i.e. at $\theta \rightarrow \hat{\theta}$, it is necessary to verify that convergence is indeed achieved at the estimated parameter $\hat{\theta}$. If $\delta_m^{\overline{v}}\left(\hat{\theta}\right)$ indeed converges to $\delta_m\left(\hat{\theta}\right)$, then $\hat{\theta}$ is truly a local minimizer of the estimating function: as $\delta_m^{\overline{v}}(\theta)$ and $\delta_m(\theta)$ are smooth in θ, $\delta_m^{\overline{v}}(\theta)$ also converges to $\delta_m(\theta)$ in the neighborhood of $\hat{\theta}$.