Industrial Organization Notes

Vasco Villas-Boas

January 15, 2026

CONTENTS

1 SIMPLEST FIRM PROBLEMS

1.1 The Monopoly Problem

We start with a quantity-setting monopolist facing a known inverse demand curve $P(Q)$ and costs $C(Q)$.

$$\pi(Q) := P(Q) \cdot Q - C(Q) - F$$

We take the FOC and derive the *Lerner Index*:

$$\begin{aligned} & \pi'(Q) = 0 \\ \implies & P'(Q) \cdot Q + P(Q) = C'(Q) \\ \implies & \frac{P(Q) - C'(Q)}{P(Q)} = -\frac{P'(Q) \cdot Q}{P(Q)} =: \frac{1}{|\epsilon_d|} \end{aligned}$$

where ϵ_d is the elasticity of demand faced by the firm.

We can also write the above as

$$P\left(1 + \frac{1}{|\epsilon_d|}\right) = MC$$

This is helpful because it shows us the important result that the monopolist never produces in the inelastic portion of the demand curve (ie., $\epsilon_d \in (-1, 0]$). We also define the market elasticity ϵ_D, which is equal to the firm elasticity ϵ_d in this monopolist setting.

1.2 Symmetric Cournot Model

Assume constant marginal cost $c_i = c$ and n equally sized firms to make life easy. We let $Q := \sum_{i=1}^{n} q_i$ the total output of the industry.

We write profits:

$$\pi_i(q_i) = (P(Q) - c_i) \cdot q_i$$

The first order is:

$$\begin{aligned} & 0 = \pi_i'(q_i) \\ \implies & 0 = (P(Q) - c_i) + q_i \cdot P'(Q) \cdot \frac{\partial Q}{\partial q_i} \end{aligned} \tag{1}$$

Cournot competition with simultaneous quantity setting implies that $\frac{\partial Q}{\partial q_i} = 1$ and $\frac{\partial q_j}{\partial q_i} = 0$ for $i \neq j$. We can then use the symmetry of the equilibrium (ie., $q_i = \frac{Q}{n}$) and marginal costs in the game to reach that

$$\begin{aligned} P(Q) + P'(Q) \cdot q_i &= P(Q) + P'(Q) \cdot \frac{Q}{n} \\ &= c \end{aligned}$$

We can rearrange this expression to form the *Lerner Index*

$$\begin{aligned} \frac{P(Q) - c}{P(Q)} &= -\frac{1}{n}\left(\frac{Q}{P(Q)}\right) P'(Q) \\ &= \frac{1}{n|\epsilon_D|} \end{aligned}$$

where $\epsilon_D := \frac{\partial Q/Q}{\partial P(Q)/P(Q)}$. In general, we see that firms face an elasticity of $\epsilon_d = n \cdot \epsilon_D$. That means in this setting firms will not increase prices even though market demand is inelastic.

1.3 Bertrand Paradox

We consider a model with symmetric marginal costs $c_i = c$ and no fixed costs where firms compete by simultaneously announcing the price they wish to set on a homogenous good. All consumers go to purchase their good from the firm that sets the lowest price. The Nash equilibrium of this game with continuous pricing occurs with $p = c$. That would mean in general that firms

This result does not explain all the firms in the world that have operating profits. To complicate this game, we can consider

- Adding capacity constraints. In this model, we have the incredible assumption that if a firm undercuts smallest market price by a small value, then that firm will face all of the demand. It is hard to believe that a firm can suddenly scale its operations to capture all demand.
- Adding other frictions. Firms may choose to have items without prices and then ask for information about the customer to set discriminatory prices.
- Adding product differentiation. The original model assumes that all firms sell an identical good but firms can advertise to make their product seem different or they can literally produce a different good.

1.4 Asymmetric Cournot Model

Now, we assume non-constant marginal costs c_i. We can build off the first order condition from (1).

$$0 = (P(Q) - c_i) + q_i \cdot P'(Q) \cdot \frac{\partial Q}{\partial q_i}$$

We now have that

$$\begin{aligned} \frac{q_i}{Q} \cdot \frac{\partial Q}{\partial q_i} &= \frac{q_i}{\sum_{j=1}^{n} q_j} \\ &=: s_i \end{aligned}$$

We can in turn write the Cournot markup / Lerner Index as

$$\frac{P(Q) - q_i}{P(Q)} = \frac{s_i}{|\epsilon_D|}$$

where $\epsilon_D := \frac{\partial Q/Q}{\partial P(Q)/P(Q)}$ as above.

1.5 Weighted Lerner Index and the Hirschman-Herfindal Index

We define the Lerner index

$$\begin{aligned} WLI &= \sum_{i=1}^{n} \frac{P - c_i}{P} s_i \\ &= \sum_{i=1}^{n} \frac{s_i^2}{|\epsilon_D|} \end{aligned}$$

With $|\epsilon_D| = 1$, we define the *Hirschman-Herfindal Index*

$$HHI = \sum_{i=1}^{n} s_i^2$$

This quantity gives us a measure of market concentration as we take s_i in percentages when computing these quantities. As rules of thumb, the DOJ and the FTC describe markets as:

- Highly concentrated: $HHI \geq 1800$ (this merits scrutiny)
- Moderately concetrated: $HHI \in [1500, 1800)$
- Unconcetrated: $HHI < 1500$

We can also work backwards from the HHI to get the effective number of symmetric firms. Here, HHI is in units $[0, 1]$ instead of $[0, 10000]$. if each of the firms were symmetric, then $s_i = \frac{1}{n}$. In that case, we have that

$$HHI = \sum_{i=1}^{n} s_i^2 = \frac{1}{n}$$

Thus, we define the effective number of symmetric firms, n^*, as $n^* := \frac{1}{HHI}$.

1.5.1 Complaints about HHI

HHI only relates to market power under Cournot assumptions. That is, we assume that firms simultaneously set quantities so that $\frac{\partial Q}{q_i} = 1$. Competition is about setting quantity rather than price, which imposes strong restrictions on cross-price elasticities. It's also unclear that quantity instead of price is the strategic variable.

Two additional complaints are that products aren't necessarily homogeneous and as a result it's sometimes hard to define markets in the first place.

1.6 Conjectural Variations to Cournot

The biggest complaint about Cournot is that we hold the quantities of competitors fixed. Suppose we did not do that so

$$\frac{\partial Q_i}{\partial q_i} = 1 + \frac{\partial Q_{-i}}{\partial q_i}$$

We can again build off Equation (1) to write that

$$\begin{aligned} 0 &= (P(Q) - c_i) + q_i \cdot P'(Q) \cdot \frac{\partial Q}{\partial q_i} \\ &= (P(Q) - c_i) + q_i \cdot P'(Q) \cdot \underbrace{\left(1 + \frac{\partial Q_{-i}}{\partial q_i}\right)}_{\theta_i} \end{aligned}$$

We have that $\theta_i = 0$ corresponds to Bertrand competition (aggregate Q is unchanged). We have that $\theta_i = 1$ corresponds to the Cournot model. We have that $\theta_i = n$ corresponds to join profit maximization. This parametrization is great for applied theory as we can nest all of these models with a single parameter.

Some issues are with some θ_i we can justify nearly anything. Also, is it possible that for any θ_i there exist consistent conjectures (ie., suppose I require firm to actually respond in the way I believe they will.)? In fact, Daughety (1985) and Lind (1992) show that Cournot with $\theta_i = 0$ is the only consistent conjecture absent some knife-edge cases.

1.7 Concentration and Market Performance Spurious Regressions

Consider the following regression equation:

$$y_i = \beta HHI_i + X_i'\gamma + \epsilon_i$$

where i is an industry, y_i is the profit in industry i, HHI_i is the concentration index in industry i, X_i are characteristics about industry i that includes a constant, and ϵ_i is a 0 mean error term that we'll carefully reason about below.

The idea is that $\beta > 0$ means that an increase in market concentration results in higher profits (or prices).

We wonder whether we believe the regression assumption that $\mathbb{E}[\epsilon_i | HHI_i, X_i'] = 0$? We expect there to be unobservable characteristics that impact both HHI_i and y_i. For instance, we expect marginal cost c_i, which we don't observe that affect both concentration and profits. We can further argue that both HHI_i and y_i are simultaneously determined by some unobserved parameters including marginal cost. As a result, we expect the regression assumption to be not valid.

Consider the next regression equation:

$$y_{if} = \beta_1 HHI_i + \beta_2 s_{if} + X_i'\gamma + \epsilon_{if}$$

where i is and industry f is a firm, y_{if} is the profit of firm f in industry i, HHI_i is concentration in industry i, s_{if} is the market share of firm f in market i, X_i holds characteristics about industry i and a constant, and ϵ_{if} is a zero-mean error term that we'll reason about below.

We again wonder whether we believe the regression assumption that $\mathbb{E}[\epsilon_{if} | HHI_i, X_i, s_{if}] = 0$? Again, we expect marginal cost c_i, which is unobserved, affect both concentration and profits invalidating this assumption.

2 HOMOGENOUS PRODUCTS

2.1 Simple Simultaneous Equation Model (Working 1927)

Consider a simple model of supply and demand for coffee where everything is linear:

$$\begin{aligned} Q_t^d &= \alpha_0 + \alpha_1 P_t + U_t \\ Q_t^s &= \beta_0 + \beta_1 P_t + V_t \\ Q_t^d &= Q_t^s \end{aligned}$$

Setting $Q_t = Q_t^d = Q_t^s$, and solving for (P_t, Q_t), we get that

$$\begin{aligned} P_t &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{V_t - U_t}{\alpha_1 - \beta_1} \\ Q_t &= \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 V_t - \beta_1 U_t}{\alpha_1 - \beta_1} \end{aligned}$$

We observe that price and quantity are functions of *both* error terms and we cannot do anything to cancel things out. We can go further and compute a few covariances

$$\begin{aligned} \operatorname{Cov}(P_t, U_t) &= -\frac{\operatorname{Var}(U_t)}{\alpha_1 - \beta_1} \\ \operatorname{Cov}(P_t, V_t) &= \frac{\operatorname{Var}(V_t)}{\alpha_1 - \beta_1} \end{aligned}$$

We see that when demand is downwards sloping (ie., $\alpha_1 < 0$) and supply is upwards sloping (ie., $\beta_1 > 0$), then price is positively correlated with the demand shifter U_t an negatively correlated with the supply shifter V_t.

Using these covariances, we can compute the covariances between (Q_t^d, Q_t^s) and P_t.

$$\begin{aligned} &\operatorname{Cov}(P_t, Q_t^d) = \alpha_1 \operatorname{Var}(P_t) + \operatorname{Cov}(P_t, U_t) \\ &\operatorname{Cov}(P_t, Q_t^s) = \beta_1 \operatorname{Var}(P_t) + \operatorname{Cov}(P_t, V_t) \\ &\implies \\ &\text{Bias}(\alpha_1) = \frac{\operatorname{Cov}(P_t, U_t)}{\operatorname{Var}(P_t)} \\ &\text{Bias}(\beta_1) = \frac{\operatorname{Cov}(P_t, V_t)}{\operatorname{Var}(P_t)} \end{aligned}$$

We can go further and compute the plim of the coefficient on price in a regression of quantity on price:

$$\begin{aligned} \frac{\operatorname{Cov}(Q_t, P_t)}{\operatorname{Var}(P_t)} &= \frac{\operatorname{Cov}(\frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 V_t - \beta_1 U_t}{\alpha_1 - \beta_1}, \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{V_t - U_t}{\alpha_1 - \beta_1})}{\operatorname{Var}(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{V_t - U_t}{\alpha_1 - \beta_1})} \\ &= \frac{\alpha_1 \operatorname{Var}(V_t) + \beta_1 \operatorname{Var}(U_t)}{\operatorname{Var}(V_t) + \operatorname{Var}(U_t)} \end{aligned}$$

from which we clearly see that the naive OLS coefficient in a regression of quantity on price is meaningless. With higher $\mathrm{Var}(V_t)$, we have a better estimate of demand and with higher $\mathrm{Var}(U_t)$ we have a better estimate of supply. Since we assume demand to be downwards sloping and supply upwards sloping, we expect the coefficient estimated in this naive regression to be biased towards 0.

2.2 Simultaneity and Identification (Angrist, Imbens, and Graddy 2000)

Angrist, Imbens, and Graddy looked at the demand for Whiting (fish) at the Fulton Fish Market in downtown Manhattan. They sought to answer questions like "what does the linear IV regression of Q on P identify, even if the true (but unknown) demand function is nonlinear?" They took a program evaluation/ treatment effects approach to understand the "causal effect" of price on quantity demanded.

They considered 4 possible different demand systems:

1. Linear system with constant coefficients:

$$q_t^d(p,z,x) = \alpha_0 + \alpha_1 p + \alpha_2 z + \alpha_3 x + \epsilon_t$$
$$q_t^s(p,z,x) = \beta_0 + \beta_1 p + \beta_2 z + \beta_3 x + \eta_t$$

2. Linear system with non-constant coefficients:

$$q_t^d(p,z,x) = \alpha_{0t} + \alpha_{1t} p + \alpha_{2t} z + \alpha_{3t} x + \epsilon_t$$
$$q_t^s(p,z,x) = \beta_{0t} + \beta_{1t} p + \beta_{2t} z + \beta_{3t} x + \eta_t$$

3. Nonlinear system with constant shape (separable):

$$q_t^d(p,z,x) = q^d(p,z,x) + \epsilon_t$$
$$q_t^s(p,z,x) = q^s(p,z,x) + \eta_t$$

4. Nonlinear system with time-varying shape (non-separable):

$$q_t^d(p,z,x) = q^d(p,z,x,\epsilon_t)$$
$$q_t^s(p,z,x) = q^s(p,z,x,\eta_t)$$

The nonlinear models in 3 and 4 permit heterogeneity in effect depending on value of p. The models in 2 and 4 permit heterogeneity in the effect of price over time, for any given p.

2.2.1 Model 1 with a Binary Instrument $z_t \in \{0,1\}$

We assume that data arises from the first model in the enumeration and assume appropriate regularity conditions on $q_t^d, q_t^s, p_t, z_t, x$ (ie., finite second moments, and finite fourth moments if we want asymptotics).

We assume that z_t is a valid instrument for p_t in the demand equation. That means, that the instrument satisfies the *exclusion restriction* (ie., conditioning on p_t means that q_t^d is not impacted by z_t):

$$\forall p, z \; q_t^d(p, z=1, x_t) = q_t^d(p, z=0, x_t) \equiv q_t^d(p, x_t)$$

It also means that the instrument satisfies the *relevance condition* (ie., z_t actually shifts supply at some point in time):

$$\exists t : q_t^s(p_t, 1, x_t) \neq q_t^s(p_t, 0, x_t)$$

We take $z_t \in \{0, 1\}$ where 1 indicates "stormy at sea" and 0 denotes "calm at sea". The idea is that offshore weather makes fishing more difficult but doesn't change onshore demand.

Assuming we condition on each value of x, we can define

$$\begin{aligned}\hat{\alpha}_{1,0} &= \frac{\hat{\mathbb{E}}[q_t|z_t = 1] - \hat{\mathbb{E}}[q_t|z_t = 0]}{\hat{\mathbb{E}}[p_t|z_t = 1] - \hat{\mathbb{E}}[p_t|z_t = 0]} \\ &\xrightarrow{\mathbb{P}} \frac{\mathbb{E}[q_t|z_t = 1] - \mathbb{E}[q_t|z_t = 0]}{\mathbb{E}[p_t|z_t = 1] - \mathbb{E}[p_t|z_t = 0]} \\ &=: \alpha_{1,0}\end{aligned} \tag{2}$$

As we're assuming that our data is generated by Model 1, we've recovered a consistent estimate of $\alpha_1 = \alpha_{1,0}$.

2.2.2 *The Other Models with a Binary Instrument $z_t \in \{0, 1\}$*

If we wish to consider the fourth model, it will be useful to add an *independence* assumption:

$$\epsilon_t \perp\!\!\!\perp z_t | x_t$$

With this assumption we disallow any non-linear relationship between our instrument and the error in demand. In order to interpret the Wald Estimator of Equation (2), it is useful to make some additional economic assumptions on the structure of the problem. First, we assume that the observed price clears the market at all times:

$$q_t(p_t) = q_t^s(p_t) \; \forall t$$

Second we assume that there are unique "potential prices" for each value of z that clear the market

$$\forall z, t \; : \tilde{p}(z, t) \text{ is such that } q_t^d(\tilde{p}(z, t)) = q_t^s(\tilde{p}(z, t), z)$$

We also will make a *monotonicity* assumption: $\tilde{p}(z, t)$ is weakly increasing in z. This will rule out "*defiers*" here and allow us to interpret the average slope as $\alpha_{1,0}$ in the demand that is traced out by the movement in p when z moves from 0 to 1 amongst the compliers. This assumption is untestable as we don't observe both potential outcomes for all periods in time.

The key result establishes that then $\alpha_{1,0}$ can be written as **COULD PROBABLY PROVE THIS**:

$$\begin{aligned}\alpha_{1,0} &= \frac{\mathbb{E}[q_t|z_t = 1] - \mathbb{E}[q_t|z_t = 0]}{\mathbb{E}[p_t|z_t = 1] - \mathbb{E}[p_t|z_t = 0]} \\ &= \frac{\frac{1}{|\{t\}|}\sum_t \mathbb{E}_t[\int_{\tilde{p}(z=0,t)}^{\tilde{p}(z=1,t)} \frac{\partial q_t^d(s)}{\partial s}]}{\mathbb{E}[p_t|z_t = 1] - \mathbb{E}[p_t|z_t = 0]}\end{aligned}$$

The range $(\tilde{p}(z = 0, t), \tilde{p}(z = 1, t))$ could differ for each t and also be dependent on the instrument z used so that depending on which instrument we use we might get a different estimate of the demand slope.

I NEED MORE SHARP UNDERSTANDING HERE

3 DIFFERENTIATED PRODUCTS BERTRAND

We consider here a multi-product Bertrand Problem where a firm f has products $\mathcal{F}_f$ for which it can set the market prices. Firm f takes the market prices p as given to recursively solve:

$$\underset{\{p_j^f : j \in \mathcal{F}_f\}}{\arg\max}\ \pi_f(p^{-f}, p^f) = \underset{p^f \in \mathcal{F}_\{}{\arg\max} \sum_{j \in \mathcal{F}_f} (p_j^f - c_j) q_j(p^{-f}, p^f)$$

Abbreviating $p := (p^{-f}, p^f)$, we can take the first order condition with respect to a particular $p_j^f = p_j$:

$$\begin{aligned}
0 &= q_j(p) + \sum_{j \in \mathcal{F}_f} (p_j - c_j) \frac{\partial q_k}{\partial p_j}(p) \\
\implies p_j &= q_j(p) \left[-\frac{\partial q_j}{\partial p_j}(p) \right]^{-1} + c_j + \sum_{k \in \mathcal{F}_f \setminus \{j\}} (p_k - c_k) \underbrace{\left(\frac{\partial q_k}{\partial p_j}(p) \right) \left[-\frac{\partial q_j}{\partial p_j}(p) \right]^{-1}}_{=: D_{jk}(p)} \\
&= \frac{1}{1 + 1/\epsilon_{jj}(p)} \left[c_j + \sum_{k \in \mathcal{F}_f \setminus \{j\}} (p_k - c_k) D_{jk}(p) \right]
\end{aligned}$$

We call $D_{jk}(p) = \frac{\frac{\partial q_k}{\partial p_j}(p)}{\left| \frac{\partial q_j}{\partial p_j}(p) \right|}$ the diversion ratio. It helps explain what percentage of the demand that's lost in good j by increasing the price of good j will move from good j to good k.

There's intuition for this last expression in how to set price the price for good j. One sets a higher price if (i) the good is more inelastic. Secondly, one sets a higher price if (ii) the marginal cost of the good is higher. Lastly, one sets a higher price if (iii) the opportunity cost of raising your price and getting diversion to your other products is larger.

It is useful to define the matrix $\Delta(p)$ by

$$\Delta_{(j,k)}(p) := -\frac{\partial q_j}{\partial p_k}(p) \mathbb{1}_{\{j,k \in \mathcal{F}_f \text{ for some } f\}} \tag{3}$$

Once we do that, the first order conditions can be written in matrix form as

$$q(p) = \Delta(p)(p - c)$$

Then, we can recover marginal costs by rearranging to

$$c = p - \Delta(p)^{-1} q(p) \tag{4}$$

assuming that we have the market prices, market sales, and elasticities. One issue here is that if we have J products, then $\Delta(p)$ has J^2 entries, which might be a lot to recover given data, especially if we wish to have heterogeneous demand elasticities that depend on other covariates too. We might also prefer to think that consumers choose products in characteristic space rather than product space. Lastly, we might also wish to have heterogeneous consumers rather than representative agents.

3.1 Advantages of Diversion Ratios

When thinking about substitution between goods, two frequently pondered quantities are cross-elasticities and diversion ratios. They are written as, respectively:

$$\epsilon_{jk} = \frac{\partial q_k / q_k}{\partial p_j / p_j},\ D_{jk} = \frac{\partial q_k / \partial p_j}{|\partial q_j / \partial p_j|}$$

To see the issue with cross-elasticities, consider a good k where $\epsilon_{jk} = 0.01$ and another good k' where $\epsilon_{jk'} = 0.03$. The question is whether good k or good k' is a better substitute for j? The answer is that we don't know because it could be that the market share of good k is massive so that a 1% increase in the price of good j increasing the quantity of good k demanded by 1% is a lot more than increasing the quantity of good k' demanded by 3%. As a result, we would like to multiply the elasticity by the quantity demanded but if we do that and divide by price, then we're at something very close to the diversion ratio.

4 Representative Consumers

As an initial benchmark when considering demand, it's useful to imagine a representative consumer that chooses and expenditure level for each good and consumes at least a little of all goods. We hope that for a flexible matrix of demand derivatives $\Delta(p)$ that satisfy some axioms of consumer theory (eg., WARP) and other nice things (eg., $\Delta(p) \neq \Delta(p')$).

4.1 Constant Elasticity of Demand (CES)

One simple candidate model for demand is a *constant elasticity demand model*. We assume that there are a continuum of J goods indexed by ω. For $\rho \in (0, 1)$, we assume that a consumer that consumes $q(\omega)$ of $\omega \in [0, J]$ has utility

$$U(q) = \left(\int_0^J (q(w))^\rho d\omega \right)^{1/\rho}$$

We impose a budget constraint $\int_0^J p(\omega) q(\omega) d\omega$.

Using Lagrangians, we can solve for demands and demand ratios (define $\sigma := \frac{1}{1-\rho} \implies \rho = \frac{\sigma}{\sigma - 1}$):

$$\begin{aligned} q(\omega) &= \left(\frac{\lambda p(\omega)}{\rho} \right)^{\frac{1}{\rho - 1}} \\ \frac{q(\omega_1)}{q(\omega_2)} &= \left(\frac{p(\omega_1)}{p(\omega_2)} \right)^{\frac{1}{\rho - 1}} \\ \implies q(\omega_1) &= q(\omega_2) \left(\frac{p(\omega_1)}{p(\omega_2)} \right)^{-\sigma} \end{aligned}$$

Define the overall price index $P := \left(\int_0^J p(\omega_1)^{\frac{\rho}{\rho - 1}} d\omega_1 \right)^{\frac{\rho - 1}{\rho}} = \left(\int_0^J p(\omega_1)^{1-\sigma} d\omega_1 \right)^{\frac{1}{1-\sigma}}$. From there, we can look at the budget constraint and re-write the demand

$$
\begin{aligned}
I &= \int_0^J p(\omega_1)q(\omega_1)d\omega_1 \\
&= q(\omega_2)p(\omega_2)^\sigma \int_0^J p(\omega_1)^{1-\sigma} d\omega_1 \\
\implies q(\omega_2) &= \frac{I \cdot p(\omega_2)^{-\sigma}}{\int_0^J p(\omega_1)^{1-\sigma} d\omega_1} \\
&= \frac{I \cdot p(\omega_2)^{-\sigma}}{P^{1-\sigma}} \\
&= I \cdot p(\omega_2)^{-\sigma} P^{\sigma-1} \\
&= \left(\frac{p(\omega_2)}{P}\right)^{-\sigma} \left(\frac{I}{P}\right)
\end{aligned}
$$

We can continue to establish the well-known *homotheticity* property of CES by plugging back into the original equation for $U(q)$.

$$
\begin{aligned}
U(q) &= \left(\int_0^J q(\omega)^\rho\right)^{1/\rho} \\
&= \left(\int_0^J p(\omega)^{\frac{\rho}{\rho-1}} I^\rho P^{\frac{\rho^2}{\rho-1}}\right)^{1/\rho} \\
&= I \cdot P^{\frac{\rho}{1-\rho}} \left(\int_0^J p(\omega)^{\frac{\rho}{\rho-1}}\right)^{1/\rho} \\
&= I \cdot P^{\frac{\rho}{1-\rho}} P^{\frac{1}{\rho-1}} \\
&= \frac{I}{P}
\end{aligned}
$$

We see that utility is just consumption divided by the price index and that it just scales with I. Next, we can look at the demand elasticity for any good implied by this model:

$$
\begin{aligned}
\frac{\partial q(\omega)}{\partial p(\omega)} &= -\sigma p(\omega)^{-\sigma-1} P^{\sigma-1} I \\
\implies \frac{\partial q(\omega)/q(\omega)}{\partial p(\omega)/p(\omega)} &= -\frac{p(\omega)}{\sigma}
\end{aligned}
$$

which means that there's one markup (and elasticity) for all goods, which is not desirable in IO.

4.2 Almost Ideal Demand System (Deaton and Muellbauer 1980)

The key ideas of the Almost Ideal Demand System model (AIDS) are to allow for separable preferences and multi-stage budgeting. The idea is that we can allocate expenditure within a group without knowing what you choose within the group. Each group has an *index price* and demand of products in other index groups respond only to changes in the *index price*, not the individual prices.

We can define the expenditure function $e(u, p)$, where p are the prices and u is the target level of normalized utility[1]. We assume that there are K goods in a particular group. We can write the log expenditure for that group

[1] This utility is normalized in the sense that $u = 0$ corresponds to achieving the subsistence utility and $u = 1$ corresponds to reaching the bliss utility

$$\log(e(u,p)) = (1-u)\log(\underbrace{a(p)}_{\text{subsistenance}}) + u\log(\underbrace{b(p)}_{\text{bliss}})$$

We assume a particular functional form for $a(\cdot)$ and $b(\cdot)$ that is second-order flexible so that

$$\log(e(u,p)) = \alpha_0 + \sum_k \alpha_k \log(p_k) + \frac{1}{2}\sum_k \sum_j \gamma^*_{kj} \log(p_k)\log(p_j) + u\beta_0 \Pi_k p_k^{\beta_k}$$

We wish to estimate $\{(\alpha_i, \beta_i, \gamma_{ij})^*) : i,j \in \{1,...,K\}\}$ and usually impose that $\sum_k \alpha_k = 1$, for every $j \in \{1,...,K\}$, $\sum_k \gamma^*_{jk} = 0$, $\sum_k \beta_k = 0$ so that demand is linearly homogeneous in p (ie., $e(u,\lambda p) = \lambda e(u,p)\ \forall \lambda > 0$). In addition we often impose that $\gamma^*_{jk} = \gamma^*_{kj}\ \forall j,k \in \{1,...,K\}$.

We denote the expenditure share for good i as w_i and can use Shepard's Lemma to reach that $w_i = \frac{\partial \log(e(u,p))}{\partial \log(p_i)}$. Applied to our specific functional form and additionally defining $\gamma_{ij} := \frac{1}{2}(\gamma^*_{ij} + \gamma^*_{ji})$

$$w_i = \alpha_i + \sum_j \gamma_{ij} \log(p_j) + \beta_i u \beta_0 \Pi_k p_k^{\beta_k}$$

Defining the price index $\log(P) := \alpha_0 + \sum_k \alpha_k \log(p_k) + \frac{1}{2}\sum_j \sum_k \gamma^*_{kj} \log(p_k)\log(p_j)$, and calling $x := e(u,p)$, we note that $u\beta_0 \Pi_k p_k^{\beta_k} = \log(\frac{x}{P})$ so that

$$w_i = \alpha_i + \sum_j \gamma_{ij} \log(p_j) + \beta_i \log\left(\frac{x}{P}\right)$$

We still have K^2 possible elasticities to estimate, which can be a lot, even with just a single group. We can further sub-divide the goods into smaller groups but that requires knowledge of how to segment the market.

4.2.1 Beer Example (Hausman, Leonard, Zona 1994)

The authors thought about estimating demand at 3 different nested-levels.

We can consider a brand-level regression just within a specific segment (ie., all light beers) of beers. Indexing this regression type by (1), we can write the regression:

$$w_i = \alpha_0^{(1)} + \sum_j \alpha_{ij}^{(1)} \log(p_j) + \beta_i \log\left(\frac{x^{(1)}}{P^{(1)}}\right) + e_i^{(1)}$$

where $x^{(1)}$ is the expenditure in the segment and $P^{(1)}$ is the price index of the segment.

We can then consider a segment-level regression across beers. Indexing this regression type by (2), we can write the regression:

$$w_s = \alpha_0^{(2)} + \sum_t \alpha_{st}^{(2)} \log(p_j) + \beta_s^{(2)} \log\left(\frac{x^{(2)}}{P^{(2)}}\right) + e_s^{(2)}$$

where $x^{(2)}$ is the expenditure in the segment of peer and $P^{(2)}$ is the price index of beer.

We can lastly consider writing a market level regression across other liquid goods (eg., wine, spirits). The regression is

$$w_m = \alpha_0^{(3)} + \sum_n \alpha_{mn}^{(3)} \log(p_m) + \beta_m^{(3)} \log\left(\frac{x^{(3)}}{P^{(3)}}\right) + e_m^{(3)}$$

where $x^{(3)}$ is the expenditure in beer and $P^{(3)}$ is the price index of the liquids.

4.2.2 Challenges in Running these Regressions

The issue with running these regressions is that the price of a good can be correlated with both unobserved product quality and demand shocks.

Hausman, in the famous *Hausman instrument*, devised using prices in one city as an instrument for prices in another for the same brand. This instrument tends to have large relevance but the exclusion restriction is less believable. A firm can run national ad campaigns whereby a demand shock in the market of interest will be correlated with the price change in the city of interest and the price change in the other city.

5 MULTINOMIAL LOGIT

Most decisions agents make are not necessarily binary. For instance, agents choose a level of schooling (or a major), choose an occupation, choose a partner, and many more things.

We consider a *multinomial discrete choice* set up where in period t, and agent picks between $\mathcal{J}_t$ alternatives. Agent i in period t choose alternative $j \in \mathcal{J}_t$ with probability s_{ijt}. For ease of notation, we will define $J_t := |\mathcal{J}_t|$ as the number of choices an agent has. Agent i receives utility U_{ijt} for choosing option j. We assume an agent must make on of the choices and their choices are mutually exclusive.

We assume that we can additively separate the utility into an observed and systematic V_{ijt} and an unobserved and stochastic ϵ_{ijt}.

$$U_{ijt} = V_{ijt} + \epsilon_{ijt}$$

Focusing on just a single period and dropping the period t subscript, we can write

$$\begin{aligned} s_{ij} &= \Pr(U_{ij} > U_{ik} \; \forall k \neq j) \\ &= \Pr(V_{ij} + \epsilon_{ij} > V_{ik} + \epsilon_{ik} \; \forall k \neq j) \\ &= \Pr(\epsilon_{ij} - \epsilon_{ik} > V_{ik} - V_{ij} \; \forall k \neq j) \\ &= \int_{\mathbb{R}^J} \mathbb{1}_{\{\epsilon_{ij} - \epsilon_{ik} > V_{ik} - V_{ij} \; \forall k \neq j\}} f(\epsilon_i) \partial \epsilon_i \end{aligned}$$

where we define $f : \mathbb{R}^J \to \mathbb{R}$ as the density of individual i's unobserved utility. The integral is a J dimensional integral over each of the J dimensions of the unobserved stochastic component of utility.

Some common parametrizations include picking

- multinomial probit: $\epsilon_i \sim \mathcal{N}(0, \Omega)$ for some covariance matrix Ω.
- multinomial logit: $\epsilon_{ij} \overset{iid}{\sim}$ Gumbel$(\mu = 0, \beta = 1)$.
- There are also heteroskedastic variants of the multinomial logit framework.

We assume that f has full support and is continuous so that s_{ij} is continuously differentiable everywhere in V_{ij} and so that we can rationalize any pattern in the data.

5.1 Basic Identification

We note that only differences in utility matter (ie,. $\Pr(\epsilon_{ij} - \epsilon_{ik} > V_{ik} - V_{ij} \; \forall k \neq j)$). That means,

- (Constants and Individual Specific Constants) Adding a constant to utility is irrelevant since for any $a \in \mathbb{R}$, $U_{ij} > U_{ik} \iff U_{ij} + a > U_{ik} + a$. Further, adding individual specific factors that enter utility of all options such as income Y_i is irrelevant since $U_{ij} > U_{ik} \iff U_{ij} + Y_i > U_{ik} + Y_i$.
- (Alternative Specific Constants) Only differences in alternative specific constants can be identified. If $U_{ib} = v_{ib} + k_b + \epsilon_{ib}$ and $U_{ic} = v_{ic} + k_c + \epsilon_{ic}$, then only $k_b - k_c$ is identified. That means we can only include $J - 1$ such ks and need to normalize one to zero, say.
- (Scale) We note that $V_{ij} + \epsilon_{ij} > V_{ik} + \epsilon_{ik} \iff \lambda V_{ij} + \lambda\epsilon_{ij} > \lambda V_{ik} + \lambda\epsilon_{ik}$ for every $\lambda > 0$. Thus, multiplying by a constant factor of $\lambda > 0$ is also irrelevant to identification of V_{ij} (we would just estimate a scaled version of it). As a result, we typically set the variance of the unobserved error to $\frac{\pi^2}{6}$ which results from setting $\beta = 1$ in the parametrization of the Gumbel distribution. Suppose we further parametrize $V_{ij} = x'_{ij}\theta$. Then, we note that θ is only identified up to a scalar multiple of unobserved component of utility. As a result, if $||\theta||$ is larger, we can view it as there being less unobserved variance in individuals' discrete choice process.

5.2 Observed Heteroskedasticity

Consider the case where $\text{Var}(\epsilon_{ib}) = \sigma^2$ and $\text{Var}(\epsilon_{ic}) = k^2\sigma^2$ for some $k > 0$. Some people interpret this as meaning that for alternative c the unobserved factors are k times larger than for segment b.

5.3 Deriving the Market Shares for the Multinomial Logit Model

5.3.1 Theorem [Daly, Zachary, Williams]

Given V_i (the vector of systematic utilities for individual i) and f, the density of unobserved tastes, an interesting task is to predict the market shares. As before, we denote $s_{ij} := \Pr(V_{ij} + \epsilon_{ij} \geq \max_{k \in \mathcal{J}} U_{ik} + \epsilon_{ik})$ as the market share of option j.

Courtesy Daly, Zachary, and Williams, we know that $s_{ij} = \frac{\partial G(V_i)}{\partial V_{ij}}$ where $G(V_i) := \mathbb{E}_f \left[\max_{j \in \mathcal{J}} V_{ij} + \epsilon_{ij}\right]$.

Proof [Sketch].

Assume for some j that V_{ij} is increased by a small quantity δ. The welfare effect on the probability mass choosing j is $\pi_j(V_i) \times \delta$. The welfare effect on the probability mass not choosing j and are still not choosing j is 0. Finally, the welfare effect on that was not choosing j but switches to choose j is on the order of $\delta \times \delta$, for δ small enough given our assumption that f is continuous. Rearranging terms, we see the desired claim that $\pi_j(V_i) = \frac{\partial G(V_i)}{\partial V_{ij}}$. □

5.3.2 Lemma [Logit Inclusive Value]

Suppose we're given V_i (the vector of systematic utilities for individual i). Also assume that f is such that $\epsilon_{ij} \overset{iid}{\sim}$ Gumbel$(\mu = 0, \beta = 1)$. Then, we have that $\max_{j\in\mathcal{J}} V_{ij} + \epsilon_{ij} =_D \log(\sum_{j\in\mathcal{J}} \exp(V_{ij})) + \epsilon$ where ϵ follows the Gumbel$(\mu = 0, \beta = 1)$ distribution.

Proof.

Denote $Z := \max_{j\in\mathcal{J}} V_{ij} + \epsilon_{ij}$. We wish to show that the CDF of Z is the same as the CDF of $\log(\sum_{j\in\mathcal{J}} \exp(V_{ij})) + \epsilon$. Now,

$$\begin{aligned}
F_Z(z) &= \Pr(Z \leq z) \\
&= \Pr(\max_{j\in\mathcal{J}} V_{ij} + \epsilon_{ij} \leq z) \\
&= \Pr(V_{ij} + \epsilon_{ij} \leq z, \forall j \in \mathcal{J}) \\
&= \Pi_{j\in\mathcal{J}} \Pr(\epsilon_{ij} \leq z - V_{ij}) \\
&= \Pi_{j\in\mathcal{J}} \exp(-\exp(-z + V_{ij})) \\
&= \exp\left(-\sum_{y\in[0:Y]} \exp(-z)\exp(V_{ij})\right) \\
&= \exp\left(-\exp\left(-z + \log\left(\sum_{j\in\mathcal{J}} \exp(V_{ij})\right)\right)\right)
\end{aligned}$$

which is indeed the CDF of $\log\left(\sum_{j\in\mathcal{J}} \exp(V_{ij})\right) + \epsilon$ where ϵ has the Gumbel distribution. □

5.3.3 Consequence of the Lemma [Computing Expected Social Welfare]

We're now ready to compute $G(V_i) := \mathbb{E}_f[\max_{j\in\mathcal{J}} V_{ij} + \epsilon_{ij}]$ where f is such that $\epsilon_{ij} \overset{iid}{\sim}$ Gumbel$(\mu = 0, \beta = 1)$. We have that

$$\begin{aligned}
G(V_i) &= \mathbb{E}_f\left[\max_{j\in\mathcal{J}} V_{ij} + \epsilon_{ij}\right] \\
&= \mathbb{E}_{\text{Gumbel}(\mu=1,\beta=1)}\left[\log\left(\sum_{j\in\mathcal{J}} \exp(V_{ij})\right) + \epsilon\right] \\
&= \log\left(\sum_{j\in\mathcal{J}} \exp(V_{ij})\right) + \mathbb{E}_{\text{Gumbel}(\mu=1,\beta=1)}\left[\epsilon\right] \\
&= \log\left(\sum_{j\in\mathcal{J}} \exp(V_{ij})\right) + \gamma, \text{ where } \gamma \text{ is Euler's constant}
\end{aligned}$$

5.3.4 Deducing Market Shares

We're now ready to apply the DZW Theorem from Section 5.3.1 to deduce the market share maps when f is such that $\epsilon_{ij} \overset{iid}{\sim}$ Gumbel$(\mu = 0, \beta = 1)$.

$$\begin{aligned} s_{ij} &= \frac{\partial G(V_i)}{\partial V_{ij}} \\ &= \frac{\exp(V_{ij})}{\sum_{k \in \mathcal{J}} \exp(V_{ik})} \end{aligned}$$

5.4 IIA Property

As a result of the multinomial logit model, we observe that $\frac{s_{ij}}{s_{ik}} = \exp(V_{ij} - V_{ik})$, which is independent of any other alternative $l \in \mathcal{J}$. That means, if we introduce a new alternative with an independent unobserved utility shock, the logit model will predict that the ratio of the market share will remain the same.

As a striking critique, suppose that there are two commute options, a car c and a blue bus bb and we fit a multinomial logit model that predicts $\frac{s^o_{ibb}}{s^o_{ic}} = 1 \implies s^o_{ibb} = s^o_{ic} = \frac{1}{2}$, where the o stands for "original". Next, suppose we introduce a red bus rb that is identical to the blue bus otherwise so that we expect $\frac{s^n_{ibb}}{s^n_{irb}} = 1$ where the n stands for "new". If we carry over our predictions from the original setting to the new setting $\frac{s^o_{ibb}}{s^o_{ic}} = \frac{s^n_{ibb}}{s^n_{ic}} = 1$, we would predict that $s^n_{ibb} = s^n_{ibb} = s^n_{ic} = \frac{1}{3}$. That the fraction of people choosing to take a car is changing from $\frac{1}{2}$ to $\frac{1}{3}$ when there's been no material change to the commute options – that sounds like a fishy prediction.

Suppose that we parametrize $V_{ij} = x'_{ij}\theta$. Since the ratio of market shares of two choices depends only on the two choices themselves, the multinomial logit model implies that we should be able to recover the same parameter θ no matter the subset of alternatives that we choose to estimate on. There's a natural specification test where one looks at the estimated parameter $\hat{\theta}$ from choices on two different subsets of alternatives. If the parameter estimates are far apart, we reject the null of the multinomial logit model.

Suppose we let $\hat{\theta}^F_{N_F}$ be the parameter estimate using the full dataset. Suppose we $\hat{\theta}^R_{N_R}$ be the parameter estimates using the restricted datasets. Then, we have that if the model is well-specified as a multinomial logit,

$$H_{N_F,N_R} = (\hat{\theta}^F_{N_F} - \hat{\theta}^R_{N_R})'[\text{Var}(\hat{\theta}^R_{N_R}) - \text{Var}(\hat{\theta}^F_{N_F})]^{\dagger}(\hat{\theta}^F_{N_F} - \hat{\theta}^R_{N_R}) \sim \chi^2_{\dim(\theta)} \text{ as } N_F \to \infty$$

NEED TO FILL IN TEST STAT HERE IN TERMS OF RANDOMNESS IN N_R

5.4.1 Own and Cross Price Elasticities and Diversion Ratios

Suppose that $V_{ij} = x'_{ij}\theta$ and we continue to assume the multinomial logit model. Then, we have that $\frac{\partial V_{ij}}{\partial x^{(1)}_{ij}} = \theta^{(1)}$. Next, we note that $\frac{\partial s_{ij}}{\partial V_{ij}} = s_{ij}(1 - s_{ij})$.[2]

As a result, we have that,

$$\begin{aligned} \frac{\partial s_{ij}}{\partial x^{(1)}_{ij}} &= \frac{\partial s_{ij}}{\partial V_{ij}} \frac{\partial V_{ij}}{\partial \theta^{(1)}} \\ &= s_{ij}(1 - s_{ij})\theta^{(1)} \\ \implies \frac{\partial \log(s_{ij})}{\partial \log\left(x^{(1)}_{ij}\right)} &= \frac{\partial s_{ij}}{\partial x^{(1)}_{ij}} \frac{x^{(1)}_{ij}}{s_{ij}} \\ &= (1 - s_{ij})x_{ij}\theta^{(1)} \end{aligned} \tag{5}$$

[2]The (1) in the superscript means we're taking indexing into position 1 of that vector.

Similarly, for cross-effects, we can find that

$$\begin{aligned} \frac{\partial s_{ij}}{\partial x_{ik}^{(1)}} &= -s_{ij} s_{ik} \theta^{(1)} \\ \implies \frac{\partial \log(s_{ij})}{\partial \log\left(x_{ik}^{(1)}\right)} &= -s_{ik} x_{ik}^{(1)} \theta^{(1)} \end{aligned} \tag{6}$$

In a demand system, if $x_{ik}^{(1)}$ is a price, and we assume that as a result $\theta^{(1)} < 0$ since we assume that people are sensitive to price, we have the following implications.

- The price elasticity for a good is increasing in its own price and characterized by a single parameter $\theta^{(1)}$. This is not ideal as we expect different goods to have different shapes of elasticities and we don't expect elasticity to be increasing at all levels of price. For instance, if a good is very cheap, we expect people to be consuming as much as they want and many of the people to be very sensitive to the price. Meanwhile, when the price of the good is very expensive, we expect there to be the cohort of the population that's purchasing it to be relatively inelastic to the price.

- As the price of a good k rises, all other goods have their shares proportionally rise. This doesn't make sense as we expect certain goods to be more related than others so that an agent will likely substitute more towards a related good.

We can also consider properties of the implied diversion ratios.

$$\begin{aligned} D_{jk} &= \frac{\frac{\partial s_{ik}}{\partial x_{ij}^{(1)}}}{\left|\frac{\partial s_{ik}}{\partial x_{ij}^{(1)}}\right|} \\ &= \frac{\theta^{(1)} s_{ik} s_{ij}}{\theta^{(1)} s_{ij}(1 - s_{ik})} \\ &= \frac{s_{ik}}{1 - s_{ij}} \end{aligned} \tag{7}$$

We see that as the price of good j rises (ie., $x_{ij}^{(1)}$ increases), we proportionally inflate the likelihood of consuming substitutes. Further, at any level of price, we are predicting a constant diversion ratio for any pair of goods.

These properties of the multinomial logit seem undesirable as they don't allow us to capture the fact that different goods may face different elasticity term structures and different substitution patterns to other goods.

6 NESTED LOGIT

The multinomial logit model is generally criticized for its unrealistic and nonflexible substitution patterns. A generalization is to allow for a block structure on the covariance of the errors so that with a block alternatives have similar substitution patterns and across blocks they have different ones.

6.1 Preparation: A Decomposition Lemma

If $\epsilon \sim$ Gumbel$(\mu = 0, \beta = 1)$, then $\epsilon =_D \alpha\eta + \zeta$ where $\eta \perp\!\!\!\perp \zeta$ and $\eta \sim$ Gumbel$(\mu = 0, \beta = 1)$ where ζ is a to-be-determined random variable. This can introduce a correlation between different groups of random variables due to the persistence of the ζ in the draw. In fact, if

$$\begin{aligned}
\epsilon_1 := \alpha\eta_1 + \zeta,\ \epsilon_2 := \alpha\eta_2 + \zeta \quad &\text{where } \eta_1 \perp\!\!\!\perp \eta_2 \\
\implies \operatorname{Cov}(\epsilon_1, \epsilon_2) &= \operatorname{Var}(\zeta) \\
&= \operatorname{Var}(\epsilon_1) - \alpha^2 \operatorname{Var}(\eta_1) \\
&= \operatorname{Var}(\epsilon_1)(1-\alpha^2),\ \text{since } \epsilon_1 =_D \eta_1
\end{aligned}$$

That means that $\operatorname{Corr}(\epsilon_1, \epsilon_2) = 1 - \alpha^2$.

The question now is whether there exists a random variable η that leads to this decomposition. We can compute the CDF $F_\epsilon(\cdot)$ of ϵ defined by $\epsilon := \alpha\eta + \zeta$ where $\eta \perp\!\!\!\perp \zeta$ and $\eta \sim$ Gumbel$(\mu = 0, \beta = 1)$.

$$\begin{aligned}
F_\epsilon(u) &= \Pr(\alpha\eta + \zeta \leq u) \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{\eta \leq \frac{u-\zeta}{\alpha}\}} | \zeta\right]\right] \\
&= \mathbb{E}\left[\exp\left(-\exp\left(\frac{\zeta}{\alpha} - \frac{u}{\alpha}\right)\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\underbrace{\exp\left(\frac{\zeta}{\alpha}\right)}_{=:Z}\underbrace{\exp\left(-\frac{u}{\alpha}\right)}_{=:t}\right)\right] \\
&= \mathbb{E}\left[\exp(-tZ)\right]
\end{aligned}$$

That implies that if Z exists, its Laplace Transform at t should coincide with the CDF of ϵ at u. And, for $F_\epsilon(\cdot)$ to be the CDF of a Gumbel$(\mu = 0, \beta = 1)$ distribution, we require that $\mathbb{E}[\exp(-Zt)] = \exp(-t^\alpha) \implies \exp(-\exp(-u)) = F_\epsilon(u)$.

An appendix in Galichon's textbook on the mathematics of discrete choice models ensures the existence of the so-called positive stable distribution of parameter α, denote $\mathcal{PS}_\alpha$, which has the desired Laplace Transform. As a result, we can concisely state the following:

If $\eta \sim$ Gumbel$(\mu = 0, \beta = 1)$, $Z \sim \mathcal{PS}_\alpha$ for $\alpha \in (0, 1]$, and $\eta \perp\!\!\!\perp Z$, then $\alpha(\eta + \log(Z)) \sim$ Gumbel$(\mu = 0, \beta = 1)$.

6.2 On to the Model

For any alternative j and individual i, let $g_{ij} \in \mathcal{G}_i$ denote the nest that alternative j belongs to. We write that utility an individual i assigns to alternative j is given by

$$\begin{aligned}
U_{ij} &= V_{ij} + \epsilon_{ij} \\
&= V_{ij} + \lambda_{g_{ij}} \log(Z_{g_{ij}}) + \lambda_{g_{ij}} \eta_{ij}
\end{aligned}$$

where $\eta_{ij} \overset{iid}{\sim}$ Gumbel$(\mu = 0, \beta = 1)$ and $Z_g \overset{indep}{\sim} \mathcal{PS}_{\lambda_g}$ for each nest g, and V_{ij} is the systematic utility agent i associates to alternative j. The indirect utility of agent i is given by

$$
\begin{aligned}
u(\epsilon_i) &= \max_{j \in \mathcal{J}} \{V_{ij} + \epsilon_{ij}\} \\
&= \max_{g \in \mathcal{G}_i} \{\lambda_g \log(Z_g) + \max_{j \in g} \{V_{ij} + \lambda_g \eta_{ij}\}\} \\
&= \max_{g \in \mathcal{G}_i} \{\lambda_g \log(Z_g) + \lambda_g \log\left(\sum_{j \in g} \exp\left(\frac{V_{ij}}{\lambda_g}\right)\right) + \lambda_g \hat{\eta}_g\}, \text{ where } \hat{\eta}_g \overset{iid}{\sim} \text{Gumbel}(\mu = 0, \beta = 1) \\
&= \max_{g \in \mathcal{G}_i} \{\lambda_g \log\left(\sum_{j \in g} \exp\left(V_{ij}/\lambda_g\right)\right) + \hat{\epsilon}_g\}, \text{ by the lemma where } \hat{\epsilon}_g \sim \text{Gumbel}(\mu = 0, \beta = 1)
\end{aligned}
$$

This result shows that the nests are picked as in a logit model where the systematic utility associated with nest g is $\lambda_g \log\left(\sum_{j \in g} \exp\left(\frac{U_{ij}}{\lambda_g}\right)\right)$. We have that the expected utility and the market shares are given by

$$
\begin{aligned}
G(V_i) &= \log\left[\sum_{g \in \mathcal{G}_i} \left(\sum_{j \in g} \exp(V_{ij}/\lambda_g)\right)^{\lambda_g}\right] \\
s_{ij} &= \left(\frac{\left(\sum_{k \in g_j} \exp(V_{ik}/\lambda_{g_j})\right)^{\lambda_{g_j}}}{\sum_{h \in \mathcal{G}_i} \left(\sum_{k \in h} \exp(V_{ij}/\lambda_h)\right)^{\lambda_h}}\right)\left(\frac{\exp(V_{ij}/\lambda_{g_j})}{\sum_{k \in g_j} \exp(V_{ik}/\lambda_{g_j})}\right)
\end{aligned} \tag{8}
$$

6.3 Substitution Patterns

Within the same nest (ie., $j, k \in g$), we have IIA and proportional substitution patterns:

$$
\frac{s_{ij}}{s_{ik}} = \frac{\exp(V_{ij})}{\exp(V_{ik})}
$$

Across nests (ie., $j \in g, j \in h$ with $g \neq h$), we have:

$$
\frac{s_{ij}}{s_{ik}} = \frac{\exp(V_{ij}/\lambda_g)\left(\sum_{l \in g} \exp(V_{il}/\lambda_g)\right)^{\lambda_g - 1}}{\exp(V_{ik}/\lambda_h)\left(\sum_{l \in h} \exp(V_{il}/\lambda_h)\right)^{\lambda_h - 1}}
$$

6.4 Some Remarks on the Nested Logit Model

We note a few special cases of the nested logit model.

- If $\lambda_g = 1 \; \forall g$, then we reduce to the simple multinomial logit case.
- If $\lambda_g \to 0 \; \forall g$, then all consumers stay within the same nest.
- The econometrician needs to assign alternatives to nests before estimating the parameters of the model that will be assumptions underlying model estimation.

- You can have multiple levels of nesting. For instance, first an individual selects a car size (ie., midsize, full-sized, sedan, etc.), then they select a manufacturer, and finally they select a car.

- You can have overlapping nests. For instance, yogurt brands are one nest, yogurt flavors are another nest. This way, strawberry yogurt competes with strawberries and with plain yogurt.

7 MIXED LOGIT OR RANDOM COEFFICIENTS LOGIT

An alternative model with more flexible substitution patterns permits individuals to have heterogeneity in their sensitivity to different covariates. For instance, we can write either of the following:

$$\begin{aligned} U_{ij} &= \beta_i x_j + \epsilon_{ij}, \text{ where } \beta_i \sim f(\beta_i|\theta) \text{ and } \epsilon_{ij} \overset{iid}{\sim} \text{Gumbel}(\mu = 0, \beta = 1) \\ U_{ij} &= \beta x_{ij} + \underbrace{\nu_i z_j + \epsilon_{ij}}_{=:\tilde{\epsilon}_{ij}} \end{aligned}$$

The x_{ij} covariates are observed and in the second framework ν_i is unobserved (and often assumed to be normally distributed). These approaches allow for heteroskedasticity in ϵ_{ij} but only of the form that we can project onto z_j. In principle, this approach will help us get a better sense of substitution behavior between alternatives.

In the two models, we can compute the market shares as follows:

$$\begin{aligned} s_{ij}(\theta) &= \int_{\mathbb{R}} \frac{\exp(x_j\beta_i)}{1 + \sum_{k\in\mathcal{J}} \exp(x_k\beta_i)} f(\beta_i|\theta)\partial\beta_i \\ &\approx \sum_{s=1}^{S} w_i^s \frac{\exp(x_j\beta_i^s)}{1 + \sum_{k\in\mathcal{J}} \exp(x_k\beta_i^s)} \end{aligned}$$

The idea is that conditional on β_i (or ν_i in the second formulation), each person follows and IIA logit model. We integrate over such individuals giving us a *mixed logit* model. These integrals don't have closed forms and generally are approximated by using quadrature methods as written above.

7.1 Population Elasticities and Diversion Ratios

At the level of an individual, substitution patterns follow a plain logit. As a result, from the conclusions of Equations (5) and (6), we have that the price elasticities implied by the model are

$$\begin{aligned} \epsilon_{s_j,p_j} &= \frac{\partial s_j}{\partial p_j}\frac{p_j}{s_j} \\ &= \frac{p_j}{s_j}\int_{\mathbb{R}} \frac{\partial s_{ij}}{\partial p_j} f(\beta_i|\theta)d\beta_i \\ &= \frac{p_j}{s_j}\int_{\mathbb{R}} \beta_i s_{ij}(1 - s_{ij})d\beta_i \\ \epsilon_{s_j,p_k} &= \frac{\partial s_j}{\partial p_k}\frac{p_k}{s_j} \\ &= \frac{p_k}{s_j}\int_{\mathbb{R}} \frac{\partial s_{ij}}{\partial p_k} f(\beta_i|\theta)d\beta_i \\ &= \frac{p_k}{s_j}\int_{\mathbb{R}} -\beta_i s_{ij} s_{ik} f(\beta_i|\theta)d\beta_i \end{aligned}$$

Similarly, leveraging the diversion ratios implied by the simple logit model written in Equation (7), we see that the estimated population diversion ratio is

$$\begin{aligned} D_{jk} &= \int_{\mathbb{R}} \frac{s_{ij}}{s_j} D_{jk,i} f(\beta_i|\theta) d\beta_i \\ &= \int_{\mathbb{R}} \frac{s_{ij}}{s_j} \cdot \frac{s_{ik}}{1 - s_{ij}} f(\beta_i|\theta) d\beta_i \end{aligned}$$

where $s_j = \int_{\mathbb{R}} s_{ij} f(\beta_i \mid \theta) d\beta_i$.

7.2 Estimation Details

There are a couple of approaches to estimate the parameters θ of this model. First, let's define some notation. Suppose that we observe $\mathcal{D} := \{(x_{ij}, \hat{y}_{ij})\}_{i=1,j=1}^{I,J}$ where x_{ij} are the characteristics associated with alternative j for individual i and $\hat{y}_{ij} := \mathbb{1}_{\{i \text{ picks } j\}}$. We can write the likelihood of the data and the log-loglikelihood of the data given the parameters θ assuming this mixed logit model as

$$\begin{aligned} \mathcal{L}(\mathcal{D}|\theta) &:= \Pi_{i=1}^{I} \Pi_{j=1}^{J} (s_{ij}(\theta))^{\hat{y}_{ij}} \\ l(\mathcal{D}|\theta) &:= \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{y}_{ij} \log(s_{ij}(\theta)) \end{aligned}$$

7.2.1 Maximum Likelihood Esimator

The maximum likelihood estimator consists of solving the problem

$$\hat{\theta}_I^{MLE} := \arg\max_{\theta} l(\mathcal{D}|\theta)$$

The first order conditions are

$$0 = \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{y}_{ij} D_\theta \log(s_{ij}(\theta)) \tag{9}$$

One issue with this approach is that to evaluate the objective, we need to be able to evaluate $s_{ij}(\theta)$, which as state before generally doesn't have a closed form in the mixed logit case. Our integral approximation will be more consequentially bad when evaluating the log of small probabilities since that's where the $\log(\cdot)$ function has a steeper derivative. With an asymptotically biased evaluation of our objective and an asymptotically biased evaluation of its gradient, our parameter estimates will be inconsistent.

To see that we can't guarantee consistency of the (scaled) objective, let us reparametrize to let

$$\hat{s}_{ij}(\theta) := s_{ij}(\theta) + e_{ij},\ e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{10}$$

We numerically compute $\hat{s}_{ij}(\theta)$ when the true value of the integral is $s_{ij}(\theta)$. We maximize the objective

$$\begin{aligned}\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} l(\mathcal{D}|\theta) &= \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{y}_{ij}\log(\hat{s}_{ij}(\theta)) \\ &= \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{y}_{ij}\log(s_{ij}(\theta)+e_{ij})\end{aligned}$$

Due to the lack of additive separability of the $\log(\cdot)$ operator, we can't in the limit as $I \to \infty$ send the error contribution to 0 and guarantee consistency of our evaluation of the objective.

7.2.2 Method of Simulated Moments

Suppose we additionally observe some variables $\{z_{ij}\}_{i=1,j=1}^{I,J}$ that co-move with some of the x_{ij}s (and possibly are equal to them). If we believe that they satisfy the moment conditions:

$$\mathbb{E}[(\hat{y}_{ij} - s_{ij}(\theta))|z_{ij}] = 0$$

we can set up a GMM estimator to estimate θ.[3] We first define the moments $g_l(\theta) = \mathbb{E}[\underbrace{\sum_{j=1}^{J}(y_{ij} - s_{ij}(\theta))z_l(z_{ij})}_{=:h_l(\theta)}] = 0$, where $z_l(\cdot)$ is a function of the instruments that helps characterize the moment. We have that $l \in \{1, ..., L\}$ where $L \geq \dim(\theta)$ so that we can in fact identify θ.[4] We can construct the sample analogue to this quantity $\hat{g}_l^{(I)}(\theta) = \frac{1}{I}\sum_{(i,j)\in I\times J}(\hat{y}_{ij} - s_{ij}(\theta))z_l(z_{ij})$ and stack them into a vector $\hat{g}^{(I)}(\theta)$. From there, for a symmetric positive semidefinite weighting matrix W, we can define the GMM estimator

$$\hat{\theta}_I^{GMM} := \arg\min_{\theta} \frac{1}{2}\hat{g}^{(I)\prime}(\theta)' W \hat{g}^{(I)}(\theta)$$

The first order conditions match those of the MLE problem when $z(z_{ij}) = \frac{\partial \log(s_{ij}(\theta))}{\partial \theta}$. In fact, these are the optimal instruments in terms of Chamberlain (1987) and Amemiya (1988) in that the asymptotic variance of the parameter estimate hits the Cramer-Rao lower bound. To see that the FOC here matches that of the MLE problem, we compute the FOC with these instruments and rearrange:

[3] The expectation is over a data-generating process parametrized by the true θ where individuals are drawn from the distribution of heterogeneous unobserved characteristics and then make a decision with logit choice probabilities given by θ.

[4] We could have also defined the moment condition $\tilde{g}_l(\theta) = \mathbb{E}[(y_{ij} - s_{ij}(\theta))z_l(z_{ij})] = 0$. The key here is that this expectation is a bit misleading as there's no marginalization over j, we're only taking the expectation over i. This expectation is written for a fixed j. Thus, if $l \in \{1, ...L\}$, expressing this moment condition means there are LJ moments in the population. The sample analogue of each moment is written $\hat{\tilde{g}}_{jl}^{(I)}(\theta) = \frac{1}{I}\sum_{i=1}^{I}(\hat{y}_{ij} - s_{ij}(\theta))z_l(z_{ij})$. These moment conditions cannot lead to an equivalence with the MLE first order conditions as one will see below (one cannot sum over J).

$$
\begin{aligned}
& 0 = \hat{G}^{(I)}(\theta)' W \hat{g}(\theta) \\
\implies & 0 = \hat{g}(\theta), \text{ assuming } \hat{G}^{(I)}(\theta)' W \text{ is invertible} \\
& = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{y}_{ij} - s_{ij}(\theta)) D_\theta \log(s_{ij}(\theta)) \\
& = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{y}_{ij} D_\theta \log(s_{ij}(\theta)) - \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} s_{ij}(\theta) D_\theta \log(s_{ij}(\theta)) \\
& = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{y}_{ij} D_\theta \log(s_{ij}(\theta)) - \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \cancel{s_{ij}(\theta)} \frac{D_\theta s_{ij}(\theta)}{\cancel{s_{ij}(\theta)}} \\
& = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{y}_{ij} D_\theta \log(s_{ij}(\theta)) - \frac{1}{I} D_\theta \left(\sum_{i=1}^{I} \sum_{\cancel{j=1}}^{J} \cancel{s_{ij}(\theta)}^{1} \right) \\
& = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{y}_{ij} D_\theta \log(s_{ij}(\theta)), \text{ since the derivative of a constant is 0}
\end{aligned}
$$

which precisely matches the first order conditions of the MLE problem in Equation (9) (up to a scalar constant which is irrelevant for parameter identification).

All of this said, the optimal instruments will lead to the same inconsistency issues as in the MLE case: the integral approximations lead to inconsistencies in our moment conditions that lead to inconsistencies in our parameter estimates. To fix the inconsistencies, we can pick our instruments to be closed-form functions of the known covariates x_{ij} and in that case we will have a consistent estimator.

Choosing closed-form functions of known covariates as instruments, to achieve a consistent (and efficient[5]) estimator, we can do a 2-step procedure where we initially estimate $\hat{\theta}_I^{(1)}$ using $W = I$, then use that $\hat{\theta}_I^{(1)}$ to consistently estimate the optimal weighting matrix $\hat{W}_o^{-1} := \hat{\mathbb{E}}_I[(h(\hat{\theta}_I^{(1)}) - \hat{\mathbb{E}}[h(\hat{\theta}_I^{(1)})])(h(\hat{\theta}_I^{(1)}) - \hat{\mathbb{E}}[h(\hat{\theta}_I^{(1)})])']$. Then, we estimate θ again with GMM using that as the weighting matrix to get a feasible and efficient GMM estimator.

$$
\hat{\theta}_I^{GMM,\, 2Step} := \arg\min_\theta \frac{1}{2} \hat{g}^{(I)}(\theta)' \hat{W}_o \hat{g}^{(I)}(\theta)
$$

With Monte Carlo approximations to the integral, given that we we're not taking logs of an approximated integral quantity, we won't be making systematic errors when evaluating our integral so that when we drive the number of individuals to infinity, the approximation error to the integral will go to 0. As a result, we will have consistent estimates for the moment conditions and parameters. To see that let's reparametrize as in Equation (10). In the limit, our estimated moment condition will be,

$$
\begin{aligned}
\lim_{I \to \infty} \hat{g}_l^{(I)}(\theta) &= \lim_{I \to \infty} \frac{1}{I} \sum_{i,j \in I \times J} (\hat{y}_{ij} - \hat{s}_{ij}(\theta)) z_l(z_{ij}) \\
&= \lim_{I \to \infty} \frac{1}{I} \sum_{i,j \in I \times J} (\hat{y}_{ij} - s_{ij}(\theta)) z_l(z_{ij}) + \lim_{I \to \infty} \frac{1}{I} \sum_{i,j \in I \times J} (-e_{ij}) z_l(z_{ij}) \\
&\to \mathbb{E}[(\hat{y}_{ij} - s_{ij}(\theta)) z_l(z_{ij})] + 0, \text{ by LLN} \\
&= \mathbb{E}[(\hat{y}_{ij} - s_{ij}(\theta)) z_l(z_{ij})]
\end{aligned}
$$

[5]This estimator will be efficient within the class of estimators that uses the same moment conditions.

7.3 Machine Learning Enhancements- Embeddings

One drawback of random coefficients logit is that all heterogeneity in substitution patterns is projected down onto the characteristics we select. That means that if those characteristics do not model the substitution patterns well, then we're stuck in how to do better.

Van Der Maaten and Weinberger (2012) had an idea to apply the embeddings literature in Computer Science to learn representation for (latent) characteristics that help model substitution patterns well. Given survey data $\mathcal{D}$ of the form "select which two elements out of $\{j,k,l\}$ are most similar, for a hyperparameter $\alpha > 0$, they seek to find characteristics $x \in \mathbb{R}^m$ for each alternative $j \in \{1,...,J\}$ to solve

$$\max_{\{x^{(j)}\}_{j=1}^{J}} \sum_{(j,k,l)\in\mathcal{D}} \log\left(\frac{\left(1+\frac{||x^{(j)}-x^{(l)}||}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1+\frac{||x^{(j)}-x^{(l)}||}{\alpha}\right)^{-\frac{\alpha+1}{2}}+\left(1+\frac{||x^{(j)}-x^{(k)}||}{\alpha}\right)^{-\frac{\alpha+1}{2}}}\right)$$

where $(j,k,l) \in \mathcal{D}$ in this order means that j and l is the most similar pair.

Conlon, Mortimer, and Sarkis (2024) had another idea to non-parametrically model substitution patterns and infer mean utility. They assume they observe a matrix $\mathcal{D}$ of diversion ratios and another matrix of first choices $\mathcal{S}$. They seek to solve the problem

$$\min_{(S,\pi)\in\mathbb{R}_+^{J\times I+I}} ||\mathcal{D}-D(S,\pi)||_2 + \lambda\,||\mathcal{S}-S\pi||_2$$
$$\text{with } ||\pi||_1 \leq 1, ||s_i||_1 \leq 1\ \forall i \in \{1,...,I\}$$

for fixed hyperparameters α, I, λ. Out-of-sample, they wish to minimize $||\mathcal{D}-D(S,\pi)||_2$. The idea of this model is that they wish to identify types of individuals $i \in \{1,...,I\}$ and their prevalence in the population and assume their substitution patterns follow logit behavior so that the observed diversion ratios are close to the true diversion ratios. We have that given logit behavior, the diversion ratios are

$$D(S,\pi) = \frac{1}{\sum_{i=1}^{I}\pi_i s_i}\sum_{i=1}^{I}\pi_i s_i\left[\frac{s_i}{1-s_i}\right]'$$

where we divide each row by $\sum_{i=1}^{I}\pi_i s_i \in \mathbb{R}^J$. We note note if utility of agent i for option j is given by $U_{ij} = V_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \overset{iid}{\sim}$ Gumbel$(\mu=0,\beta=1)$, then $\log(s_{ij}) - \log(s_{i0}) = V_{ij} + V_{i0}$ so that in this model we identify the systematic utility of each alternative for each type of individual i relative to the outside good.

8 NUMERICAL OPTIMIZATION

Often, we are interested in solving problems of the following two forms:

- Root finding: $f(x) = 0$
- Optimization: $\arg\min_x f(x)$

These two problems are related because if the function f is convex we find the minimum by setting $f'(x) = 0$.

8.1 Newton's Method for Root Finding

Consider the Taylor series for $f(x)$ approximated around $f(x_0)$:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + f''(x_0)(x - x_0)^2 + o_p(x^3)$$

Suppose we seek x^* such that $f(x^*) = 0$. We can consider using just up to the linear terms in the equation above:

$$\begin{aligned} 0 &= f(x_0) + f'(x_0)(x_1 - x_0) \\ \implies x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} \end{aligned}$$

This motivates an iterative scheme to find x^*.

1. Start with some x_0.
2. Update using $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$
3. Stop when $|x_{k+1} - x_k| < \epsilon$

8.2 Newton's Method for Minimization

Suppose that we're trying to minimize $f : \mathbb{R}^K \to \mathbb{R}$ where f is known to be globally convex so that the Hessian of f, $D_{xx'}f$, is positive semi-definite, so that the second order condition is satisfied. We wish to solve

$$x^* := \arg\min_x f(x) \tag{11}$$

The optimum is characterized by the first order condition $D_x f(x) = 0$. We approximate $D_x f$ using a one level Taylor approximation around some initial x_0:

$$D_x f(x) \approx D_x f(x_0) + D_{xx'} f(x_0)(x - x_0) + o_p(x^2)$$

If we wish to find x^* so that $D_x f(x^*) = 0$, we can consider using just up to the linear terms in the expansion above:

$$\begin{aligned} 0 &= D_x f(x_0) + D_{xx'} f(x_0)(x_1 - x_0) \\ \implies x_1 &= x_0 - \left[D_{xx'} f(x_0)\right]^{-1} D_x f(x_0) \end{aligned}$$

This motivates an analogous scheme to find the optimum

1. Start with some x_0.
2. Update using $x_{k+1} = x_k - \left[D_{xx'} f(x_0)\right]^{-1} D_x f(x_0)$.
3. Stop when $||D_x f(x_k)||_2 < \epsilon$

8.3 Quasi Newton Methods

We can generalize this to Quasi-Newton methods that update parameters to solve problems like in Equation (11):

$$x_{k+1} = x_k - \lambda_k A_k D_x f(x_k)$$

where λ_k is the step length (ie., the learning rate), $d_k := A_k D_x f(x_k)$ is the step direction where we often rescale it to be of unit length.

- When $\lambda_k = 1$ and $A_k = [D_{xx'} f(x_0)]^{-1}$, this is a full Newton step.
- When $A_k = I_K \; \forall k \in \mathbb{N}$, this is gradient descent with a learning schedule of $\{\lambda_k : k \in \mathbb{N}\}$.
- When we're minimizing the negative log-likelihood as is relevant when solving an MLE problem, under the assumption of correct specification, we can exploit the Fisher information matrix identity. Specifically, if we're optimizing $\arg\min_\theta -l(\theta) = \arg\min_\theta \sum_{i=1}^n \log(f(x_i, \theta))$, we know that $-\mathbb{E}[D_{\theta\theta'} \log(f(X_i, \theta))] = \mathbb{E}[D_\theta \log(f(X_i, \theta))(D_\theta \log(f(X_i, \theta)))']$ $\forall \theta$. Thus, a sensible choice for $A_k := -\left[\frac{1}{n} \sum_{i=1}^n D_\theta f(x_i, \theta_k)(D_\theta f(x_i, \theta_k))'\right]$ is a sensible choice in that it's consistent approximation to the matrix recommended by the Newton step and it's less computationally expensive to compute.

Generally, the costly part of this optimization is computing and updating the Hessian matrix so these solvers use tricks to make that more efficient.

8.4 A Quick Note on Machine Learning (ML) Approaches

Inverting Hessian for models with a large number of parameters becomes quite infeasible numerically. Many ML models can be highly non-convex and there may be many potential local minima. Most models do some form of gradient descent (ie., $A_k = I_K$). In addition, many models to a randomly batched gradient update with a small learning rate so that (a) the gradient is less expensive to compute and (b) the model generalizes better to unseen data.

9 Demand – Aggregate Data and Endogeneity

There are I individuals (indexed by i) making choices from a set of J (indexed by j) alternatives, and an outside option given by index 0. We assume that individuals are *exchangeable* in the sense that any unobserved heterogeneity takes the form of an independent random draw from one preference distribution.

In this case, before drawing their unobserved preference shocks, ex-ante, they have the same choice probabilities $s_{ij} = s_j$. Suppose that $\hat{y}_{ij}$ is the choice made by individual i. Define $\hat{q}_j := \sum_{i=1}^I \hat{y}_{ij}$ to be the total number of individuals making choice j. We have that

$$(\hat{q}_1, ..., \hat{q}_J, \hat{q}_0) \sim \text{Multinomial}(I, s_1, ..., s_J, s_0)$$

If I becomes large enough, by a LLN, we have that

$$(\frac{\hat{q}_1}{I}, ..., \frac{\hat{q}_J}{I}, \frac{\hat{q}_0}{I}) \xrightarrow{\mathbb{P}} (s_1, ..., s_J, s_0)$$

Our goal is to find parameters θ that let us "equate" (as closely as possible) observed market shares $\frac{\hat{q}_1}{I}$ to model choice probabilities $s_j(x, \theta)$. Two challenges are that we (a) don't observe q_0 as it's hard to know how many people considered

picking one of the J options but decided against it.[6]. As an implication of not observing q_0, we also don't observe I. Also, with just data on a single market, it may be hard to tease out the kind of heterogeneity present in the underlying population for various product characteristics from unobserved taste preferences for the products.

9.1 Semiparametric Extension (Fox, Kim, Ryan, Bajari 2011)

Suppose we have data on $t \in \{1, ..., T\}$ markets where market t has I individuals making choices. We let $\hat{S}_{jt}$ be the observed market share of alternative j in market t. We also observe characteristics x_{jt} for alternative j in market t.

Fox, Kim, Ryan, Bajari (2011) present a semiparametetric approach where they can get a sense of the heterogeneity in in taste preferences for the underlying population making choices. There idea proceeds as follows:

1. Draw a large number (call it I) of β_i from a prior distribution $g(\beta_i)$.
2. Compute choice probabilities $s_{ijt}(\beta_i, x_t) = \frac{\exp(x'_{jt}\beta_i)}{1+\sum_{k=1}^{J}\exp(x'_{kt}\beta_i)}$.
3. Estimate the following contstrained least squares problem, which we note will give us sparse solutions.[7]

$$\min_{\pi \in \Delta^{I-1}} \sum_{j=1}^{J}\sum_{t=1}^{T}\left(\hat{S}_{jt} - \sum_{i=1}^{I}\pi_i s_{ijt}(\beta_i, x_t)\right)^2$$

9.2 Inversion in Multinomial Logit

Suppose that individual i's utility associated with alternative $j \in \{1, ..., J\}$ (with outside good indexed by 0) at time $t \in \{1, ..., T\}$ is given by

$$u_{ijt} = \underbrace{x'_{jt}\beta - \alpha p_{jt} + \xi_{jt}}_{=:\delta_{jt}} + \epsilon_{ijt}, \text{ and } s_{jt}(\delta_t) = \frac{\exp(\delta_{jt})}{1 + \sum_{k=1}^{J}\exp(\delta_{kt})}$$

Errors are distributed as in Section 5.3.4 so that we indeed have the multinomial logit choice probabilities. The idea is that ξ_{jt} is observed to the firm when prices are set but is not observed to us econometricians – consumers agree on this value. It reflects unobserved product quality that can be correlated with price $\text{Corr}_t(\xi_{jt}, p_{jt}) \neq 0$ but we operate under the assumption that $\mathbb{E}_t[\xi_{jt}|x_t] = 0$. The idea is that conditional on having the same characteristics, a car such as a BMW can still be "better" than a Peugeot that might lead to higher prices. The assumption that $\mathbb{E}_t[\xi_{jt}|x_t] = 0$ says that everything in this unobserved quality is mean independent (and as a result orthogonal) from the known characteristics of *any* of the alternatives.

Taking logs on both sides of the expression for $s_{jt}(\delta_t)$ above, we get that

$$\begin{aligned}
\log(s_{jt}(\delta_t)) &= \left[x'_{jt}\beta - \alpha p_{jt} + \xi_{jt}\right] - \log\left(1 + \sum_{k=1}^{J}\exp(x'_{kt}\beta - \alpha p_{kt} + \xi_{kt})\right) \\
\log(s_{0t}(\delta_t)) &= -\log\left(1 + \sum_{k=1}^{J}\exp(x'_{kt}\beta - \alpha p_{kt} + \xi_{kt})\right) \\
\implies \log(s_{jt}(\delta_t)) - \log(s_{0t}(\delta_t)) &= x'_{jt}\beta - \alpha p_{jt} + \xi_{jt}
\end{aligned}$$

[6]For instance, how can we know how many people walked down the cereal aisle and chose not to buy cereal? That is very challenging

[7]We note that this is a model with a LASSO penalty since we're requiring each π_i to be non-negative.

If we assume that we observe, $\log(\hat{s}_k)$ for $k \in \{0, 1, ..., J\}$, then we can think about running the regression

$$\log(\hat{s}_{jt}) - \log(\hat{s}_{0t}) = x'_{jt}\beta - \alpha p_{jt} + \xi_{jt}$$

with some appropriate estimating assumptions if we can get an appropriate instrument for p_{jt}.

9.3 Inversion in Nested Logit

We now consider a problem where individual i's the utility associated with alternative $j \in \{1, ..., J\}$ (with outside good indexed by 0) at time $t \in \{1, ..., T\}$ is given by

$$u_{ijt} = \underbrace{x'_{jt}\beta - \alpha p_{jt} + \xi_{jt}}_{=:\delta_{jt}} + \lambda_{g_j} Z_{g_j} + \lambda \eta_{ijt}, \text{ and } s_{jt}(\delta_t) = \left(\frac{(\sum_{k \in g_j} \exp(\delta_{jt}/\lambda_{g_j}))^{\lambda_{g_j}}}{1 + \sum_{h \in G} (\sum_{k \in h} \exp(\delta_{kt}/\lambda_h))^{\lambda_h}} \right) \left(\frac{\exp(\delta_{jt}/\lambda_{g_j})}{\sum_{k \in g_j} \exp(\delta_{kt}/\lambda_{g_j})} \right)$$

Errors are distributed as in Section 6.2 so that we indeed have the nested logit choice probabilities. Good j belongs to nest g_j and the outside good is assigned to its own nest. Once again, we assume that ξ_{jt} is unobserved to the econometrician, $\text{Corr}(\xi_{jt}, p_{jt}) \neq 0$, and $\mathbb{E}_t[\xi_{jt}|x_t] = 0$.

Then, we have that (where $\rho_g := 1 - \lambda_g$),

$$\log(s_{jt}(\delta_t)) - \log(s_{0t}(\delta_t)) = x'_{jt}\beta - \alpha p_{jt} + \rho_g \log(s_{jt|g}(\delta_t)) + \xi_{jt}$$

If we assume that we observe $\log(\hat{s}_{kt})$ for $k \in \{1, ..., J\}$ and $t \in \{1, ..., T\}$, then we can think about running the regression

$$\log(\hat{s}_{jt}) - \log(\hat{s}_{0t}) = x'_{jt}\beta - \alpha p_{jt} + \rho_g \log(\hat{s}_{jt|g}) + \xi_{jt}$$

if we can get an appropriate instrument for p_{jt} and $\log(\hat{s}_{jt|g})$.

9.4 Inversion in Random Coefficients Logit

We now consider a problem where individual i's utility associated with alternative $j \in \{1, ..., J\}$ (with outside good indexed by 0) at time $t \in \{1, ..., T\}$ is given by

$$u_{ijt} = \underbrace{x'_{jt}\beta - \alpha p_{jt} + \xi_{jt}}_{=:\delta_{jt}} + \mu_{ijt} + \epsilon_{ijt} \tag{12}$$

Generally, we assume that $\mu_{ijt} = x'_{jt}[\Pi y_i + \Sigma v_i]$ where $y_i \overset{iid}{\sim} \mathcal{N}(\mu_y, \Sigma_y)$ are unobserved demographics and $v_i \overset{iid}{\sim} \mathcal{N}(\mu_v, \Sigma_v)$ are unobserved heterogeneity. We label $\theta_2 := [\text{vec}_R(\Pi)', \text{vec}_R(\Sigma)', \mu'_v, \text{vec}_R(\Sigma_v)', \mu'_y, \text{vec}_R(\Sigma_y)']'$.[8] Making an independent Gumbel assumption on the error as in Section 7, we get that[9]

[8] We often assume that μ_y, Σ_y, $\mu_v = 0$, and $\Sigma_v = I$ reducing the parameters in θ_2 to $[\text{vec}_R(\Pi)', \text{vec}_R(\Sigma)']'$.

[9] Often, we choose to not include μ_y, μ_v in θ_2. That means they're estimated in the same step as α, β are estimated in Algorithm 1 below.

$$s_{jt}(\delta_{jt}, \theta_2) = \int \frac{\exp(\delta_{jt} + \mu_{ijt})}{1 + \sum_{k=1}^{J} \exp(\delta_{kt} + \mu_{ikt})} f(\mu_{ijt}|\theta_2) d\mu_{ijt}$$

For a fixed value of θ_2, we can consider the system of equations given by

$$\hat{s}_{jt} = s_{jt}(\delta_{jt}, \theta_2), \forall j, t$$

where the unknowns are δ_{jt} $\forall j, t$. The solutions at each t can be written as the solution to the following convex optimization problem

$$\min_{\delta_t \in \mathbb{R}^J} \sum_{s=1}^{S} w_s \log \left(1 + \sum_{j=1}^{J} \exp(\delta_{jt} + \mu_{ijt}^s(\theta_2)) \right) - \sum_{k=1}^{J} \delta_{kt} \hat{s}_{kt} \tag{13}$$

where w_s are Gauss-Hermite quadrature nodes that help evaluate the integral. Given the convexity of the problem in δ_t, we know it has a unique solution for δ_t.

9.5 BLP Pseudocode

Here, in Algorithm 1 we present pseudocode for the BLP algorithm under the random coefficient utility specification of Equation 12.

Algorithm 1 BLP Random Coefficients Coordinate Ascent

Input: : $\hat{s}_{jt}, x_{jt}, z_{jt}, p_{jt}, \theta_2^{(0)}, tol, S, W_1, W_2$
Output: : α, β, θ_2
Start Algorithm:
$i \leftarrow 0$
$\theta_2^{(i)} \leftarrow \theta_2^{(0)}$
$\theta_2^{(i-1)} \leftarrow ([\infty] * (\dim(\theta_2)))'$
while $||\theta_2^{(i)} - \theta_2^{(i-1)}|| \geq tol$ **do**
  $\delta_t^{(i)} \leftarrow \arg\min_{\delta_t \in \mathbb{R}^J} \sum_{s=1}^{S} w_s \log \left(1 + \sum_{j=1}^{J} \exp(\delta_{jt} + \mu_{ijt}^s(\theta_2)) \right) - \sum_{k=1}^{J} \delta_{kt} \hat{s}_{kt}$
  $g^{(i)} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J} (\delta_{jt} - x_{jt}'\beta + \alpha p_{jt}) z_{jt}$ # set up moment conditions
  $(\hat{\alpha}^{(i)}, \hat{\beta}^{(i)}) \leftarrow \arg\min_{\alpha,\beta} \frac{1}{2} g^{(i)\prime} W_1 g^{(i)}$ # can do 2-Step GMM
  $\hat{\xi}_{jt}^{(i)} \leftarrow \delta_{jt}^{(i)} - x_{jt}'\hat{\beta}^{(i)} - \hat{\alpha}^{(i)} p_{jt}$
  $h^{(i)} \leftarrow \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J} \hat{\xi}_{jt} z_{jt}$
  $\theta_2^{(i+1)} \leftarrow \arg\min_{\theta_2} \frac{1}{2} (h^{(i)})' W_2 (h^{(i)})$ # can do 2-Step GMM
  $i \leftarrow i + 1$
end while
return $\hat{\alpha}^{(i)}, \hat{\beta}^{(i)}, \theta_2^{(i)}$

9.6 Dube, Fox, and Su (2012)

With the advent of better computation, Dube, Fox, and Su (2012) devised an minimization problem (DFS) that estimates all of the parameters in one-shot.

$$\begin{aligned}
&\arg\min_{\theta_2,\alpha,\beta,\xi,\psi} \frac{1}{2}\psi' W \psi \\
&\text{s.t.} \quad \psi = Z'\xi \\
&\qquad \xi_{jt} = \delta_{jt} - x'_{jt}\beta - \alpha p_{jt} \\
&\log(\hat{s}_{jt}) = \log(s_{jt}(\theta_2, \delta))
\end{aligned}$$

This optimization problem is known as a mathematical problem with equilibrium constraints (MPEC). The unknown parameters in the BLP approach satisfy the same set of first order conditions as in the DBS approach, not only asymptotically but also in finite sample. Thus, we have that $\hat{\theta}_2^{BLP} \approx \hat{\theta}_2^{DBS}$, except for some numerical differences in the optimization routines. Thus, we should pick the routine that is more numerically convenient for our problem.

9.7 Advantages and Disadvantages of each Algorithm

The BLP Algorithm has the following set of advantages:

- In the outer optimization problem, we've concentrated out all of the linear in utility parameters so that we only need to search over θ_2 of reduced dimension.

- When T (ie., the number of markets is large), we can solve for δ_t in each market in parallel.

The BLP Algorithm has the following set of disadvantages:

- Small numerical errors in the inner loop can be amplified in the outer loop so that we must have very tight tolerance in the inner optimization problem.

- We need to take on some mathematical effort to compute a Jacobian with respect to θ_2 of the outer optimization via the implicit function theorem which can be a bit annoying.

The DFS Algorithm has the following set of advantages:

- The problem scales better in $\dim(\theta_2)$.

- Because all constraints hold at the optimum only, there's a smaller impact of numerical error in tolerance or numerical integration.

- The derivatives with respect to parameters in the problem are easier to compute than those of BLP for the outer maximization.

The DFS Algorithm has the following set of disadvantages:

- We are no longer concentrating out parameters so that the single optimization problem has a lot more of them.

- Parallelizing derivatives is tricker than in the BLP case.

10 ADDING SUPPLY

We will now aim to jointly estimate the parameters that govern demand and supply in these models. We name the parameters as follows:

- θ_1: The linear exogenous demand parameters.

- θ_2: The parameters including price and random coefficients (eg., endogenous and nonlinear parameters). We make a modification to the above definition to also include α in θ_2 as it will be a necessary parameter to identify markups.

- θ_3: The linear exogenous supply parameters.

Recall the first order conditions for the multi-product Bertrand problem discussed in Section 3. We had that after defining the matrix (as in Equation (3)),

$$\Delta_{(j,k)}(p) := -\frac{\partial q_j}{\partial p_k} \mathbb{1}_{\{j,k \in \mathcal{F}_f \text{for some } f\}}$$

the marginal costs can be recovered by

$$c = p - \underbrace{\Delta(p)^{-1} q(p)}_{=:\eta(p,q,\theta_2)}$$

as in Equation (4). Next, to recover marginal costs, we tend to assume a functional form for $c_{jt}(x_{jt}, \mathrm{w}_{jt}, w_j t)$, where x_{jt} are characteristics of the product and market that are important to supply and demand, w_{jt} are characteristics just important to supply, and w_{jt} is an error term, without restrictions at this point.

We specify marginal cost as some generalized linear model[10] with an additively separable error term

$$f(c_{jt}) = [x'_{jt}, \mathrm{w}_{jt}]'\theta_3 + w_{jt}$$

We make the exogeneity assumption $\mathbb{E}_t[w_{jt}|z_t^s] = 0$ where the expectation is taken over realizations of markets for the J goods.

10.1 Bringing Supply and Demand Together

To bring it together, we will write out Algorithm 2 to jointly estimate the parameters $(\theta_1, \theta_2, \theta_3)$.

[10] Generally $f(\cdot)$ is either the identiy function or $\log(\cdot)$.

Algorithm 2 Joint Supply and Demand Estimation

Input: : $\hat{s}_{jt}, x_{jt}, \mathrm{w}_{jt}, z^D_{jt}, z^S_{jt}, p_{jt}, \theta_2^{(0)}, tol, S, W_1, W_2, f(\cdot)$
Output: : $\beta, \theta_2, \theta_3$
Start Algorithm:
$i \leftarrow 0$
$\theta_2^{(i-1)} \leftarrow ([\infty] * (\dim(\theta_2)))'$
while $||\theta_2^{(i)} - \theta_2^{(i-1)}|| \geq tol$ **do**
  $\delta_t^{(i)} \leftarrow \arg\min_{\delta_t \in \mathbb{R}^J} \sum_{s=1}^S w_s \log\left(1 + \sum_{j=1}^J \exp(\delta_{jt} + \mu^s_{ijt}(\theta_2))\right) - \sum_{k=1}^J \delta_{kt}\hat{s}_{kt}$
  $g^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^T \sum_{j=1}^J (\delta_{jt} - x'_{jt}\beta + \theta_2^{(i)}[\alpha]p_{jt}) z^D_{jt}$ # set up demand moment conditions
  $\eta^{(i)} \leftarrow \eta(p, \hat{s}, \theta_2^{(i)}, \theta_2^{(i)}[\alpha])$
  $c_{jt}^{(i)} \leftarrow p_{jt} - \eta_{jt}^{(i)}$
  $l^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^T \sum_{j=1}^J (f(c_{jt}^{(i)}) - [x'_{jt}, \mathrm{w}'_{jt}]'\theta_3) z^S_{jt}$ # set up supply moment conditions
  $(\hat{\beta}^{(i)}, \hat{\theta}_3^{(i)}) \leftarrow \arg\min_{\beta,\theta_3} \frac{1}{2}[g^{(i)\prime}, l^{(i)\prime}] W_1 [g^{(i)\prime}, l^{(i)\prime}]'$ # can do 2-Step GMM
  $\hat{\xi}_{jt}^{(i)} \leftarrow \delta_{jt}^{(i)} - x'_{jt}\hat{\beta}^{(i)} - \theta_2^{(i)}[\alpha]p_{jt}$
  $\hat{w}_{jt}^{(i)} \leftarrow f(c_{jt}^{(i)}) - [x'_{jt}, \mathrm{w}'_{jt}]'\hat{\theta}_3^{(i)}$
  $h^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^T \sum_{j=1}^J \hat{\xi}_{jt} z^D_{jt}$
  $k^{(i)} \leftarrow \frac{1}{T}\sum_{t=1}^T \sum_{j=1}^J \hat{w}_{jt}^{(i)} z^S_{jt}$
  $\theta_2^{(i+1)} \leftarrow \arg\min_{\theta_2} \frac{1}{2}[h^{(i)\prime}, k^{(i)\prime}] W_2 [h^{(i)\prime}, k^{(i)\prime}]'$ # can do 2-Step GMM
  $i \leftarrow i + 1$
end while
return $\hat{\beta}^{(i)}, \theta_2^{(i)}, \hat{\theta}_3^{(i)}$

10.2 Instrument Selection

Recall the exogeneity assumption when estimating demand:

$$\begin{aligned} 0 &= \mathbb{E}_t\left[\xi_{jt} \mid x_t, z_t\right] \\ \implies 0 &= \mathbb{E}_t\left[f(x_t, z_t)\xi_{jt}\right] \end{aligned}$$

where z instruments for price. We can consider using any function $f(\cdot)$ such as $x^2_{jt}, x^3_{jt}, x_{jt}z_{jt}$ as a valid instrument. The question is how should we select $f(\cdot)$ and what are good choices for z?

To answer the question, we write out expanded versions of the equilibrium demand equations. We here add characteristics v_{jt} that only affect demand but not supply.

$$\begin{aligned} \delta_{jt}(\hat{s}_t, \theta_2) &= [x'_{jt}, \mathrm{v}'_{jt}]'\beta - \alpha p_{jt} + \xi_{jt} \\ f(c_{jt}) &= [x'_{jt}, \mathrm{w}'_{jt}]'\theta_3 + w_{jt} \end{aligned}$$

The first place to look for instruments is to look at something in the other equation. In other words, to instrument for price in the demand equation, we can use w_{jt} as we know it's correlated with the price that firms set but in principle is uncorrelated with the unobserved product-time error.[11]

[11] MacKay and Miller (2022) propose using $\mathrm{Cov}(\xi_{jt}, w_{jt}) = 0$ as the identification assumption but this rules out the idea that cars products that are unobservably expensive to produce also have unobservably high demand.

The equilibrium markup is obviously a function of *everything* (ie., $\eta_{jt}(p_t, \hat{s}_t, \xi_t, \omega_t, x_t, \mathrm{w}_t, \mathrm{v}_t, \theta_2, \mathrm{y}_t)$) and is endogenous. For markup shifters, we simply hope to make the best possible exclusion assumption that still moves the markup. We can consider using cross-market variation in the number or strength of competitor. For instance,

- We can pick w_{-jt} as a measure of the strength of the products of the other firms in the market.
- We can pick $\bar{x}_{-j_f t} := \frac{1}{|\mathcal{F}_j|-1} \sum_{k \in \mathcal{F}_j \setminus \{j\}} x_{kt}$ as a measure of the firm characteristics for other products it controls.
- We can pick $\frac{1}{|\mathcal{J} \setminus \mathcal{F}_j|} \sum_{k \in \mathcal{J} \setminus \mathcal{F}_j} x_{kt}$
- Higher order interactions between these terms.
- We don't want to pick p_{-j} as we often believe that firms are playing a simultaneous move game to set their prices in the market.

We also need instruments for the random coefficient parameters Σ that are part of θ_2. For these coefficients, Gandhi and Houde (2026?) propose using differentiation instruments. They define:

$$d_{jt}^k = |x_{kt} - x_{jt}|$$

They use the distances to define the following instruments:

$$\begin{aligned}
DIV_{j1} &:= \sum_{k \in \mathcal{J}} (d_{jt}^k)^2 \\
DIV_{j2} &:= \sum_{k \in \mathcal{J} \setminus \mathcal{F}_j} (d_{jt}^k)^2 \\
DIV_{j3}(c) &:= \sum_{k \in \mathcal{J}} \mathbb{1}_{\{d_{jt} < c\}} \\
DIV_{j4}(c) &:= \sum_{k \in \mathcal{J} \setminus \mathcal{F}_j} \mathbb{1}_{\{d_{jt} < c\}}
\end{aligned}$$

To identify Σ, we want to observe market shares respond to changes in market conditions for goods that are similar. For intuition, suppose that we have two goods j and k that are very similar along one characteristic and k has very small market share so that its deterministic utility is rather small. Suppose that we observe a super high increase in price for j and lots of people substitute from j to k. Then, we would determine that there's high variance in the heterogeneity for that characteristic as that's how one can explain the subtstitution.

10.3 Optimal Instruments

Chamberlain (1987) asks how we can choose instruments to obtain the semi-parametric efficiency bound with conditional moment restrictions in GMM. Suppose we have the conditional moment restriction,

$$\mathbb{E}[g(x_i, \theta_0) \mid z_i] = 0$$

That implies that $\mathbb{E}[f(z_i) g(x_i, \theta_0)] = 0 \, \forall f$. What f should we pick? Chamberlain (1987) shows the answer is to pick the instruments:

$$f(z_i) = \mathbb{E}\left[\frac{\partial g(x_i, \theta_0)}{\partial \theta} \mid z_i\right]' \mathbb{E}\left[g(x_i, \theta_0) g(x_i, \theta_0)' \mid z_i\right]^{-1}$$

10.3.1 Simple IV Problem

Consider the simplest IV problem:

$$\begin{aligned} y_i &= \beta x_i + \gamma v_i + u_i,\ \mathbb{E}[u_i \mid v_i, z_i] = 0 \\ g(y_i, x_i, v_i; \beta, \gamma) &:= y_i - \beta x_i, -\gamma v_i \end{aligned}$$

which gives that the optimal instruments are related to, ignoring the covariance term:

$$\begin{aligned} \mathbb{E}\left[\frac{\partial g(y_i, x_i, v_i; \beta, \gamma)}{\partial \beta} \mid v_i, z_i\right] &= -v_i \\ \mathbb{E}\left[\frac{\partial g(y_i, x_i, v_i; \beta, \gamma)}{\partial \gamma} \mid v_i, z_i\right] &= -\mathbb{E}\left[x_i \mid v_i, z_i\right] \end{aligned}$$

This looks a lot like the first stage of 2SLS. Note that nothing says that $\mathbb{E}\left[x_i \mid v_i, z_i\right]$ needs to be linear – we can fit a neural network to estimate the conditional mean.

10.3.2 Optimal IV in Two-Sided BLP

Recall that the conditional moment restrictions in GMM in BLP are given by

$$\begin{aligned} 0 &= \mathbb{E}[\xi_{jt} \mid z_t^D] \\ 0 &= \mathbb{E}[w_{jt} \mid z_t^S] \end{aligned}$$

Under the assumption that the demand instruments don't impact supply at all and that the supply instruments don't impact demand at all[12]

$$\underbrace{\mathbb{E}\left[\left(\frac{\partial \xi_{jt}}{\partial \theta}, \frac{\partial w_{jt}}{\partial \theta}\right) \mid z_t^D, z_t^S\right]}_{=:G'} \underbrace{\mathbb{E}\left[\left(\xi_{jt}, w_{jt}\right)\left(\xi_{jt}, w_{jt}\right)' \mid z_t^d, z_t^s\right]^{-1}}_{=:\Omega^{-1}}$$

In practice, to compute these instruments feasibly people

1. Fix $\hat{\theta} = (\hat{\theta}_1', \hat{\theta}_2', \hat{\theta}_3')'$ and draw (ξ, w) from the empirical density.
2. Solve the firms FOCs for $\hat{p}(\xi, w, \hat{\theta})$
3. Solve for the shares $s_t(x_t, \hat{p}_t, \hat{\theta})$

[12] This is believable. For instance, if we're using stormy seas to get exogenous variation in price to estimate demand, we don't believe that stormy seas will impact the pricing function in a way that's unobserved to supply:

4. Compute necessary Jacobians.

5. Average over values of sampled (ξ, w).

11 MICRO DATA

Data now is a lot better than they were back in 1995. It's now common to observe choices of some individuals, demographic characteristics such as the in Nielsen dataset.

11.1 Minimum Distance Pseudo Likelihood Estimator (MDPLE) Approach

Consider a setting where random utility is given by

$$u_{ijt} = \underbrace{x_{jt}'\beta - \alpha p_{jt} + \xi_{jt}}_{=:\delta_{jt}} + \mu_{ijt} + \epsilon_{ijt},\ \epsilon_{ijt} \overset{iid}{\sim} \text{Gumbel}(\mu = 0, \beta = 1)$$

We assume that $\mu_{ijt} = x_{jt}'\left[\Pi y_i + \Sigma v_i\right]$ where $y_i \overset{iid}{\sim} \mathcal{N}(\mu_y, \Sigma_y)$ are sometimes observed demographics and $v_i \overset{iid}{\sim} \mathcal{N}(\mu_v, \Sigma_v)$ are unobserved heterogeneity in preferences. We throw all of the parameters into a vector θ.

We assume that in each market $t \in \{1, ..., T\}$, N_t individuals make decisions following this utility model. If we let $\hat{s}_{ijt} := \mathbb{1}_{\{i \text{ picks } j \text{ in } t\}}$ be a dummy, we can define the aggregate market share of good j in market t as $\hat{s}_{jt} := \frac{1}{N_t}\sum_{i=1}^{N_t}\hat{s}_{ijt}$.

We assume that as a researcher, we view a subset S_t of the consumers in market t where we let $D_{it} := \mathbb{1}_{\{i \text{ is observed in } t\}}$. If we observe an individual i, then we observe there choice $\{\hat{s}_{ijt}\}_{j=1}^{J}$ for the individual and their demographics y_{it}.

Let $s_{ijt}(x_t, p_t, \theta \mid y_{it})$ be the individual choice probability assuming we know their demographics. Let $s_{jt}(x_t, p_t, \theta)$ be the aggregate choice probability. we can write the log-likelihood of the choices as

$$\begin{aligned} l(\theta, \delta) &:= \sum_{t=1}^{T}\sum_{j=0}^{J}\sum_{i=1}^{N_t}\hat{s}_{ijt}\left[D_{it}\log(s_{ijt}(x_t, p_t, \theta \mid y_{it})) + (1 - D_{it})\log(s_{jt}(x_t, p_t, \theta))\right] \\ &= \sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{i=1}^{N_t} D_{it}\hat{s}_{ijt}\log\left(\frac{s_{ijt}(x_t, p_t, \theta \mid y_{it})}{s_{jt}(x_t, p_t, \theta)}\right) + \sum_{t=1}^{T} N_t \sum_{j=1}^{J}\hat{s}_{jt}\log\left(s_{jt}(x_t, p_t, \theta)\right) \end{aligned}$$

Assuming that we have demand side instruments z_{jt}^d, we can construct moments to identify β. That is,

$$g(\theta, \delta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J} z_{jt}^D(\delta_{jt} - x_{jt}'\beta - \alpha p_{jt})$$

We can combine these moments with our data to form the objective

$$\min_{\theta,\delta} -l(\theta, \delta) + \frac{\lambda}{2}g(\theta, \delta)' W g(\theta, \delta)$$

for some hyperparameter $\lambda > 0$ that governs the tradeoff in the losses in case the GMM overidentifies β.

11.2 Micro BLP

Suppose that we have aggregate data generated market-by-market where markets are indexed by t. Each market has:

- We have products $j \in \mathcal{J}_t$ that have observed characteristics x_{jt} and unobserved quality ξ_{jt}
- There are consumer types $i \in \mathcal{I}_t$ that have *observed* demographics y_{it} and unobserved preferences v_{it}.
- We observe individual type market shares $\hat{s}_{ijt}$ and have knowledge of consumer weights w_{it} that help us form aggregate market share $s_{jt} = \sum_{i \in \mathcal{I}_t} w_{it} s_{ijt}$.

We have micro data that's generated dataset-by-dataset d conditional on aggregate data. The results are:

- $\{(t_n, j_n, y_{i_n t_n})\}_{n \in \mathcal{N}_d}$ from independent survey of selected consumers. In other words, we interview someone from a market, we observe their choice, and their demographics at the time of the market.
- Each consumer n is surveyed with known probability $w_{di_n j_n t_n}$

Often, we only have access to or are only willing to use summary statistics from the survey:

- Those are smooth functions $f(\bar{v}_d)$ of averages $\bar{v}_d = \frac{1}{N_d} \sum_n v_{di_n j_n t_n}$

For instance, we may have access to the survey result: "What is the mean age of the people who purchased this specific Nissan Car?" In this case, we know $w_{di_n j_n t_n} = 0$ for individuals who didn't purchase the car. We also know that $v_{dijt} = \text{age}(y_{it})$.

We can construct the GMM estimator:

$$\begin{aligned}
\hat{\theta} &= \arg\min_{\theta} \hat{g}(\theta)' W \hat{g}(\theta) \\
\hat{g}(\theta) &= [\hat{g}_A(\theta)', \hat{g}_A(\theta)']' \\
\hat{g}_A(\theta) &= \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{T} (\hat{\delta}_{jt}(\theta) - x'_{jt}\beta) z^D_{jt} \\
\hat{g}^{(k)}_M &= f_k(\bar{v}) - f_k(v(\theta)) \text{ for } k \in \{1, ..., L\}
\end{aligned}$$

With the share constraint still giving the mean utility of each alternative in each market:

$$\hat{s}_{jt} = \sum_{i \in \mathcal{I}_t} w_{it} s_{ijt}(\hat{\delta}, \theta \mid y_{it})$$

If moment k is like the example above about the Nissan car, the model predicted moment will be computed as (where j is the Nissan car and t is the market in which the survey d takes place):

$$f_k(v(\theta)) = \frac{\sum_{i \in I_t} w_{it} s_{ijt}(x_t, p_t, \theta | y_{it}) \text{age}(y_{it}) w_{dijt}}{\sum_{i \in \mathcal{I}_t} w_{it} s_{ijt}(x_t, p_t, \theta | y_{it}) w_{dijt}}$$

12 ANTITRUST IN HORIZONTAL MARKETS

In anti-trust, historically there's been a debate about whether governments should maximize social surplus or just focus on consumer surplus. Today, debates often center on political power, "bigness", and profit of small firms.

12.1 Sherman Act (1890)

- Section 1: "Every contract, combination in the form of trust or otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations, is declared to be illegal"

- Section 2: "Every person who shall monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce among several states, or with foreign nations, shall be deemed guilty of a felony"

The two sections deem illegal price fixing, horizontal market division (ie., you sell in this region and I'll sell in this other region), and refusals to deal to harm competition.

12.2 Clayton Act (1914)

- Section 2: Prohibits some forms of price discriminatino, but only when it lessens competition.

- Section 3: Prohibits sales based on the condition that the buyer not buy from your competitor (includes tying and exclusive dealing), but only when the effect may be to substantially lessen competition.

- Section 7: Prohibits mergers where the effect of such acquisition may be to substantially lessen competition, or tend to create a monopoly in any line of commerce.

- Section 8: Prevents a person from being a director of multiple competing firms. Now there's a question about whether VC companies that have different employees on the boards of different competiting firms is okay.

12.3 Horizontal Antitrust DOJ/FTC Outline

The DOJ and FTC take the following steps to evaluate a merger.

1. They define the market. This will be described more below.

2. The measure the concentration of the market recalling the definition of HHI in Section 1.5. The DOJ can ask for divestitures as a remedy if there are a few problematic markets in an otherwise uncontroversial merger. Additional discussion here will be defered to Section 1.5.

3. Simulate the merger and see what happens to welfare and its decomposition.

12.4 Market Definition

This step is often the most contentious step in the antitrust process. For instance:

- Merging parties wish to argue that the market is large. When Whole Foods and Wild Oats merged in 2007, they argued that they competed with Walmart and Target so that their importance in the market was small. The DOJ ended up breaking up the merger in 2009 due to reduced competition in the organic foods space.

- A monopolist may argue that its product is substitutable with others such as in US v. Dupont (1956). This case gave rise to the *Cellophane Fallacy* whereby economists noted that aluminum foil was a substitute for cellophane at the monopoly price (set by Dupont) but would not have been if the cellophane market was competitive.

Courts often rely on the *hypothetical monopolist test* to define a market:

1. Start with a candidate market such as the market for Coca-Cola.

2. If a monopolist controlled Coca-Cola only, could it raise prices by 5-10% without losing too many customers and profitably. If not, then expand firm to another product too such as Pepsi in our example.

3. Keep expanding the firm until it can profitably raise prices by 5-10%. At that point, the market is defined as all products the firm controls.

There are many problems with the *hypothetical monopolist test*:

- To avoid the Cellophane Fallacy, we'd like to ask how substitutable products are at competitive prices (which we don't see) as opposed to at observed market prices.

- The order in which we add products to the hypothetical monopolist may change the market definition. We could think about removing products from the hypothetical monopolist to see what is the smallest hypothetical monopolist that could sustainably increase prices. Also, in reality in our example, Coca-Cola owns many products (eg., Coke, Sprite, Diet Coke, Powerade) and we should maybe take that into account. Do we add firms or products to the hypothetical monopolist?

- Two different people can pose candidate markets that are constructed by the hypothetical monopolist test and overlap very little.

12.4.1 Aggregate Diversion Ratio

For any candidate market $\mathcal{M}$, one can define the aggregate diversion ratio:

$$D_{j\to 0}^{\mathcal{M}} = 1 - \frac{\sum_{k\in\mathcal{M}} \frac{\partial q_k}{\partial p_j}}{\frac{\partial q_j}{\partial p_j}}$$

The idea here is that we're looking at what percentage of the market leaves by increasing the price of a good. If there's high diversion to the outside good, it means everyone leaves the market and maybe the market definition is too small. Meanwhile, if there's low diversion to the outside good, that means the market recaptures customers after to a price increase of one good. One can use this quantity to help define the market and to know if a hypothetical monopolist in the market could increase prices by 5-10%.

That said, if one has access to diversion ratios, one can more directly think about whether a new ownership structure will let a firm raise prices 5-10% above the perfectly competitive level in the new equilibrium as in Section 12.6.1 and avoid a "market definition" step all together.

12.5 Upwards Pricing Pressure

Recall the result of the first order condition in the differentiated products Bertrand game of Section 3.

$$p_j = \frac{1}{1 + 1/\epsilon_{jj}(p)} \left[c_j + \sum_{k\in\mathcal{F}_f\setminus\{j\}} (p_k - c_k) D_{jk}(p) \right]$$

where $D_{jk}(p)$ is the diversion ratio between product j and k at the observed market prices. If firm f acquires firm g that controls products $\mathcal{F}_g$, we define the upwards pricing pressure of the acquisition on product j as

$$UPP_j = \Delta c_j + \sum_{k \in \mathcal{F}_g} (p_k - c_k) D_{jk}(p)$$

We evaluate the diversion ratios $D_{jk}(p)$ at the old market prices and think about how the acquisition changes the opportunity cost of increasing the price of the good as more diversion is reclaimed.

12.6 Mergers and Counterfactual Prices

Recall the differentiated products Bertrand setup in Section 3. We had that when defining the ownership-elasticity matrix as in Equation (3) with typical element

$$\Delta_{(j,k)}(p) = \frac{\partial q_j}{\partial p_k} \mathbb{1}_{\{j,k \in \mathcal{F}_f \text{ for some } f\}}$$

We can write the first order condition of this game in matrix form as

$$\begin{aligned} q(p) &= \Delta(p) \cdot (p - c) \\ \implies c &= p - \Delta(p)^{-1} q(p) \end{aligned}$$

A merger changes the ownership-elasticity matrix by changing the results of the indicator random variables since different products will be now controlled by the same firm.

12.6.1 Simulating a Merger

To fully simulate a merger, one can take the following steps under the assumption that firms compete under the differentiated products Bertrand game:

1. Recover the marginal costs as $c := p - \Delta(p)^{-1} q(p)$.
2. Adjust (possibly) marginal cost to reset it as $\tilde{c}_j = c_j \cdot (1 - e)$ for products j where there may be some sort of efficiency gain that's measured by e.
3. Change the ownership matrix $\Delta^{pre}(p) \to \Delta^{post}(p)$.
4. Solve for p^{post} via $p = \tilde{c} - \Delta(p)^{-1} q(p)$. This step is very tricky because we have to solve an implicit system of equations with p on both sides.

To partially simulate a merger, one can take the following steps under the same competition assumption:

1. Hold all irrelevant prices fixed at the pre-merger prices.
2. Adjust the marginal costs for potential efficiencies as in Step 2 above.
3. Solve for the new prices given the change in the products controlled by the merging firm. In particular, if we are just imagining a firm acquiring a single firm that produces a single product, then we just need to solve a single equation.

12.6.2 Solution Methods for Full Merger Analysis

When solving the merger analysis, we're interested in a specialization of the more general problem. Consider finding a root of $F(x) = 0$ where F has m nonlinear equations in $x \in \mathbb{R}^m$ unknowns. Expanding, we wish to solve:

$$
\begin{aligned}
0 &= F_1(x_1, ..., x_m) \\
0 &= F_2(x_1, ..., x_m) \\
&... \\
0 &= F_m(x_1, ..., x_m)
\end{aligned}
$$

12.6.2.1 Gauss-Jacobi: Simultaneous Best Reply The Gauss-Jacobi method is an iterative scheme to solve this problem that attempts to compute a simultaneous best reply to the current iterate. To be more precise, suppose the current iterate is $x^n = (x_1^n, ..., x_m^n)$. We compute the next iterate x^{n+1} by solving one equation in one variable using only values from x^n:

$$
\begin{aligned}
0 &= F_1(x_1^{n+1}, x_2^n, ..., x_m^n) \\
0 &= F_2(x_1^n, x_2^{n+1}, ..., x_m^n) \\
&... \\
0 &= F_m(x_1^n, x_2^n, ..., x_m^{n+1})
\end{aligned}
$$

12.6.2.2 Gauss-Seidel: Iterated Best Response The Gauss-Seidel method is an iterative scheme to solve this problem that computes a one-by-one iterative best reply to the current iterate. To be more precise, suppose the current iterate is $x^n = (x_1^n, ..., x_m^n)$. We compute the next iterate x^{n+1} by solving one equation in one variable and updating as we go:

$$
\begin{aligned}
0 &= F_1(x_1^{n+1}, x_2^n, ..., x_m^n) \\
0 &= F_2(x_1^{n+1}, x_2^{n+1}, ..., x_m^n) \\
&... \\
0 &= F_m(x_1^{n+1}, x_2^{n+1}, ..., x_m^{n+1})
\end{aligned}
$$

This one can sometimes be sped up by reordering equations. For instance, it makes sense to put the merging parties at the top.

12.6.2.3 Newton-Raphson The Newton-Raphson method is also an iterative scheme to solve this problem. It goes as follows:

1. Take an initial guess x^0
2. Take a Newton-Step by $x^{n+1} = x^n - [D_x F(x^n)]^{-1} F(x^n)$ analogously that to in Section 8.2.
3. Stop when $||F(X^n)||_2 < ftol$ or $||x^{n+1} - x^n||_2 < rtol$.

Often computing $[D_x F(x^n)]^{-1}$ is hard as it requires computing the derivative of the markup with respect to prices, which is a second derivative of demand with respect to prices. That said, this is the recommended method for general problems.

12.6.2.4 Exploiting the Logit Formula For the mixed logit utility specification, the price-derivative-ownership matrix becomes

$$
\Delta_{(j,k)}(p) = \begin{cases} \int \alpha_i s_{ij}(1 - s_{ij}) \partial F_i & \text{if } j = k \\ -\left(\mathbb{1}_{\{j,k \in \mathcal{F}_f \text{ for some} f\}}\right) \int \alpha_i s_{ij} s_{ik} \partial F_i & \text{if } j \neq k \text{ (ie., otherwise)} \end{cases}
$$

For the plain logit utility specification, assuming one firm owns all the products,[13] and defining $s(p)$ as the vector of market share, we can factor that into

$$\Delta(p) = \underbrace{diag(\alpha s(p))}_{=:\Lambda(p)} - \underbrace{\alpha s(p) s(p)'}_{=:\Gamma(p)}$$

That implies that

$$\begin{aligned} p - c &= \Delta^{-1}(p) s(p) \\ &= (\Lambda(p) - \Gamma(p))^{-1} s(p) \\ \implies (\Lambda(p) - \Gamma(p))(p - c) &= s(p) \\ \implies (p - c) &= \Lambda(p)^{-1}\Gamma(p) \cdot (p - c) + \Lambda(p)^{-1} s(p) \end{aligned}$$

One can solve for p by initially updating c per the post-merger marginal cost update and then iterate on $p - c$, keeping c fixed. The iterative scheme

$$(p - c)^{(n+1)} = \Lambda\left(p^{(n)}\right)^{-1} \Gamma\left(p^{(n)}\right) \cdot (p - c)^{(n)} + \Lambda\left(p^{(n)}\right)^{-1} s(p^{(n)})$$

is very fast and reliable.

13 CONDUCT

A foundational question in Industrial Organization is: how do we observe data on price and quantity to infer which model of firm behavior generated those outcomes?

We begin with the assumption that we have a relatively standard BLP-style differentiated products setup:

- We have data on markets indexed by t
- The markets have products indexed by j
- We have product market characteristics $\mathcal{X}_t := \{x_{jt}, \mathrm{v}_{jt}, \mathrm{w}_{jt} : j \in \mathcal{J}_t\}$ where x_{jt} are characteristics of the product and market that affect both supply and demand, w_{jt} are characteristics of the product and market that affect only supply, and v_{jt} are characteristics of the product and market that affect only demand.
- We as econometricians observe market shares $\{\hat{s}_{jt} : j \in \mathcal{J}_\sqcup \cup \{0\}\}$.
- We as econometricians observe market prices p_t
- We have information on the demographics present in market t given by y_t

We consider a variant of the differentiated products Bertrand game of Section 3 that's characterized by a conduct parameter κ. The value $\kappa_{fg} \in [0, 1]$ defines how much firm f cares about firm g's profits. If firm f and firm g merge, then we write that $\kappa_{fg} = 1$. Firm f solves:

[13] Note that it's not a hard extension to look at a case where there's a more interesting ownership structure. One just needs to appropriately use a Hadamard product to decompose the price derivative part from the ownership part.

$$\arg\max_{p^f \in \mathcal{F}_\{} \sum_{j \in \mathcal{F}_f} (p_j^f - c_j) q_j(p^{-f}, p^f) + \kappa_{fg} \sum_{j \in \mathcal{F}_g} (p_j - c_j) q_j(p^{-f}, p_f)$$

The first order condition with respect to price p_j set by firm f is

$$0 = q_j(p) + \sum_{k \in \mathcal{F} \cup \mathcal{F}_g} \kappa_{fg}(p_k - c_k) \frac{\partial q_k}{\partial p_j}(p)$$

It is useful to define the matrix $\Delta(p)$ by

$$\Delta_{(j,k)}(p, \kappa) := -\frac{\partial q_j}{\partial p_k}(p) k_{f_j g_j}$$

Once we do that, we can write the first order conditions in matrix form as:

$$\begin{aligned} q(p) &= \Delta(p, \kappa)(p - c) \\ \implies c &= p - \underbrace{\Delta(p, \kappa)^{-1} q(p)}_{=:\eta(p,q,\kappa)} \end{aligned}$$

13.1 Reasons for Deviation from Static Bertrand Game

We note that $\kappa_{fg} > 0$ is not necessarily evidence of malfeasance between firm f and firm g – it's just evidence of a deviation from *static Bertrand pricing*. Here are some possible explanations for deviations:

- For any work, one needs to have good estimates of the cross price derivatives. If these are biased, than all conduct conclusions will be biased as well.

- Often, it's a combination of the producing firm and retailers that set supermarket prices. Retailers may soften downstream competition.

- The Bertrand game is a simultaneous move one-shot pricing game. There are lots of alternatives such as Stackelberg leader-follower.

- Forward looking firms may not set static Nash prices. They may temporarily create sales so as to attract more customers.

- There may be an unmodeled supergame where firms are legally and tacitly colluding to set higher prices.

13.2 Towards a Framework for Testing Conduct

We can consider a BLP like model for demand and supply with structural error terms:

$$\begin{aligned} \delta_{jt}(\hat{s}_t, \theta_2) &= h_d(x_{jt}, \mathrm{v}_{jt}, \theta_1) + \xi_{jt} \\ f(c_{jt}) &= f(p_{jt} - \eta_{jt}(p_t, \hat{s}_t, \theta_2, \kappa)) \\ &= h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + w_{jt} \end{aligned}$$

The big issue in estimating this model is that p_{jt} and η_{jt} are functions of (ξ_t, w).

13.2.1 Approach 1: Estimating Demand Side Alone

We can think to estimate θ_2 from demand alone by

$$\delta_{jt}(\hat{s}_t, \theta_2) = h_d(x_{jt}, \mathrm{v}_{jt}, \theta_1) + \xi_{jt},\ \mathbb{E}[\xi_{jt} \mid x_t, \mathrm{v}_t, \mathrm{w}_t] = 0$$

as in Algorithm 1. Once we recover θ_2, we can recover marginal costs as $c_{jt} = p_{jt} - \eta_{jt}(p_t, \hat{s}_t, \theta_2, \kappa)$. The challenge here is that we can always produce a vector of marginal costs that rationalize that which we observe in other words, we cannot non-parametrically identify κ without more restrictions. In practice, we can discount some models if we recover $c_{jt} \leq 0$ as that's not believable.

13.2.2 Approach 2: Simultaneous Supply and Demand

We can think to estimate θ_2 from both supply and demand:

$$\begin{aligned}\delta_{jt}(\hat{s}_t, \theta_2) &= h_d(x_{jt}, \mathrm{v}_{jt}, \theta_1) + \xi_{jt},\ \mathbb{E}[\xi_{jt} \mid x_t, \mathrm{v}_t, \mathrm{w}_t] = 0 \\ f(c_{jt}) &= h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + w_{jt},\ \mathbb{E}[w_{jt} \mid x_t, \mathrm{v}_t, \mathrm{w}_t] = 0\end{aligned}$$

using an approach like in Algorithm 2. We can try to add additional moment restrictions to tease out different ownership structures characterized by κ. The biggest issue here is that if we run some test for κ it is hard to distinguish whether we are testing the ownership structure or just the functional form of our supply model.

13.2.3 Approach 3: Testing a Single Model of Conduct

We can think to estimate θ_2 from demand:

$$\delta_{jt}(\hat{s}_t, \theta_2) = h_d(x_{jt}, \mathrm{v}_{jt}, \theta_1) + \xi_{jt},\ \mathbb{E}[\xi_{jt} \mid x_t, \mathrm{v}_t, \mathrm{w}_t] = 0$$

At that point, we run the regression:

$$p_{jt} = h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + \lambda\eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa) + w_{jt},\ \mathbb{E}[w_{jt} \mid x_t, \mathrm{w}_t, \mathrm{z}_t] = 0$$

If the model of conduct κ held, then we would expect $\lambda = 1$ so that we can run this test. Again, this approach suffers from the fact that it's hard to distinguish whether we are testing the ownership structure or just the functional form of our supply model.

13.2.4 Approach 4: Goodness of Fit Tests

We can think to estimate θ_2 from demand:

$$\delta_{jt}(\hat{s}_t, \theta_2) = h_d(x_{jt}, \mathrm{v}_{jt}, \theta_1) + \xi_{jt},\ \mathbb{E}[\xi_{jt} \mid x_t, \mathrm{v}_t, \mathrm{w}_t] = 0$$

At that point, we can non-linear least squares to estimate for two models of conduct κ_1, κ_2:

$$f(p_{jt} - \eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_1)) = h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + w_{jt}^{(1)}$$
$$f(p_{jt} - \eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_2)) = h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + w_{jt}^{(2)}$$

We can compute $Q(\kappa_i) = \sum_{(j,t)\in\mathcal{J}\times T} \hat{w}_{jt}^{(i)}$ and employ the non-nested test of Rivers and Vuong (2002) to test between the two hypotheses. Again, this approach suffers from the fact that it's hard to distinguish whether we are testing the ownership structure or just the functional form of our supply model.

13.2.5 Approach 5: Backus, Conlon, and Sinkinson (2022) Motivations

We can think to estimate θ_2 from demand:

$$\delta_{jt}(\hat{s}_t, \theta_2) = h_d(x_{jt}, \mathrm{v}_{jt}, \theta_1) + \xi_{jt},\ \mathbb{E}[\xi_{jt} \mid x_t, \mathrm{v}_t, \mathrm{w}_t] = 0$$

At that point, for two models of conduct κ_1, κ_2 we do GMM and compare the violations of the unconditional moments:

$$f(p_{jt} - \eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_1)) = h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + w_{jt}^{(1)},\ \mathbb{E}[w_{jt}^{(1)} \mid z_t] = 0$$
$$f(p_{jt} - \eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_2)) = h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + w_{jt}^{(2)},\ \mathbb{E}[w_{jt}^{(2)} \mid z_t] = 0$$

We can again employ the non-nested test of Rivers and Vuong (2002) to test between the two hypotheses of conduct.

The issue here again is that the test will depend on the choice of unconditional moment restrictions chosen from the conditional moment restriction. In addition, it will depend on the choice of $h_s(\cdot)$.

13.2.6 Approach 6: Backus, Conlon, and Sinkinson (2022) Solution

The authors again estimate θ_2 from demand. They then assume that $f(x) = x$ and that one model of conduct κ_1 is correctly specified and the other κ_2 is not.

$$\begin{aligned} & p_{jt} - \eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_1) = h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + w_{jt}^{(1)},\ \mathbb{E}[w_{jt}^{(1)} \mid z_t] = 0 \\ \implies & p_{jt} - \eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_2) = h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + \underbrace{\eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_1) - \eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_2) + w_{jt}^{(1)}}_{=: w_{jt}^{(2)}} \end{aligned}$$

If they estimate the second model under the assumption that $\hat{\mathbb{E}}[w_{jt}^{(2)} \mid z_t] = 0$, they will be estimating a mis-specified model that has the omitted variable $\eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_1) - \eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_2)$. The idea of the test is to detect the presence of the omitted variable.

The authors suggest using the unconditional moment $\mathbb{E}[w_{jt}^{(i)}\mathbb{E}[\eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_1) - \eta_{jt}(p, \hat{s}, \hat{\theta}_2, \kappa_2) \mid z_t]] = 0$ in the GMM problem of Section 13.2.4. That is one of the optimal instruments when estimating the model:

$$p_{jt} = h_s(x_{jt}, \mathrm{w}_{jt}, \theta_3) + \tau\eta_{jt}^A + (1-\tau)\eta_{jt}^B + w_{jt},\ \mathbb{E}[w_{jt} \mid z_t] = 0$$

per the analysis of Section 10.3. The problem now reduces to estimating $\hat{w}_{jt}^{(i)}$ for $i \in \{1,2\}$ and $\mathbb{E}[\eta_{jt}(p,\hat{s},\hat{\theta}_2,\kappa_1) - \eta_{jt}(p,\hat{s},\hat{\theta}_2,\kappa_2) \mid z_t]$.

Algorithm 3 Testing Conduct

Input: : $\hat{s}_{jt}, x_{jt}, z_{jt}, p_{jt}, \theta_2, \kappa_1, \kappa_2$
Output: : $\mathcal{T}$ # Rivers and Vuong (2002) test statistic
Start Algorithm:
Split the markets t into 70% train and 30% test
On the training sample, train neural network g to predict $\eta_{jt}(p,\hat{s},\theta_2,\kappa_1) - \eta_{jt}(p,\hat{s},\theta_2,\kappa_2)$ from z_t.
On the training sample, train neural networks $h_s^{(i)}$ to predict $p_{jt} - \eta_{jt}(p,\hat{s},\theta_2,\kappa_i)$ from x_{jt} and w_{jt}, for $i \in \{1,2\}$.
On the test sample compute $\hat{w}_{jt}^{(i)} = (p_{jt} - \eta_{jt}(p,\hat{s},\theta_2,\kappa_i)) - h_s^{(i)}(x_{jt},\mathrm{w}_{jt})$
On the test sample, compute $Q(\kappa_i) \leftarrow \left(\frac{1}{T}\sum_{j,t}\hat{w}_{jt}^{(i)} g(z_t)\right)^2$
Use Rivers and Vuong (2002) to compute $\mathcal{T} \leftarrow \frac{\sqrt{T}(Q(\kappa_1)-Q(\kappa_2))}{\hat{\sigma}_{Q(\kappa_1)-Q(\kappa_2)}}$ and we know $\mathcal{T} \sim \mathcal{N}(0,1)$ asymptotically.
return T

14 Willingness to Pay and Healthcare

Industrial Organization economists study health care to answer the following sorts of questions:

- How do hospital (systems) and insurers interact?
- What is the value of adding a hospital to an insurer's network?
- What determines the market power of insurers? hospitals?
- Can steering incentives (in network / out of network) be effective in reducing costs?

14.1 Ex-Post Willingness to Pay (WTP)

What is the value that individual i places on having a hospital j included in network G? That can be modeled as individual i receives the following utility from option j

$$\begin{aligned} U_{ij} &= \alpha R_j + H_j' \Gamma X_i + \tau_1 T_{ij} + \tau_2 T_{ij} \cdot X_i + \tau_3 T_{ij} \cdot R_j - \gamma(Y_i, Z_i) P_j(Z_i) + \epsilon_{ij} \\ &= U(H_j, X_i, \lambda_i) + \epsilon_{ij} \end{aligned}$$

where

- $H_j := [R_j, S_j]$ is partitioned into generic R_j and diagnosis specific S_j characteristics.
- $X_i := [Y_i, Z_i]$ is partitioned into demographic characteristics Y_i and diagnosis specific characteristics Z_i.
- $T_{ij} := T_j(\lambda_i)$ is the distance from i's home to hospital j.
- $P_j(Z_i)$ is the price to someone with diagnosis characteristics Z_i of going to hosptial j.
- The parameters are $[\alpha, \Gamma, \tau]'$ and $\gamma(X_i)$.

Under the assumption that $\epsilon_{ij} \overset{iid}{\sim}$ Gumbel$(\mu = 0, \beta = 1)$, we have that the choice probabilities for individual i of alternative j is given by

$$s_{ij}(G) = \frac{\exp[U(H_j, X_i, \lambda_i)]}{\sum_{g \in G} \exp[U(H_g, X_i, \lambda_i)]}$$

and by Section 5.3.3 we know that

$$\mathbb{E}_{\epsilon_i}\left[\max_{j \in G} U(H_j, X_i, \lambda_i) + \epsilon_{ij}\right] = \log\left(\sum_{j \in G} \exp[U(H_j, X_i, \lambda_i)]\right)$$

up to translations. What happens when remove hospital j from the choice set after we know the diagnosis Z_i but before we draw ϵ_i. One can compute that

$$\mathbb{E}_{\epsilon_i}\left[\max_{j \in G} U(H_j, X_i, \lambda_i) + \epsilon_{ij}\right] - \mathbb{E}_{\epsilon_i}\left[\max_{k \in G \setminus \{j\}} U(H_k, X_i, \lambda_i) + \epsilon_{ij}\right] = \log\left(\frac{1}{1 - s_{ij}(G)}\right)$$

To express this difference in dollars, we must divide it by $\gamma(X_i)$. We define

$$\begin{aligned}\Delta W_{ij}^{EP}(G, Y_i, \lambda_i) &= \frac{\mathbb{E}_{\epsilon_i}\left[\max_{j \in G} U(H_j, X_i, \lambda_i) + \epsilon_{ij}\right] - \mathbb{E}_{\epsilon_i}\left[\max_{k \in G \setminus \{j\}} U(H_k, X_i, \lambda_i) + \epsilon_{ij}\right]}{\gamma(X_i)} \\ &= \frac{\log\left(\frac{1}{1 - s_{ij}(G)}\right)}{\gamma(X_i)}\end{aligned}$$

where EP stands for ex-post.[14]

14.2 Ex-Ante Willingness to Pay (WTP)

At the beginning of the year, when one chooses insurance, we don't observe Z_i. Thus, we must integrate out over Z_i:

$$\Delta W_{ij}^{EA}(G) = \int_{\mathcal{Z}} \Delta W_{ij}^{EP}(G, Y_i, \lambda_i) f(Z_i \mid Y_i, \lambda_i) dZ_i$$

where EA stands for ex-ante. As an insurer, you also wish to integrate over demographics and locations:

$$\Delta W_{j}^{EA}(G) = \int_{\mathcal{Y} \times \mathcal{Z} \times *} \Delta W_{ij}^{EP}(G, Y_i, \lambda_i) f(Z_i, Y_i, \lambda_i) dY_i dZ_i d\lambda_i$$

Next, we focus on the problem of the insurer, if they choose to include j in their network, they gain

[14] As a remark, one notes that in the logit model, the WTP for an alternative by an individual is *only* a function of the market share of that alternative. It's not dependent on whether or not there are close substitutes, which might be a bad property.

$$\pi_j = \alpha(N\Delta W_j^{EA}(G) - \Delta C_j(G)) + u_j$$

where $\Delta W_j^{EA}(G)$ comes from the extra willingness to pay of consumers, N is the population, $\Delta C_j(G)$ comes from extra costs of including j in the network, $\alpha \in [0,1]$ is the share of the surplus they get (ie., their bargaining weight), and u_j is an error term without specification at this point.

In estimating this model and running this regression to find α, one generally assumes that $\gamma(X_i) = \gamma_p$ and that $\Delta C_j(G) = 0$.

14.3 Hospital Merger

If hospitals j and k merge, we note that

$$\Delta W_{i(j+k)}^{EP}(G, Y_i, \lambda_i) = \frac{1}{(1 - s_{ij}(G) - s_{ij}(G))\,\gamma(X_i)}$$

so that the WTP for the combo is larger. The merger will result in a large WTP increase when j and k are close substitutes or affect the same sets of customers.

14.4 Bargaining between Insurers and Hospitals

In Ho (2009), she aimed to understand the following questions:

- How do insurers and healthcare providers (hospitals) divide profits?
- How is the division of profits related to the networks of hospitals offered by insurers?
- Given we know about WTP and demand, how can we treat supply decisions of endogenous networks and negotiations between firms seriously?

Ho (2009) assumed the following order to a game:

1. Hospitals make price offers to plans.
2. Plans choose their hospital networks.
3. Plans set premiums.
4. Consumers and employers jointly choose plans.
5. Sick consumers visit hospitals; plans pay hospitals for services provided.

Plan choice of quality and products, together with teh hospital's choices of capacity, location, services, and quality are outside the model (exogenous). The author specifies the following model utility model. Individual i receives utility u_{ijlm} for choosing hospital h in market m with diagnosis l

$$u_{ihlm} = \eta_h + x_h'\alpha + x_h' V_{il}\beta + \epsilon_{ihlm},\ \epsilon_{ihlm} \overset{iid}{\sim} \text{Gumbel}(\mu = 0, \beta = 1)$$

where V_{il} are consumer characteristics (eg., diagnosis, location) and x_h are observed hospital characteristics and η_h is unobserved hospital quality. The ex-ante expected utility of individual i in market m for plan j is

$$EU_{ijm} = \sum_{l \in \mathcal{L}} p_{il} \log \left(\sum_{h \in H_j} \exp\left(\eta_h + x_h' \alpha + x_h' V_{il} \beta\right) \right)$$

where p_{il} is the ex-ante probability that consumer i receives diagnosis l. The author then specifies the following utility individual i receives when choosing plan j in market m

$$\tilde{u}_{ijm} = \xi_{jm} + z_{jm}' \lambda + \gamma_1 EU_{ijm} + \gamma_2 \frac{prem_{jm}}{y_i} + \omega_{ijm},\ w_{ijm} \overset{iid}{\sim} \text{Gumbel}(\mu = 0, \beta = 1)$$

where ξ_{jm} is unobserved plan quality, z_{jm} includes size of the network, plan age, consumer reported availability and speed scores, etc. The variable $prem_{jm}$ is the premium of plan j in market m and y_i is the income of individual i. The outside good is uninsured care. Variables used to instrument for $prem_{jm}$ include average hospital wage, average weekly nurse wage across markets, etc. which are cost shifters such that ξ_{jm} is assumed to be mean-independent of them.

The surplus generated by plan j in market m with network H_j (against rivals with networks H_{-j}) is

$$S_{jm}(H_j, H_{-j}) = \sum_i \left(n_i s_{ijm}(H_j, H_{-j}) \left[prem_{jm} - p_i \sum_{h \in H_j} s_{ihm}(H_j) cost_h \right] \right)$$

where

- n_i is the population per demographic group
- p_i is the admission probability of an individual of type i

14.4.1 Supply Intuition

An interesting question to think about is why don't all hosptials reach agreement with all plans?

- A hospital choosing not to enter into an insurance plan changes both hospital demand and plan demand.
- Star hospitals benefit from selecting consumers that have high WTP.
- Hosptials may already be at capacity without contracting with all insurance plans.

The profit for an insurer is the surplus generated less costs paid to hospitals + some unspecified error term

$$\pi_{jm}^P(H_j, H_{-j}, \mathcal{X}, \theta) = S_{jm}(H_j, H_{-j}) - c_{jm}^{hosp}(H_j, H_{-j}, \mathcal{X}, \theta) + \phi_{jm}$$

where H_{-j} is unobserved when an insurer picks which hospitals to contract with. To be at an equilibrium, plan j's expected profits from network H_j must exceed its expected profits from an alternative network where it drops or adds a hospital h, holding all else fixed. Ie.,

$$\mathbb{E}\left[\pi_{jm}^P(H_j, H_{-j}, X, \theta) \mid I_{jm}\right] \geq \mathbb{E}\left[\pi_{jm}^P(H_j^h, H_{-j}, X, \theta) \mid I_{jm}\right]$$

where I_{jm} is plan j's information set in market m. At this point, we can consider combining moment conditions $\mathbb{E}[z_{jt}\xi_{jt}] = 0$ with these moment inequalities with the assumption that we penalize violations of moment equalities in both directions and moment inequalities only in one direction. Parameters will then be often set identified rather than point identified. Estimation of these models is out of scope of these notes.

15 DYNAMIC ESTIMATION

15.1 Markov Decision Process

A Markov decision process is characterized by a time index $t \in \{0, 1, ..., T\}$ with $T \leq \infty$, a state space S, a decision space A, a family of constraint sets $A_t(s_t) \subseteq A$, a family of transition probabilities $p_{t+1}(\cdot \mid s_t, a_t)$, a family of discount factors $\beta_t(s_t, a_t) \geq 0$ and a single period utility function $u_t(s_t, a_t)$ such the the utility functional U has an additive separability decomposition:

$$U(s, a) = \sum_{t=0}^{T} \left[\Pi_{j=0}^{t-1} \beta_j(s_j, a_j)\right] u_t(s_t, a_t)$$

Markov decision processes vary in that they can have a finite horizon (ie., $T < \infty$), infinite horizon (ie., $T = \infty$), a continuous decision time or discrete decision time, a discrete state space or infinite state space, and a discrete action space or infinite action space.

If utilities, transition probabilities, and discount factors are such that they're time invariant (ie., $u_t(s_t, a_t) = u(s, a)$, $p_{t+1}(\cdot \mid s_t, a_t) = p(\cdot \mid s, a)$, $\beta_t(s_t, a_t) = \beta(s, a)$),[15] we call the Markov decision problem stationary. In this case, it's often useful to consider the Bellman equation

$$V(s) = \max_{a \in \Phi(s)} \left[u(s, a) + \beta \int_S V(s') p(ds' \mid s, a) \right] \tag{14}$$

This is a functional equation and V represents a fixed point to that equation. If we define the Bellman operator Γ as the operation on V on the RHS, we search for a fixed point $V = \Gamma(V)$.

15.2 Formal Existence and Uniqueness Arguments for Value Function Deferred

Conditions under which a solution to functional equation exists, is unique, etc. is deferred to Jarda's notes in Chapter 3.3.3 (Characterization of the value function under bounded returns). Generally, these conditions include well-behaved policy correspondences $\Phi : S \to 2^A$, $\beta \in (0, 1)$, a continuous and bounded $u(\cdot, \cdot)$. These conditions help show that Γ is a strict contraction mapping under the infinity norm with contraction factor β (ie., $||\Gamma V - \Gamma W||_\infty \leq \beta ||V - W||_\infty$ for any V, W), which in turn guarantees a unique function solution to the functional equation since the set of functions on which Γ is a continuous self-map (ie., the set of continuous bounded (in the infinity norm) functions) is complete.

15.3 Solution Approaches

15.3.1 Value Function Iteration

Suppose that all the necessary conditions hold for the unique solution to the Bellman equation. We can consider iterating on the value function until convergence.

[15] We make an additional simplification that the discount factor is also state and action invariant.

As a simplification to solve the problem numerically, suppose that $|S| < \infty$ and $|A| < \infty$.[16] We can solve for the value function V by value function iteration as described in Algorithm 4.

Algorithm 4 Value Function Iteration

Input: : $u(\cdot,\cdot)$, β, S, A, $\Phi : S \to 2^A$, tol, $p(\cdot \mid \cdot,\cdot)$
Output: : V # value function for each state
Start Algorithm:
$V^{(-1)}(s) \leftarrow \infty \; \forall s \in S$
$V^{(0)}(s) \leftarrow 0 \; \forall s \in S$
$i \leftarrow 0$
while $||V^{(i)} - V^{(i-1)}|| \geq tol$ **do**
  $V^{(i+1)}(s) \leftarrow \max_{a \in \Phi(s)} u(s,a) + \beta \sum_{s' \in S} p(s' \mid s,a) V^{(i)}(s') \; \forall s \in S$
  $i \leftarrow i + 1$
end while
return $V^{(i)}$

Suppose that V^* is the unique solution to the functional equation. Also suppose that payoffs in each period are bounded above by 1. Then we can bound the number of steps n needed to have an approximation of the optimal value function to any level of precision. To see that, we can write that $||V^* - V^{(0)}|| = ||V^*||_\infty \leq \frac{1}{1-\beta}$. In turn, by the contraction mapping result, we have that

$$\begin{aligned}
\frac{\beta^n}{1-\beta} &\geq \beta^n ||V^*||_\infty \\
&= \beta^n ||V^* - V^{(0)}|| \\
&\geq ||\Gamma^n V^{(0)} - \Gamma^n V^*||_\infty \\
&= ||V^n - V^*||_\infty
\end{aligned}$$

If we wish to bound the error by ϵ, by transitivity we can set

$$\begin{aligned}
\epsilon &\geq \frac{\beta^n}{1-\beta} \\
\implies n &\geq \frac{1}{|\log(\beta)|} \log\left(\frac{1}{(1-\beta)\epsilon}\right)
\end{aligned}$$

We see that the number of steps is increasing in $\beta \in (0,1)$.

15.3.2 Policy Iteration

Again, suppose that all the necessary conditions hold for the unique solution to the Bellman equation. Again, to solve the problem numerically, we assume that $|S| < \infty$ and $|A| < \infty$.[17] We solve the problem by policy function iteration as in Algorithm 5.

[16] We can also assume that these sets are continuous and we construct a discrete approximation of them.
[17] We can also assume that these sets are continuous and we construct a discrete approximation of them.

Algorithm 5 Policy Function Iteration

Input: : $u(\cdot,\cdot)$, β, S, A, $\Phi : S \to 2^A$, tol, $p(\cdot,\cdot)$
Output: : h # policy function for each state
Start Algorithm:
$h^{(-1)}(s) \leftarrow a_{\Phi(s)[-1]}\ \forall s \in S$
$h^{(0)}(s) \leftarrow a_{\Phi(s)[0]}\ \forall s \in S$
$i \leftarrow 0$
while $||a^{(i)} - a^{(i-1)}|| \geq tol$ **do**
  $\tilde{P} \leftarrow$ matrix formed as $p(s' \mid a^{(i)}(s), s)$ for $s, s' \in S$
  $V^{(i)} \leftarrow [I_S - \beta\tilde{P}]^{-1}(u(a^{(i)}(s), s))_{s\in S}$
  $a^{(i+1)}(s) \leftarrow \arg\max_{a\in\Phi(s)} u(s,a) + \beta\sum_{s'\in S} p(s' \mid s, a)V^{(i)}(s')\ \forall s \in S$
  $i \leftarrow i + 1$
end while
return $V^{(i)}$

This algorithm speed is independent of β and very fast when S is relatively small. The only challenge is inverting the matrix for which special care needs to be taken for top speed.

15.3.3 Linear Programming

If the set of states and the set of actions are finite, we can write the Bellman equation problem as a linear program with dual: **NEED TO UNDERSTAND THE PRIMAL PROBLEM HERE BETTER**

$$\begin{aligned}&\min_V \sum_{s\in S} V(s)\\ &s.t.\ V(s) \geq u(s,a) + \beta\sum_{s'\in S} p(s' \mid s, a)\ \forall s \in S, a \in \Phi(s)\end{aligned}$$

This problem is now linear in V and can be solved using linear programming.

15.3.4 Collocation Method

For problems with a continuous state space and a discrete or continuous action space, we can consider using a basis function approximation to the value function. The idea is that we write an approximate value function $\tilde{V}(s;c) := \sum_{i=1}^K c_i\phi(x_i)$ for some basis functions $\{\phi_1, ..., \phi_K\}$. We can consider solving this problem by gradient descent on the following problem for some chosen points $\tilde{S}$

$$\min_c \sum_{s\in\tilde{S}} w(s)\left|\left|\sum_{i=1}^K c_i\phi_i(s) - \max_{a\in\Phi(s)}\left[u(s,a) - \int_S\left(\sum_{i=1}^K c_i\phi_i(s')\right)p(ds' \mid s, a)\right]\right|\right|^2$$

There are some complications here, namely, how to differentiate $\max_{a\in\Phi(s)}\left[u(s,a) - \int_S\left(\sum_{i=1}^K c_i\phi_i(s')\right)p(ds' \mid s, a)\right]$ with respect to c. If A is discrete (and finite), we can compute a sub-gradient that corresponds to the maximal quantity of a. If A is a continuous set, then we need to think about how to differentiate the measure with respect to a to find the argmax to then differentiate the resulting maximal quantity with respect to c. This problem is not in general convex so solving it is not easy.

15.4 Approximation and Discretization

Many problems we're interested in have continuous states and continuous decisions. We can consider an approximation to the problem in Equation (14):

$$\hat{V}(s) = \max_{a \in \hat{\Phi}(s)} u(s,a) + \beta \sum_{k=1}^{N} \hat{V}(s_k) \hat{p}(s_k \mid s, a)$$

There's a huge literature on how to efficiently generate grid points s_k and the transition matrix $\hat{p}$. Generally, one wants more grid points where one expects more curvature in the Value function as else where one will likely be interpolating.

15.5 Finite Horizon Problem

We can also consider a finite horizon problem where $T < \infty$ so that we do not expect a stationary solution to the Bellman equation. In that case we can write the recursive equation for $t \in \{0, ..., T\}$ where we usually use the convention that $V_{T+1}(s) = 0 \; \forall s \in S$.

$$V_t(s) = \max_{a \in \Phi_t(s)} u_t(s,a) + \beta \int_S V_{t+1}(s') p_{t+1}(ds' \mid s, a)$$

Solving for the optimal value and policy function here is very easy to do efficiently when we assume a finite state space and action space. We describe an approach Algorithm 6.

Algorithm 6 Backward Induction Algorithm

Input: : S, $p_t(\cdot \mid s, a)$, $u_t(\cdot,)$, $\Phi_t(\cdot)$
Output: : Value functions V and policy function h
Start Algorithm:
$V_{T+1}(s) \leftarrow 0 \; \forall s \in S$
for $t \leftarrow T$ **to** 1 **by** -1 **do**
  $h_t(s) \leftarrow \arg\max_{a \in \Phi_t(s)} \{u(s,a) + \sum_{s' \in S} p_{t+1}(s' \mid s, a) V_{t+1}(s')\} \; \forall s \in S$
  $V_t(s) \leftarrow \max_{a \in \Phi_t(s)} \{u(s,a) + \sum_{s' \in S} p_{t+1}(s' \mid s, a) V_{t+1}(s')\} \; \forall s \in S$
end for
return h, V

15.6 Why Dynamic Estimation?

Suppose that we're considering how to estimate the demand elasticity for laundry detergent as in Hendel and Nevo (2006). The long-run elasticity for laundry detergent might be super inelastic since we all need some detergent. That said, if detergent goes on sale periodically, we might see a spike in sales in infer that demand is very elastic. The issue here is that decision markers have a dynamic decision problem where they might buy a bunch of laundry detergent when it's on sale so that they don't need to buy it when it's at its normal price.

Solving dynamic problems is hard since there's a technological burden as the state space gets large. It's also tricky to disentangle serially correlated unobservables and unobserved heterogeneity from state dependent actions. A story here is that when modeling unemployment over time, it's hard to distinguish between the causes of someone being unemployed two periods in a row: (a) you could be unemployed two periods in a row because when you're unemployed one period it's hard to get a job the next (b) you have a bad unobserved characteristic that makes you unemployed in both periods. It's also hard to think about how to model expectations of agents over the state transitions given their actions.

15.7 Adding Heterogeneity with an Example [Rust]

We assume that an individual solves

$$\max_{\{a_t:t\geq 1\}} \mathbb{E}\left[\sum_{t=1}^{\infty} \beta^{t-1}\pi_\theta(x_t, a_t)\right]$$

where $a_t \in \{0, 1\}$ and

$$\pi_\theta(x_t, a_t) = \begin{cases} -c_\theta(x_t) & \text{if } a_t = 0 \\ -c_\theta(0) - RC_\theta & \text{if } a_t = 1 \end{cases}$$

and $c_\theta(x_t)$ is the maintenance cost for a bus engine that has x_t miles and the function is parametrized by θ, RC_θ is the cost of replacing the bus engine. The variable $\beta \in (0, 1)$ is a discount factor.

It's useful to write the choice-specific value function

$$\tilde{V}_\theta^t(x_t, a_t) = \begin{cases} \pi_\theta(x_t, 0) + \beta\mathbb{E}_t[V_\theta^{t+1}(x_{t+1}, \epsilon_{t+1})] & \text{if } a_t = 0 \\ \pi_\theta(x_t, 1) + \beta\mathbb{E}_t[V_\theta^{t+1}(0, \epsilon_{t+1})] & \text{if } a_t = 1 \end{cases} \tag{15}$$

We then define the value function as a function of the choice-specific value functions and unspecified errors ϵ_{it}

$$V_\theta^t(x_t, \epsilon_t) = \max\{\tilde{V}_\theta^t(x_t, 0) + \epsilon_{0t}, \tilde{V}_\theta^t(x_t, 1) + \epsilon_{1t}\}$$

We next make some assumptions to move forward. The first assumption is a Markovian assumption. That is

$$p(x_{t+1}, \epsilon_{t+1} \mid x_t, \epsilon_t, a_t, ..., x_0, a_0, \epsilon_0) = p(x_{t+1}, \epsilon_{t+1} \mid x_t, \epsilon_t, a_t)$$

The second assumption is a conditional independence assumption: the transition density of the controlled processes $\{x_t, \epsilon_t\}$ factor as

$$p(x_{t+1}, \epsilon_{t+1} \mid x_t, \epsilon_t, a_t) = q(\epsilon_{t+1} \mid x_{t+1})p(x_{t+1} \mid x_t, a_t)$$

These assumptions are powerful since they mean that we don't need to treat ϵ_t as a state variable, which is nice since it's unobserved. It's controversial because we could expect there to be persistent unobserved heterogeneity as in the unemployment story of Section 15.6. We next also make an iid Gumbel assumption:

$$\epsilon_{t+1} \mid x_{t+1} =_d \epsilon_{t+1} \overset{iid}{\sim} \text{Gumbel}(\mu = 0, \beta = 1)$$

Finally, we also assume that the problem is stationary so that we can drop all remaining subscripts from the variables. This stationarity assumption is key for the data-generating process when doing asymptotics. Given these assumptions, we can use the result of Section 5.3.1 to write that

$$\begin{aligned}\Pr(a=1 \mid x,\epsilon,\theta) &= \frac{\exp\left(\tilde{V}_\theta(x,1)\right)}{\exp\left(\tilde{V}_\theta(x,0)\right)+\exp\left(\tilde{V}_\theta(x,1)\right)}, \text{ using the assumptions} \\ &= \frac{\exp\left(\pi_\theta(x,1)+\beta V_\theta(0)\right)}{\exp\left(\pi_\theta(x,1)+\beta V_\theta(0)\right)+\exp\left(\pi_\theta(x,0)+\beta \mathbb{E}_x\left[V_\theta(x')\right]\right)}\end{aligned} \tag{16}$$

We note that the conditional value function here plays the same rule as the systematic utility of a choice in the traditional logit setting.

15.8 Estimating θ in the Example [Rust]

The likelihood function for a single bus can be written as

$$\begin{aligned}\mathcal{L}(x_1,...,x_T;a_1,...,a_T \mid x_1;\theta) &= \Pr(a_1 \mid x_1;\theta)\Pi_{t=2}^T \Pr(a_t,x_t \mid x_1,a_1,...,x_{t-1},a_{t-1};\theta) \\ &= \Pr(a_1 \mid x_1;\theta)\Pi_{t=2}^T \Pr(a_t,x_t \mid x_{t-1},a_{t-1};\theta), \text{ by stationary Markov assumption} \\ &= \Pi_{t=1}^T \Pr(a_t \mid x_t;\theta)\Pi_{t=2}^T \Pr(x_t \mid x_{t-1},a_{t-1}), \text{ by conditional independence}\end{aligned}$$

As a result, the log-likelihood is additively separable into two components

$$l(\theta) = \sum_{t=1}^T \log(\Pr(a_t \mid x_t;\theta)) + \sum_{t=2}^T \log\left(\Pr(x_t \mid x_{t-1},a_{t-1})\right)$$

We can discretize mileage into intervals and compute empirical transition probabilities between intervals to construct the transition probabilities between intervals. At this point, we seek to estimate the parameters θ that characterize the flow payoff. We can select β arbitrarily and then do a non-nested hypothesis test to decide between multiple choices. We seek to find θ to solve the following problem:

$$\begin{aligned}\hat{\theta}_T^{MLE} &= \arg\max_\theta l(\theta) \\ &= \arg\max_\theta \sum_{i=1}^N \sum_{t=1}^T \log(\Pr(a_{it} \mid x_{it};\theta))\end{aligned} \tag{17}$$

where we've introduced other busses ($i \in \{1,...,N\}$) that we assume that have transition and choice probabilities that are identically and independently distributed.

15.8.1 Nested Fixed Point Approach

In Algorithm 7, we present a nested fixed point approach to find $\hat{\theta}_{NT}^{MLE,NFP}$ that solves the problem in Equation (17). For this algorithm, we find that it's useful to iterate on the *value function*. To show the algorithm, we first generalize a little

bit and consider a general case where we have a finite state of actions S and a finite set of actions A. For any $x \in S$, we can define the value function

$$\begin{aligned} V(x) &:= \max_{a \in A}\{\pi(x,a) + \beta \mathbb{E}_{x'|x,a}[V(x')] + \epsilon_a\}, \epsilon_a \overset{iid}{\sim} \text{Gumbel}(\mu = 0, \beta = 1) \quad (18) \\ &= \max_{a \in A}\{\pi(x,a) + \beta \sum_{x' \in S} P_{x',xa} V(x') + \epsilon_a\}, \text{ assuming transition probabilities } P_{x',xa} \\ &= \log\left(\sum_{a \in A} \exp\left(\pi(x,a) + \beta \sum_{x' \in S} P_{x',xa} V(x')\right)\right), \text{ up to a scalar constant (eg., } \gamma) \quad (19) \end{aligned}$$

We now present the algorithm. The Bellman operator in the inner loop of the algorithm is in fact a contraction mapping. For suggestions on how one might prove that, see Chapter 3.3.3 of Jarda's notes and Section 6.4.1 of my Applied Microeconometrics Notes.

Algorithm 7 Rust – Nested Fixed Point

Input: : $\pi, \beta, S, A, tol_l, tol_V, P, \{x_{it}, a_{it}\}_{i=1,t=1}^{N,T}, \eta$
Output: : $\hat{\theta}_T^{MLE,NFP}$
Start Algorithm:
$\theta^{(0)} \leftarrow [0, ..., 0]$
$\theta^{(-1)} \leftarrow [-1, ..., -1]$
$D_\theta l(\theta^{(-1)}) \leftarrow [\infty, ..., \infty]$
$n \leftarrow 0$
while $||D_\theta l(\theta^{(n-1)})|| \geq tol_l$ **do**
  $V_{\theta^{(n)}}^{(0)}(x) = 0 \; \forall x \in S$
  $V_{\theta^{(n)}}^{(-1)}(x) = -\infty \; \forall x \in S$
  $j \leftarrow 0$
  while $||V_{\theta^{(n)}}^{(j)} - V_{\theta^{(n)}}^{(j-1)}|| \geq tol_V$ **do**
    $V_{\theta^{(n)}}^{(j+1)}(x) \leftarrow \log\left(\sum_{a \in A} \exp\left(\pi_{\theta^{(n)}}(x,a) + \beta \sum_{x' \in S} P_{x',xa} V_{\theta^{(n)}}^{(j)}(x')\right)\right)$
    $j \leftarrow j + 1$
  end while
  $l(\theta^{(n)}) \leftarrow \sum_{i=1}^{N} \sum_{t=1}^{T} \log\left(\Pr(y_{it} \mid x_{it}, \theta^{(n)})\right)$ where $\Pr(a_{it} \mid x_{it}, \theta^{(n)}) = \frac{\exp\left[\pi_{\theta^{(n)}}(x_{it}, a_{it}) + \beta V_{\theta^{(n)}}^{(j)}(x_{it})\right]}{\sum_{a' \in A} \exp\left[\pi_{\theta^{(n)}}(x_{it}, a') + \beta V_{\theta^{(n)}}^{(j)}(x_{it})\right]}$
  Compute $D_\theta l(\theta^{(i)})$ using the analytically using the IFT or numerically using FD
  $\theta^{(n+1)} \leftarrow \theta^{(n)} + \eta D_\theta l(\theta^{(n)})$
  $n \leftarrow n + 1$
end while
return $\theta^{(n-1)}$

15.8.1.1 Comments This nested fixed point approach has the issue that we need to very precisely solve the nested fixed point problem for any iteration of the outer loop. If we don't then the errors in the solution to the inner loop will propagate to the derivatives in the solution to the outer loop and the outer loop solver may not converge. As explained in Section 15.3.1, the speed of the inner loop is determined by β and this can be rather slow if β is near 1. As for deriving $l(\theta)$ with respect to θ, see Section 6.3.2 of my Applied Microeconometrics Notes.

15.8.2 Hotz and Miller

We define the *choice specific value function*[18]

[18] This is like the quantity in Equation (15) except with new more condensed notation for ease of writing.

$$v_a(x) := \pi(x,a) + \beta \mathbb{E}_{x'|x,a}[V(x')]$$

As explained in Section 5.1, we only estimate differences in utility so that the choice probabilities are invariant to shifting the systematic utility by a constant amount. That is, taking the problem in Equation (18), we can write the action choice probabilities from state $x \in S$ as

$$\begin{aligned} p_a(x) &:= \frac{\exp(v_a(x))}{\sum_{a \in A} \exp(v_a(x))} \\ &= \frac{\exp(v_a(x) - v_0(x))}{\sum_{a \in A} \exp(v_a(x) - v_0(x))} \end{aligned} \tag{20}$$

where $v_0(x)$ is the utility associate with some reference action in state $x \in S$. We next notice that this implies:

$$\log(p_j(x)) - \log(p_0(x)) = v_j(x) - v_0(x) \tag{21}$$

Lemma 1 (Arcidiacano-Miller)**.** For any state-action pair (x,a), there exists a function ψ such that

$$\bar{V}(x) = v_a(x) + \psi_a(\mathrm{p}(x))$$

where the expected state value function, $\bar{V}$ is defined in Section 15.8.1.

Proof.

We have that

$$\begin{aligned} \bar{V}(x) &= \mathbb{E}_\epsilon \left[\max_{\epsilon_a} \{v_j(x) + \epsilon_j\} \right] \\ &= \mathbb{E}_\epsilon \left[\max_{\epsilon_a} \{ \underbrace{v_j(x) - v_a(x)}_{=:\phi_{ja}(\mathrm{p}(x))} + \epsilon_j \} \right] + v_a(x) \\ &= \underbrace{\mathbb{E}_\epsilon \left[\max_{\epsilon_a} \{\phi_{ja}(\mathrm{p}(x)) + \epsilon_j\} \right]}_{=:\psi_a(\mathrm{p}(x))} + v_a(x) \\ &= \psi_a(\mathrm{p}(x)) + v_a(x) \end{aligned}$$

The fact that $\mathrm{p}(x)$ is a sufficient statistic for x in $v_j(x) - v_a(x)$ comes from the observation in Equation (21). One can then use Section 5.3.2 to show that $\psi_a(x) = -\log(\mathrm{p}_a(x)) + \gamma$. □

Again, assuming that $|S| < \infty$, let's "normalize" $\pi(x,0) = 0 \; \forall x \in S$.[19] Let's also define F_0 to be the transition matrix post choosing action 0. We then have that, stacking equations for each state and defining $\pi_a = (\pi(x,a))_{x \in S}$ for any $a \in A$:

$$\begin{aligned} v_0 &= \cancelto{0}{\pi_0} + \beta F_0 \bar{V} \\ \implies \bar{V} - \psi_0(\mathrm{p}) &= \beta F_0 \bar{V} \\ \implies \bar{V} &= (I_{|S|} - \beta F_0)^{-1} \psi_0(\mathrm{p}) \end{aligned}$$

[19] This is not a real normalization. If there are $|S|$ states and $|A|$ actions, then there are $|S| * |A|$ state specific flow utilities and in the logit case we generally innocuously set one of those to be 0. In this case, we're setting $|S|$ of them to be 0, which isn't a real normalization since ideally we would be able to have a different utility associated with the reference action in different states. The issue is that there are only $|S| * (|A| - 1)$ linearly independent choice probabilities in the data so that we need to restrict our utility function for non-parametric identification.

To estimate the utility functions of other actions:

$$\begin{aligned} & v_a = \pi_a + \beta F_a \bar{V} \\ \implies & \bar{V} - \psi_a(\mathrm{p}) = \pi_a F_a \bar{V} \\ \implies & \pi_a = -\psi_j(\mathrm{p}) + (I_{|S|} - \beta F_j)^{-1}(I_{|S|} - \beta F_0)^{-1}\psi_0(\mathrm{p}) \end{aligned}$$

Since we can identify $\psi_{a'}(\mathrm{p})$ from the data $\forall a' \in A$ as seen in Lemma 1, we can estimate the flow utilities associated with each action non-parametrically.

In the continuous state case, we cannot apply the Arcidiacano-Miller trick since we must interpolate the value function at states between approximate nodes.

15.8.3 Forward Simulation

We note that the choice-specific value function $v_a(x;\theta)$ can be expressed as follows, approximately:

$$\begin{aligned} v_a(x;\theta) := \pi(x,a;\theta) + \beta \mathbb{E}_{x^{(1)}|x,a}[\mathbb{E}_{a^{(1)}|x}[\mathbb{E}_{\epsilon^{(1)}|a^{(1)},x^{(1)}}[\pi(x^{(1)},a^{(1)};\theta) + \epsilon^{(1)} \\ + \beta(... + \beta \mathbb{E}_{X^{(T)}|X^{(T-1)},a^{(T-1)}}[\mathbb{E}_{d^{(T)}|x^{(T)}}[\mathbb{E}_{\epsilon^{(T)}|a^{(T)},x^{(T)}}[\pi(x^{(T)},a^{(T)};\theta) + \epsilon^{(T)}]]])]]] \end{aligned}$$

for some large T. One sees that if $\epsilon_i \overset{iid}{\sim}$ Gumbel$(\mu = 0, \beta = 1)$, then we have that

$$\begin{aligned} \mathbb{E}[\epsilon_i \mid i = \arg\max_{j \in A}\{U_j + \epsilon_j\}] &= \mathbb{E}[\max_{j \in A}\{U_j + \epsilon_j\} - U_i \mid i = \arg\max_{j \in A}\{U_j + \epsilon_j\}] \\ &= \mathbb{E}[\max_{j \in A}\{U_j + \epsilon_j\} \mid i = \arg\max_{j \in A}\{U_j + \epsilon_j\}] - U_i \\ &= \log\left(\sum_{j \in A} \exp(U_j)\right) + \gamma - U_i, \text{ using Appendix A} \\ &= \gamma - \Pr\left(i = \arg\max_j\{U_j + \epsilon_j\}\right) \end{aligned}$$

These two results suggest the following algorithm to estimate θ given data $\{\hat{a}_t, x_t\}_{s=1}^N$. We simulate paths $s \in \{1, ..., S\}$ from each (x, a) so as to estimate the choice specific value functions:

$$\begin{aligned} v_a^{(s)}(x;\theta) &\approx \pi(x,a;\theta) + \beta[\pi(x_s^{(1)}, a_s^{(1)};\theta) + \gamma - \hat{p}(a_s^{(1)} \mid x_s^{(1)}) + \beta(... + \beta(\pi(x_s^{(T)}, a_s^{(T)};\theta) + \gamma - \hat{p}(a_s^{(T)} \mid x_s^{(T)})))] \\ v_a(x;\theta) &\approx \frac{1}{S}\sum_{s=1}^{S} v_a^{(s)}(x;\theta) \end{aligned}$$

The choices $\hat{a}^{(t)} \mid \hat{x}^{(t)}$ are sampled iid from the empirical distribution $\hat{p}(a_t \mid x_t)$ and $\hat{x}^{(t+1)} \mid \hat{x}^{(t)}, \hat{a}^{(t)}$ are sampled iid from the empirical distribution $\hat{G}(x_{t+1} \mid x_t, a_t)$. From here, we can compute the model choice probabilities $p(a_t \mid x_t;\theta)$ using Equation (20). We solve the following optimization problem:

$$\min_\theta ||\hat{p} - p(\theta)||^2$$

I would recommend solving this problem using an auto-differentiation package like PyTorch.

15.8.4 Mathematical Programming with Equilibrium Constraints (MPEC) Approach

We defer the MPEC estimation of θ to section 6.4.2 in my Applied Microeconometrics Notes as there's a very clever way to vectorize everything and then use the augmented lagrangian method to solve the problem.

15.9 Can We Distinguish Between a Static Model and a Dynamic Model?

Generally we cannot. The issue is that we can write an observationally equivalent static model and the same set of parameters θ can be estimated for the static and dynamic model. To test between the two models, we can consider doing a non-nested hypothesis test. If we can find exclusion restrictions that move the flow utility without moving the future utility, we can think about testing the dynamic nature of the model as in Magnac and Thesmar (2002).

As one final remark, suppose that we parametrize $c_\theta(x) = \theta_1 + \theta_2 x$. If we estimate this as a dynamic model and separately as a static model, do we expect to see θ_2 to be bigger or smaller in the dynamic case? Similarly, do we expect to see RC_θ bigger or smaller in the dynamic case? The answer is that we expect to see θ_2 bigger in the static case and RC_θ smaller in the static case. The intuition is that to rationalize the engine replacements in the static case, it must generally be that the cost of maintenance exceeds the cost of replacement, which pushes the cost of replacement to be smaller and the cost of maintenance to be larger.

16 SWITCHING COSTS

Suppose we have a static random coefficients logit utility specification:

$$u_{ijt} = x'_{jt}\beta_i - \alpha_i p_{jt} + \xi_{jt} + \epsilon_{ijt}$$

and have panel-data on each consumer i's purchase choices among alternatives and we note that a particular individual often repeats their consumption of the same choice. It could be that they have a persistent taste captured β_i for some characteristic of the choice. It could also be that there's an unmodeled switching cost that keeps the individual picking the same alternative as last period. Distinguishing between these two possibilities is hard as state dependence to $p-1$ periods back in time and unobserved persistent taste is observationally equivalent to state dependence to p periods back in time.

16.1 Why do we care about switching costs?

On the consumer and policy side, switching costs are a real friction in the economy. Consumers are often highly persistent in their product choices because

- they really like the product.
- they are unaware of alternatives.
- they are lazy.
- they have a psychological cost associated with change.
- there's cost associated with learning about other products.

These considerations are extremely important in the market for health insurance. For instance, individuals are often concerned about if their doctor is still covered by another insurance plan. Individuals also can have a hard time figuring out how much they expect to spend on medical expenses and deciding which plan is best accordingly.

On the firm side, if firms are competing in an undifferentiated products setting, switching costs are ways that they can escape the Bertrand competition trap where they all sell prices at marginal cost.

Separately, discerning between switching costs and persistent heterogeneity is important because the existence of the former encourages the desire to harvest profits in a dynamic way as in the Cabral model of Section 16.1.1. The latter encourages market segmentation and the desire to statically extract surplus from consumers by price discrimination that way.

16.1.1 Dynamic Firm Model – Cabral (2008)

Consider a dynamic optimization problem faced by a price-setting single-product firm i with a vector of current market prices p, previous market shares q', exogenous switching costs s that just consumers face with respect to their product, and exogenous marginal costs c. Suppose that future profits are discounted by β. Firm i's value is given by

$$V_i(\mathrm{q}', s) = \max_{p_i} (\mathrm{p}_i - c_i) \cdot \mathrm{q}_i(\mathrm{q}', \mathrm{p}, s) + \beta V_i(\mathrm{q}, s)$$

The first order condition of the firm is

$$\begin{aligned} 0 &= \mathrm{q}_i(\mathrm{q}', \mathrm{p}, s) + (\mathrm{p}_i - c_i) \cdot \partial_{\mathrm{p}_i} \mathrm{q}_i(\mathrm{q}', \mathrm{p}, s) + \beta \partial_{\mathrm{p}_i} V_i(\mathrm{q}, s) \\ &= \mathrm{q}_i(\mathrm{q}', \mathrm{p}, s) + (\mathrm{p}_i - c_i) \cdot \partial_{\mathrm{p}_i} \mathrm{q}_i(\mathrm{q}', \mathrm{p}, s) + \beta \left(\partial_{\mathrm{q}_i} V_i(\mathrm{q}, s)\right) \left(\partial_{\mathrm{p}_i} \mathrm{q}_i(\mathrm{q}', \mathrm{p}, s)\right) \\ \implies \mathrm{p}_i - \mathrm{q}_i(\mathrm{q}', \mathrm{p}, s) &= \underbrace{\frac{\mathrm{q}_i(\mathrm{q}', \mathrm{p}, s)}{-\partial_{\mathrm{p}_i} \mathrm{q}_i(\mathrm{q}', \mathrm{p}, s)}}_{\text{harvesting}} - \underbrace{\beta \left(\partial_{\mathrm{q}_i} V_i(\mathrm{q}, s)\right)}_{\text{investment}} \end{aligned}$$

The first term on the RHS reflects pressure on firm i to harvest profits and raise prices. If quantity does not drop much when it raises prices, then it will want to increase its price. The second term on the RHS reflects pressure on firm i to invest for tomorrow's profits. If we assume there are switching costs, then firm i can set lower prices today, attract more consumers to the product, and then have more sticky consumers next period.

We note that $V(\mathrm{q}, s)$ is increasing in s and $\partial_{\mathrm{q}_i} V_i(\mathrm{q}, s = 0) = 0$ since then firms face a static problem and don't need to invest for the future. We also note that the higher switching costs are, the lower $|-\partial_{\mathrm{p}_i} \mathrm{q}_i(\mathrm{q}', \mathrm{p}, s)|$ so that firms will want to raise prices more now. It's unclear whether $\mathrm{q}_i(\mathrm{q}', \mathrm{p}, s)$ is increasing or decreasing in s: it could be that it induces the firm to set higher prices and then sell less or sell more.

16.2 Modeling State Dependence

To move towards a utility specification that captures switching costs and $p = 1$ state dependence,[20] we often write the following modified utility specification:

$$u_{ijt} = x'_{jt}\beta_i - \alpha_i p_{jt} + \xi_{jt} + \gamma_i \mathbb{1}_{\{y_{i(t-1)}=j\}} + \epsilon_{ijt} \tag{22}$$

where $y_{i(t-1)}$ is the choice individual i made in period $t-1$. We can think of the switching costs as either (a) increasing utility for j if that was previously purchased or (b) providing additional cost if $y_{it} \neq y_{i(t-1)}$. These models are often *time-inconsistent* since at time $t-1$ individuals don't internalize the impact their decision now will have on their utility at time t.

[20] We could also include two lagged choices if we so desired.

16.2.1 Identification – Dube, Histch, and Rossi (2009)

These authors look at panel demand data for orange juice using the utility model in Equation (22). They define the distribution of heterogeneous parameters $\theta := [\alpha_i, \beta_i', \gamma_i]'$. For each individual i, they assume that $\theta_i \overset{iid}{\sim} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$ so that it's drawn from a mixture of K normals.

They choose to estimate the demand parameters using Markov chain Monte Carlo (MCMC)[21] so that they need to specify priors on π, μ_k, Σ_k for $k \in \{1, ..., K\}$.

Intuitively, to understand where identification between heterogeneity and state-dependence enters, we can consider the following simple example. Suppose that we observe the sequence of purchases in Table 1 between P, C where H denotes the product is being sold at a high price and L denotes the product is being sold at a low price.

Time	Price(C)	Price(P)	Choice
$t = 1$	H	H	C
$t = 2$	H	L	C
$t = 3$	H	H	C
$t = 4$	L	H	P
$t = 5$	H	H or L	P

Table 1. State-Dependence Identification

In period $t = 5$ we would declare evidence of state-dependence since the individual had the same decision earlier (no matter if Price(C) $= H$ or L) and picked C but now is picking P.

They do find evidence of state-dependence in their estimation and also persistent heterogeneity in taste. They proceed to do a clever robustness check for if they mis-specified the agent's indirect utility. They consider estimating the same model but when they've permuted the choices over time of individuals. If the model was mis-specified, they would still expect to find persistent heterogeneity but also spuriously large γ_i. They in fact continue to find persistent heterogeneity but they don't find state-dependence, supporting their original finding.

16.3 Switching Costs and Adverse Selection in Health Care Insurance

Handel (2013) looks at the health care insurance market and asks the question if the reduction of switching costs can worsen adverse selection faced by insurers and potentially unravel the market. He then looks at how consumer welfare changes holding prices constant and allowing for equilibrium price changes.

The idea behind the paper is that when one lowers switching costs between insurance plans by providing more information in the market, the enrollees with the lowest costs and lowest risk aversion will reallocate to less comprehensive plans. That will mean that comprehensive plans will contain a riskier pool so that insurance companies will raise those premia due to adverse selection. In very bad cases, the "Market for Lemons" death spiral can occur. As an implication, it's possible that a reduction in switching costs ends up harming social welfare.

16.3.1 Data

The data used in the paper is from one large, self-insured employer. They have employee health plan choices, claim-level employee utilization and expenditure data, employee demographics (eg., job characteristics, age, gender, job tenure, ...). They assume that at time t_0, all enrollees actively select a new plan (with no default option) so that there's no switching cost. In time t_1 and onwards, employee picks plan with default as the choice from the previous year. Plan prices adjust in time t_1 and onwards and they use claims data (diagnoses and spending) to construct an ex-ante out-of-pocket expense measure.

[21] See page 66 of these notes for a nice explanation of MCMC.

16.3.2 Descriptive Statistics

They look to compare employees who make active choices in a year to those who made an active choice in prior cohorts to confirm that they're similar in observed demographics. They note that at time t_1 a PPO_{250} plan[22] becomes strictly dominated at some income levels and family sizes meaning that at any level of medical expense it would be preferable to choose the dominating plan. They look across all years and find that higher health risk individuals tend to choose more comprehensive plans (so that there's adverse selection in some form).

16.3.3 Cost Model

To assess what's the expected amount of out-of-pocket expenses for a family unit k under plan j in year t, Handel (2013) enters past diagnoses and payments into a Johns Hopkins model to predict future medical and pharmacy expenses and divides the sample into groups based on predicted expenses.

Handel (2013) assumes that (1) consumers' beliefs match the cost model's estimates (no private information) and (2) that there's no moral hazard meaning that consumers don't go to the hospital more because they have a more comprehensive plan.

16.3.4 Demand Model

They define the following demand model. Family unit k under plan j at time t has the following utility:

$$
\begin{aligned}
u_{kjt} &= \int_0^\infty u_k\left(W_k, OOP, p_{kjt}, \mathbb{1}_{\{j=y_{k(t-1)}\}}, X_k^A, X_{kt}^B\right) f_{kjt}(OOP) dOOP, \\
u_k\left(W_k, OOP, p_{kjt}, \mathbb{1}_{\{j=y_{k(t-1)}\}}, X_k^A, X_{kt}^B\right) &= -\frac{1}{\gamma_k(X_k^A)} \exp\left(-x\gamma_k(X_k^A)\right), \text{ where} \\
x &= W_k - p_{kjt} - OOP + \eta(X_{kt}^B, Y_k)\mathbb{1}_{\{j=y_{k(t-1)}\}} + \delta_k(Y_k)\mathbb{1}_{\{j=PPO_{1200}\}} + \epsilon_{kjt}
\end{aligned}
$$

where $y_{k(t-1)}$ is the plan chosen by family unit k in period $t-1$, p_{kjt} is the premium associated with plan j for family k in period t, X_{kt}^B are time-varying family unit k demographics, X_k^A are time-invariant demographics, W_k is family unit k wealth, and Y_k is an indicator for if family unit k is single or a family.[23]

For their main results they assume that $\gamma_k(X_k^A) \overset{iid}{\sim} \mathcal{N}(\mu + (X_k^A)'\beta, \sigma_\gamma^2)$ and $\eta(X_{kt}^B, T_k) = \eta_0 + (X_{kt}^B)'\eta_1 + \eta_2 Y_k$. Finally, they assume that $\epsilon_{kjt} \overset{iid}{\sim} \mathcal{N}(0, \sigma_j^2(Y_k))$.

16.3.5 Supply Model

The total premium is set as the average plan cost for the plan's enrollee's in the prior year plus an administrative markup L. That is, the total premium (TP_{jt}^y) conditional on income / family level y is set as:

$$
TP_{jt}^y = \left(\frac{1}{|K_{j(t-1)}^y|} \sum_{k \in K_{j(t-1)}^y} PP_{kj(t-1)}\right) + L
$$

where $PP_{kj(t-1)}$ is the plan cost for family unit k in period $t-1$ and $K_{j(t-1)}^y$ is the set of family units with income y that chose plan j in period $t-1$.

[22] The subscript of 250 is the deductible. Plans with lower deductibles generally have higher premia and are generally more comprehensive.

[23] There is one term missing in the wealth specification that's dropped here for clarity of exposition but is not meaningful to the analysis of the paper.

16.3.6 Results

They find the following things:

- Switching costs are higher for families than for singles. This might make sense since family's have more at stake so they're more careful about switching.

- There's little heterogeneity in the risk-aversion parameter.

- In a counter-factual analysis that reduces switching costs, the healthier switch first, consumer choices conditional on price improve, but there's worse adverse selection and the endogenous premium for the comprehensive PPO_{250} rises since the pool of individuals selecting it is less healthy. They find that total welfare falls, where welfare is measured as the certainty equivalent wealth to give you the same utility as that which you receive from the plan. This counterfactual analysis assumes that switching costs are totally psychological and shouldn't be included in welfare calculations at baseline. Naturally, it is the healthier switchers who benefit and the less healthy non-switchers who are worse off.

16.3.7 Comments

To opine on the results, it's important to think about what is a switching cost. If switching costs are purely psychological, then maybe it doesn't make sense to include them in welfare calculations. If they're actual learning costs or transaction costs, then maybe they should partially be included in welfare calculations. I find it hard to believe that one can totally distinguish between these two sorts of costs: it could be that people experience a psychological cost associated with switching that requires a large financial gain to compensate it. In other words, it could be a psychological cost that's compensated by a large financial gain as opposed to a transaction cost. That said, if one does notice that medical costs of switchers are similar to those of the population, then one might be less inclined to believe that there are large transaction costs.

A potential policy conclusion from this paper is that we should keep consumers uninformed about the medical plans to prevent worsening of adverse selection by healthy people switching to less comprehensive plans. This takeaway is obviously controversial since healthier people are worse off financially when they're pooled with the less-healthy people since they face higher premia. The state of the world where consumers are uninformed about medical plans is not Pareto superior to a state where they are informed.

17 DYNAMIC DEMAND

In Sections 7 and 9, we looked at static demand problems where consumers make a one-time decision to purchase a product and that's the end of their participating in the model. In reality, demand for goods takes a dynamic form:

- Individuals often care if they can sell goods on a secondary market when making a purchase and it's interesting to understand if this option is good or bad for initial sellers? Does it improve overall consumer welfare?

- Consumers purchase a good often understanding that it will have a certain lifetime. Firms must manage a tradeoff between building durable goods that last longer and thus merit a higher price but face less frequent purchases or building less durable goods and selling them at a lower price more frequently.

- Firms offer temporary sales for some goods. Are these sales forms of pure-price discrimination where firms will sell to more attentive and budget conscious consumers at a lower price? Are they for attracting product switchers as discussed in Section 16.

17.1 Lifetime Utility for Durable Good

We now say that the utility for an individual i of purchasing a good j at time t (omitting any disturbance term) is given by

$$u_{ijt} = \sum_{\tau=t}^{\infty} \beta^{\tau-t} f_{ij\tau}$$

where $\beta \in (0,1)$ is a discount factor and $f_{ij\tau}$ is a flow payment from the good at time τ. We can even consider wrapping this utility in an expectation if there's a probability ρ_τ that it breaks in any period, say. If the secondary market is frictionless, and we wish to parametrize $f_{ij\tau}$ it's often sensible to label $\Delta p_{jt} := p_{jt} - p_{j(t-1)}$ as the per-period price of owning the good for $t-1$ to t.

17.2 Durable Goods

Adoption of high-tech goods is often intuitively pondered using a curve like in Figure 1. The idea is that there are some initial adopters of the goods, those who adopt it in the modal period of adoption, and then some laggards who adopt it late. If one were to estimate demand as a function of price, even using an appropriate IV for price, one might find that demand falls with price in the early period, rises with price in the mid period, and again falls with price in the late period. In these settings, it helps to think about the large picture when thinking about demand.

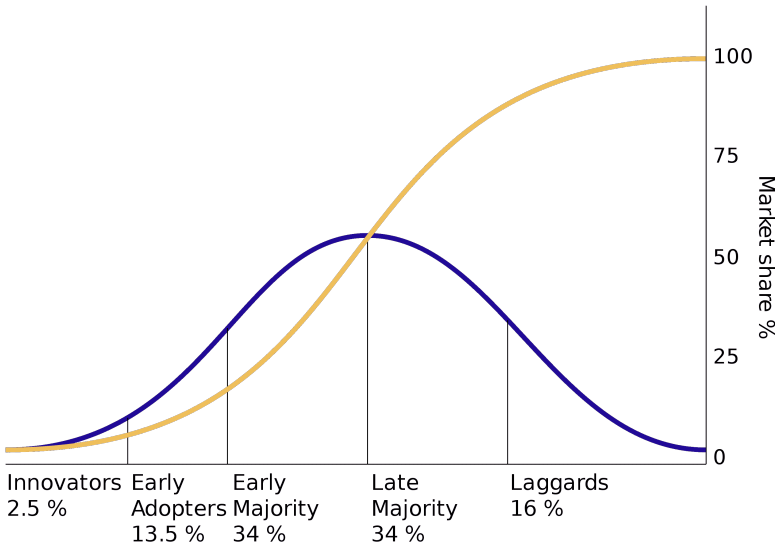

Figure 1. Adoption of Durable Good

17.3 Dynamic Demand for Durables: An Infeasible Static Approach

Consider the following utility specification for individual i purchasing goods $j \in \{0, ..., J\}$ at time t:

$$u_{ijt} = \alpha_i x_{jt} + \xi_{jt} + \epsilon_{ijt}, \text{ for } j \in \{1, ..., J\}$$
$$u_{i0t} = \bar{u}_{i0t} + \epsilon_{i0t}$$

The problem is that $\bar{u}_{i0t}$ is unobserved and $\bar{u}_{i0t} = 0 \; \forall i, t$ is a *bad* assumption. The reason is that (a) some individuals may have a substitute product in good condition at home and (b) may anticipate a good sale tomorrow. If we knew $\bar{u}_{i0t}$, then we could in good conscious estimate this model and expect meaningful results.

These models are challenging to reasonably estimate without this knowledge and even if we try some ad-hoc approach to control for the deterministic utility of the outside option (eg., $\bar{u}_{i0t} = \eta_{it} + \gamma_{0i} + \gamma_{1i} t + \gamma_{1i} t^2 ...$). We still probably

would not be confident about any counterfactual analysis. The reason is that baked into $\bar{u}_{i0t}$ is equilibrium beliefs about the future, which we'd expect to change in any counterfactual analysis like a price increase today. In contrast, note that we don't expect for demand primitives of individuals, such as preferences towards a particular characteristic, to change in any counterfactual analysis so that our previous discussions of counterfactual analyses are sound under this assumption.

17.4 Dynamic Demand for Durables: A Simple 2 Period Model

Consider a simple two period market where consumers purchase at most one unit of an (infinitely) durable good. After purchasing the good, they leave the market forever and have no interest in purchasing a second unit of the good. Utility of individual i in period $t \in \{1, 2\}$ for purchasing good $j \in \{0, ..., J\}$ is given by

$$\begin{aligned} u_{ijt} &= \alpha_i x_{jt} + \xi_{jt} + \epsilon_{ijt} \\ u_{i0t} &= \epsilon_{i0t} \end{aligned}$$

where $\epsilon_{ijt} \overset{iid}{\sim} \text{Gumbel}(\mu = 0, \beta = 1)$. We present now two feasible approaches to estimating the demand model.

17.4.1 A Naive Static Approach

Suppose that we treat each period $t \in \{1, 2\}$ as a separate market and close our eyes and do BLP. We will get wrong that there's no substitution across markets even though buying a good in period $t = 2$ is a substitute to buying it in period $t = 1$. For people who don't purchase the good in period $t = 2$, we will incorrectly assign 0 deterministic utility to the outside good to those individuals even though they have infinite utility for the outside good. The distribution of α_i is likely different in $t \in \{1, 2\}$ since people who purchase the good in the first period leave the market but we'll be throwing everyone into the estimation of that parameter in period 2.

17.4.2 Multi-period Static Demand with Complete Information

We consider another estimation approach where we suppose that consumers have full information about all product characteristics and shocks $(x_{j1}, x_{j2}, \xi_{j1}, \xi_{j2}, \epsilon_{ij1}, \epsilon_{ij2})$ ex-ante for both periods. We could think about estimating a static demand model where all goods across both periods are treated as choices for an individual. That is, purchasing good j in $t = 1$ is treated as a separate good from purchasing good j in $t = 2$.

One issue with the estimation approach is that in reality individuals don't have perfect foresight of the second period characteristics and shocks in the first period. A second issue is that we may wish to consider the fact that if one purchases in the first period one has two periods of consumption and any consumption in the second period should probably be discounted.

17.4.2.1 Some Incomplete Information Suppose initially that only $\Omega_{i,t=1} := \{x_{j1}, x_{j2}, \xi_{j1}, \xi_{j2}, \epsilon_{ij1}\}$ is known at period $t = 1$. We say that $\Omega_{i,t=1}$ is i's *information set* at period $t = 1$. Then, we can write that the utility of the outside option in period $t = 1$ is given by

$$\begin{aligned} u_{i01} &= \mathbb{E}_{\epsilon_{i2}} \left[u_{ij2} \mid \Omega_{i,t=1} \right] \\ &= \log \left(\sum_j \exp \left(\alpha_i x_{j2} + \xi_{j2} \right) \right) \end{aligned}$$

If we wish for x_{j2} or ξ_{j2} to be not-known at period one, we can write that

$$\begin{aligned} u_{i01} &= \mathbb{E}_{\epsilon_{i2},\xi_{jt},x_{j2}}\left[u_{ij2} \mid \Omega_{i,t=1}\right] \\ &= \mathbb{E}_{\xi_{j2},x_{j2}}\left[\log\left(\sum_j \exp\left(\alpha_i x_{j2} + \xi_{j2}\right)\right)\right] \end{aligned}$$

and integrate out over ξ_{j2}, x_{j2} given some distribution assumption in $\Omega_{i,t=1}$. Lastly, note that if we had the possibility of purchase over more periods than two periods, we could form nested log-sum-exps into the utility of the outside option in period $t = 1$ to account for the ability to wait for more periods.

17.4.2.2 Rational Expectations When estimating these models, we generally assume rational expectations. That means that we write a regression model, where we assume ξ_{j2}, x_{j2} are known, for simplicity:

$$\mathbb{E}_t\left[\log\left(\sum_j \exp\left(\alpha_i x_{j2} + \xi_{j2}\right)\right) \mid \Omega_{it}\right] = \log\left(\sum_j \exp\left(\alpha_i x_{j2} + \xi_{j2}\right)\right) + \nu_{it},\ \mathbb{E}_t\left[\nu_{it} \mid \Omega_{it}\right] = 0$$

so that we can use any moment of the form $\mathbb{E}_t\left[\nu_{it} f(\Omega_{it})\right] = 0$ for any f to consistently and unbiasedly estimate model parameters.

17.4.2.3 Rewriting the Demand Model We can now think of rewriting the demand model of Section 17.4 as follows:

$$\begin{aligned} u_{ij1} &= \alpha_i x_{j1} + \xi_{j1} + \epsilon_{ij1}, \text{for } j \in \{1, ..., J\} \\ u_{i01} &= \beta \log\left(\sum_j \exp\left(\alpha_i x_{j2} + \xi_{j2}\right)\right) + \nu_{i1} + \epsilon_{i01} \\ u_{ij2} &= \alpha_i x_{j2} + \xi_{j2} + \epsilon_{ij2}, \text{for } j \in \{1, ..., J\} \\ u_{i02} &= \epsilon_{i02} \end{aligned}$$

where $\beta \in (0, 1)$ is a discount factor. To model the changes in the distribution of α_i in the population from period $t = 1$ to period $t = 2$, we might think that consumers belong to a discrete set of types so that an individual i in type G_i has mass $w_{G_i,t=1}$ in the population at time $t = 1$. In period $t = 2$, the unnormalized mass in the population moves according to the law of motion $w_{G_i,t=2} = w_{G_i,t=1} s_{G_i 0,t=1}$ where $s_{G_i 0,t=1}$ is the mass of individuals of type G_i that go for the outside option in period $t = 1$. Lastly, we could think about multiplying the deterministic component of utility u_{ij1} by $(1 + \beta)$ to reflect the fact the consumers in the first period get the utility over two periods.

17.5 Dynamic Demand for Storables: Same Idea

We write a similar demand model for storable goods. We say that the utility of individual i of purchasing goods $j \in \{0, ..., J\}$ at time t is given by

$$\begin{aligned} u_{ijt} &= \alpha_i x_{jt} + \xi_{jt} + \epsilon_{ijt} \\ u_{i0t} &= \bar{u}_{i0t} + \beta \mathbb{E}_t\left[\max_{j \in \{0,...,J\}} u_{ij(t+1)}\right] + \nu_{it} + \epsilon_{ijt} \end{aligned}$$

where we interpret $\bar{u}_{i0t}$ as the amount of the storable good individual i has remaining. When estimating the model, it's often reasonable to again partition the market into types that have a certain amount of laundry left.

17.6 Durables versus Storables Comments

We generally expect to see $\text{Cov}_t(u_{i0t}, p_{jt}) > 0$ for storable goods and $\text{Cov}_t(u_{i0t}, p_{jt}) < 0$ for durable goods. The reason is that for storable goods, when the price is high, we expect less individuals to consume the good as they likely bulk-purchased it in a previous period during a sale. Meanwhile, for durable goods, we expect to see this negative covariance during the period of majority adoption as shown in Figure 1 as individuals are adopting the good despite the increase in price.

17.7 Hendel and Nevo (2006)

When a supermarket cuts the cost of laundry detergent for a week, there's a huge increase in sales, which would leads an economist to conclude that consumers are extremely elastic with respect to price. Meanwhile, when a supermarket makes a permanent price cut to laundry, there's little sales impact in the long run, which leads economists to think that consumers look inelastic with respect to price.

Hendel and Nevo thought about this situation and concluded that consumers respond to temporary price reductions by stockpiling inventory and spend down their inventories in high price periods. They also found that consumers have variables storage costs and price sensitivities. This result would have implications for high-low inter-termporal pricing strategies to get the low-storage cost high price sensitivity individuals during periods of low prices and the high-storage cost low price sensitivity individuals during periods of high prices.

17.7.1 Data

Hendel and Nevo looked at 9 supermarkets in Chicago. At the store-level for each of $j \in \{1, ..., 13\}$ ($J := 13$) brands in week t, they looked at purchases of $x \in \{32, ..., 256\}$ ounces of laundry detergent. For each brand, size, and week, they had access to price p_{jxt}, quantity q_{jxt}, and promotions a_{jxt}. They had access to household level data for $h \in \{1, ..., H\}$ where household h has utility $u(c_{ht}, \nu_{ht}; \theta_h)$ where current consumption $c_{ht} = \sum_{j=1}^{J} c_{hjt}$ is *not* brand specific. They assumed a shock ν_{ht} that affects marginal utility of consumption and they endowed purchase decisions $d_{hjxt} = \mathbb{1}_{\{h \text{ purchases brand } j \text{ at size } x \text{ at } t\}}$.

17.7.2 Dynamic Utility Model

They assumed that household h facing state $s_{ht} := (i_{ht}, a_t, p_t, \nu_{ht}, \epsilon_{ht})$ solves the sequence problem

$$
\begin{aligned}
V(s_{ht}) = \max_{\{c_{h\tau}(s_h^\tau), d_{hjx\tau}(s_h^\tau)\}_\tau, s_h^\tau} & \sum_{\tau=t}^{\infty} \beta^{\tau-t} \mathbb{E}_t [u(c_{h\tau}(s^\tau), \nu_{h\tau}(s_h^\tau), \theta_h) - C_h(i_{h,\tau+1}(s_h^\tau); \theta_h) \\
& + \sum_{j,x} d_{hjx\tau}(\alpha_h^p p_{jx\tau} + \xi_{hjx} + \alpha_h^a a_{jx\tau} + \epsilon_{hjx\tau}) \mid s_{ht}] \\
\text{s.t. } & i_{h(\tau+1)}(s_h^\tau) = i_{h\tau}(s_h^\tau) + \left(\sum_{j,x} d_{hjx\tau}(s_h^\tau) \cdot x_{h\tau} \right) - c_{h\tau}(s^\tau), \ \sum_{j,x} d_{hjx\tau}(s_h^\tau) = 1, \ i_{h(\tau+1)}(s_h^\tau) \geq 0
\end{aligned}
$$

where $C(\cdot)$ is the cost of storage. Note that we assume that a household receives a brand-size specific shock to utility $\xi_{jx\tau}$ only at the time of purchase, after that they're indifferent in their brand choice when consuming the laundry detergent. We're fairly general here and define s_h^τ to be a sequence of states for t to τ and in the most general problem require household h to make a choice along all these contingent paths.

For estimating the demand model, Hendel and Nevo make the following three simplifying assumptions:

- ν_{ht} are iid over time and consumer types.

- p_{jxt} and a_{xjt} follow an exogenous first-order Markov process.

- $\epsilon_{hjxt} \overset{iid}{\sim}$ Gumbel$(\mu = 0, \beta = 1)$.

These assumptions allow them to drop $(\nu_{ht}, \epsilon_{ht})$ from the state so that $s_{ht} = (i_{ht}, a_t, p_t)$ and solve a recursive problem. They can write the probability of a choice of household h at time t as

$$\Pr(d_{hjxt} \mid i_{ht}, p_t, a_t, \nu_{ht}) = \frac{\exp\left(\alpha_h^p p_{jxt} + \alpha_h^a a_{jxt} + \xi_{hjx} + M(i_{ht}, p_t, a_t, j, x)\right)}{\sum_{k,y} \exp\left(\alpha_h^p p_{kyt} + \alpha_h^a a_{kyt} + \xi_{hky} + M(i_{ht}, p_t, a_t, k, y)\right)}$$
$$M(p_t, a_t, j, x) = \max_{c_{ht}} u(c_{ht}, \nu_{ht}) - C(i_{h(t+1)}) + \beta \mathbb{E}_t \left[V(s_{t+1} \mid d_{hjxt} = 1, c_{ht}, s_t)\right]$$

They can integrate over ν_{ht} to write the unconditional probability of the choice

$$\Pr(d_{hjxt} \mid i_{ht}, p_t, a_t) = \int_{\mathbb{R}} \Pr(d_{hjxt} \mid i_{ht}, p_t, a_t, \nu_{ht}) f_\nu(\nu_{ht}) d\nu_{ht}$$

Finally, they can write the likelihood of any sequence of purchases as

$$l(\theta) = \Pi_{h=1}^{H} \Pi_{t=1}^{T} \Pr(d_{hjxt} = 1 \mid i_{ht}, p_t, a_t)^{d_{hjxt}}$$

17.7.3 Estimation

They propose a 3-Step estimator to recover the demand parameters. In the first step, they look at brand choice *conditional on size* in order to recover $\{\alpha^a, \alpha^p, \xi\}$. They concede that estimating the parameters in parts leads to a reduction in efficiency of the estimator but they propose this strategy to make estimation more tractable and intuitive. In the first step, the key is that conditional on purchase size $M(i_{ht}, p_t, a_t, j, x) = M(i_{ht}, p_t, a_t, k, x)\ \forall j, k \in \{1, ..., J\}$ so that dynamics drop out of the brand-choice equation. The choice probabilities (conditional on ν_{ht}) reduce to

$$\Pr(d_{hjxt} \mid i_{ht}, p_t, a_t, \nu_{ht}, x) = \frac{\exp\left(\alpha_h^p p_{jxt} + \alpha_h^a a_{jxt} + \xi_{hjx}\right)}{\sum_k \exp\left(\alpha_h^p p_{kxt} + \alpha_h^a a_{kxt} + \xi_{hkx}\right)}$$

so that they can recover $\{\alpha^a, \alpha^p, \xi\}$ from maximizing a likelihood conditional on the consumption size across households and time. Note that each household's parameter estimation is separable in this setting though it's also reasonable to let $\xi_{hjx} = \xi_{jx}\ \forall h \in \{1, ..., H\}$ so that we don't estimate household specific brand-size shocks. Next, they define the size x specific inclusive values at t for household h by

$$\omega_{hxt} = \log\left(\sum_j \exp\left(\alpha_h^p p_{jxt} + \alpha_h^a a_{jxt} + \xi_{hjx}\right)\right)$$

They then rewrite the problem for household h as

$$V(i_{ht}, \omega_{ht}, \nu_{ht}) = \max_{c,x} \{u(c_{ht}, \nu_{ht}; \theta_h) - C(i_{h(t+1)}; \theta_h) + \omega_{hxt} + \epsilon_{xt} + \beta \mathbb{E}_t[V(i_{h(t+1))}, \omega_{h(t+1)}, \nu_{h(t+1)})]\}$$

using the dynamic of the original problem in Section 17.7.2. Notice that they've implicitly assumed that ω_t is a sufficient statistic for the dynamics from $\omega_t \to \omega_{t+1}$ when computing the transition probabilities.

Once the inclusive values and transition probabilities are known, they estimate the estimate the dynamic parameters, and the storage parameters using a nested fixed point approach as in Algorithm 7. Separately for each household, given θ_h, they solve for the value function by value function iteration, they then compute compute the likelihood of the choices and the gradient of the likelihood with respect to θ_h, and do gradient descent to find the optimal θ_h. Note that they do *not* observe the consumption c_{ht} in this estimation procedure, which is rather spectacular. Instead, they intuit consumption from dynamic purchase decisions (ie., if a household purchases two times in short order, they intuit they have high storage costs for instance).

For more details on estimation and identification, see Section 4.3.1 of their paper.

REFERENCES

Chris Conlon. Grad io — fall 2025 lecture notes and syllabus. Online, 2025. Accessed: 2026-01-15. URL: https://chrisconlon.github.io/gradio.html.

APPENDIX A

Here, we show the solution to Exercise 2.1 in Alfred Galichon's Mathematical Methods of Discrete Choice Models textbook that shows that the maximum quantity of the max of a bunch of Gumbel random variables is independent of the argmax of those Gumbel random variables.

Exercise 2.1

Let $(\epsilon_1, \epsilon_2) \overset{iid}{\sim}$ Gumbel$(\alpha = 0, \beta = 1)$.

(i)

Take any $U_1, U_2 \in \mathbb{R}$. Here, I wish to show that the distribution of $U_1 + \epsilon_1 | U_1 + \epsilon_1 \geq U_2 + \epsilon_2$ coincides with the distribution of $C := \log(\exp(U_1) + \exp(U_2)) + \epsilon$ where $\epsilon \sim$ Gumbel$(\alpha = 0, \beta = 1)$.

As a first step, for any $z \in \mathbb{R}$, let's compute $\Pr(U_1 + \epsilon_1 \geq U_2 + \epsilon_2 | U_1 + \epsilon_1 \leq z)$. Let f_{ϵ_i} and F_{ϵ_i} denote the pdf and cdf of ϵ_i, respectively.

$$
\begin{aligned}
\Pr(U_1 + \epsilon_1 \geq U_2 + \epsilon_2 | U_1 + \epsilon_1 \leq z) &= \int_{-\infty}^{z - U_1} \int_{-\infty}^{U_1 - U_2 - x_1} f_{\epsilon_2}(x_2) f_{\epsilon_1 | U_1 + \epsilon_1 \leq z}(x_1) \mathrm{d}x_2 \mathrm{d}x_1 \\
&= \frac{1}{F_{\epsilon_1}(z - U_1)} \int_{-\infty}^{z - U_1} F_{\epsilon_2}(U_1 - U_2 + x_1) f_{\epsilon_1}(x_1) \mathrm{d}x_1 \\
&= \frac{1}{F_{\epsilon_1}(z - U_1)} \int_{-\infty}^{z - U_1} \exp(-e^{-U_1 + U_2 - x_1}) \exp(-x_1 - e^{-x_1}) \mathrm{d}x_1 \\
&= \frac{1}{F_{\epsilon_1}(z - U_1)} \int_{-\infty}^{z - U_1} \exp(-x_1) \exp\left[-\exp(-x_1) e^{-U_1 + U_2} - \exp(-x_1) \right] \mathrm{d}x_1 \\
&\text{Define } t = \exp(-x_1) \implies \mathrm{d}t = -\exp(-x_1)\, \mathrm{d}x_1 \\
&= \frac{1}{F_{\epsilon_1}(z - U_1)} \int_{\exp(U_1 - z)}^{\infty} \exp(-t(1 + e^{-U_1 + U_2})) \mathrm{d}t \\
&= \frac{1}{F_{\epsilon_1}(z - U_1)} \frac{\exp(U_1)}{\exp(U_1) + \exp(U_2)} \left[\exp(-t(1 + e^{-U_1 + U_2})) \right]_{\exp(U_1 - z)}^{\infty} \\
&= \frac{1}{F_{\epsilon_1}(z - U_1)} \frac{\exp(U_1)}{\exp(U_1) + \exp(U_2)} \left[\exp(\exp(U_1 - z)(1 + e^{-U_1 + U_2})) \right] \\
&= \frac{1}{F_{\epsilon_1}(z - U_1)} \frac{\exp(U_1)}{\exp(U_1) + \exp(U_2)} \left[\exp(e^{U_1 - z} + e^{U_2 - z}) \right] \\
&= \frac{1}{F_{\epsilon_1}(z - U_1)} \frac{\exp(U_1)}{\exp(U_1) + \exp(U_2)} \left[\exp(e^{-z + \log(\exp(U_1) + \exp(U_2))}) \right]
\end{aligned}
$$

We note that $F_{\epsilon_1}(z - U_1) = \Pr(U_1 + \epsilon_1 \leq z)$ and by the DZW theorem that $\Pr(U_1 + \epsilon_1 \geq U_2 + \epsilon_2) = \frac{\exp(U_1)}{\exp(U_1) + \exp(U_2)}$. Thus, by Bayes' rule,

$$\begin{aligned}
\Pr(U_1+\epsilon_1 \leq z | U_1+\epsilon_1 \geq U_2+\epsilon_2) &= \frac{\Pr(U_1+\epsilon_1 \geq U_2+\epsilon_2 | U_1+\epsilon_1 \leq z)\Pr(U_1+\epsilon_1 \leq z)}{\Pr(U_1+\epsilon_1 \geq U_2+\epsilon_2)} \\
&= \frac{\cancel{\frac{1}{F_{\epsilon_1}(z-U_1)}}\cancel{\frac{\exp(U_1)}{\exp(U_1)+\exp(U_2)}}\left[\exp(e^{-z+\log(\exp(U_1)+\exp(U_2))})\right]\cancel{F_{\epsilon_1}(z-U_1)}}{\cancel{\frac{\exp(U_1)}{\exp(U_1)+\exp(U_2)}}} \\
&= \exp(e^{-z+\log(\exp(U_1)+\exp(U_2))})
\end{aligned}$$

This precisely coincides with the CDF of $C = \log(\exp(U_1)+\exp(U_2)) + \epsilon$ where $\epsilon \sim$ Gumbel$(\alpha = 0, \beta = 1)$.

(ii)

Let $\tilde{y} = 1$ iff $U_1 + \epsilon_1 > U_2 + \epsilon_2$ and $\tilde{y} = 2$, otherwise. I wish to show that $\tilde{y} \perp\!\!\!\perp U_{\tilde{y}} + \epsilon_{\tilde{y}}$. I have that for any $z \in \mathbb{R}$,

$$\begin{aligned}
\Pr(\tilde{y}=1, U_{\tilde{y}}+\epsilon_{\tilde{y}} \leq z) &= \Pr(U_1+\epsilon_1 \leq z, U_1+\epsilon_1 > U_2+\epsilon_2) \\
&= \Pr(U_1+\epsilon_1 \leq z | U_1+\epsilon_1 > U_2+\epsilon_2)\Pr(U_1+\epsilon_1 > U_2+\epsilon_2) \\
&= \exp(e^{-z+\log(\exp(U_1)+\exp(U_2))})\left(\frac{\exp(U_1)}{\exp(U_1)+\exp(U_2)}\right) \\
&= \Pr(U_{\tilde{y}}+\epsilon_{\tilde{y}} \leq z)\Pr(\tilde{y}=1)
\end{aligned}$$

$$\begin{aligned}
\Pr(\tilde{y}=2, U_{\tilde{y}}+\epsilon_{\tilde{y}} \leq z) &= \Pr(U_2+\epsilon_2 \leq z, U_2+\epsilon_2 > U_1+\epsilon_1) \\
&= \Pr(U_2+\epsilon_2 \leq z | U_2+\epsilon_2 > U_1+\epsilon_1)\Pr(U_2+\epsilon_2 > U_1+\epsilon_1) \\
&= \exp(e^{-z+\log(\exp(U_1)+\exp(U_2))})\left(\frac{\exp(U_2)}{\exp(U_1)+\exp(U_2)}\right) \\
&= \Pr(U_{\tilde{y}}+\epsilon_{\tilde{y}} \leq z)\Pr(\tilde{y}=2)
\end{aligned}$$

Thus, indeed by the definition of independence, I have that $\tilde{y} \perp\!\!\!\perp U_{\tilde{y}} + \epsilon_{\tilde{y}}$.

(iii)

Suppose now that $\epsilon_i \overset{iid}{\sim}$ Gumbel$(\alpha = 0, \beta = 1)$ for $i \in \{1, ..., Y\}$ where $[Y] := \{1, ..., Y\}$. For $U \in \mathbb{R}^Y$, define $\tilde{y} := \arg\max_{y \in [Y]} U_y + \epsilon_y$. I wish to show that $\tilde{y} \perp\!\!\!\perp U_{\tilde{y}} + \epsilon_{\tilde{y}}$. We have that

$$
\begin{aligned}
\Pr(U_{\tilde{y}} + \epsilon_{\tilde{y}} \leq z, \tilde{y} = y) &= \Pr(U_y + \epsilon_y \leq z, U_y + \epsilon_y \geq \max_{z \neq y} U_z + \epsilon_z) \\
&= \Pr(U_y + \epsilon_y \leq z | U_y + \epsilon_y \geq \max_{z \neq y} U_z + \epsilon_z) \Pr(U_y + \epsilon_y \geq \max_{z \neq y} U_z + \epsilon_z) \\
&= \Pr\left(U_y + \epsilon_y \leq z | U_y + \epsilon_y \geq \log\left(\sum_{z \neq y} \exp(U_z)\right) + \epsilon\right) \Pr(U_y + \epsilon_y \geq \max_{z \in [Y]} U_z + \epsilon_z) \\
&\text{since } \max_{z \neq y} U_z + \epsilon_z \overset{d}{=} \log\left(\sum_{z \neq y} \exp(U_z)\right) + \epsilon \text{ where } \epsilon \sim \text{Gumbel}(\alpha = 0, \beta = 1) \text{ and } \epsilon \perp\!\!\!\perp (\epsilon_1, ..., \epsilon_Y) \\
&= \exp(e^{-z + \log(\exp(U_y) + \exp(\log(\sum_{z \neq y} \exp(U_z))))}) \left(\frac{\exp(U_y)}{\sum_{z \in [Y]} U_z}\right), \text{ using part (i)} \\
&= \exp(e^{-z + \log(\sum_{z \in [Y]} \exp(U_z))}) \left(\frac{\exp(U_y)}{\sum_{z \in [Y]} U_z}\right) \\
&= \Pr(U_{\tilde{y}} + \epsilon_{\tilde{y}} \leq z) \Pr(\tilde{y} = y)
\end{aligned}
$$

Thus indeed, by the definition of independence, $\tilde{y} \perp\!\!\!\perp U_{\tilde{y}} + \epsilon_{\tilde{y}}$.