The Price is (not) Right: Incorporating Unobserved Price Heterogeneity in Demand Models

Laura Grigolon, Liana Jacobi and Michelle Sovinsky*

January 17, 2020

Preliminary and Incomplete

Abstract

This paper presents a method to incorporate price heterogeneity when researchers don't observe individual transaction prices. Prices may not be available in sufficient detail if, for example, the purchase is illegal, the sales price is private information, the individual did not make a purchase, or the prices are aggregated. Estimates based on inadequate pricing data can lead to biased elasticities resulting in misguided policy recommendations. We show how to overcome this problem by supplementing the pricing data with (commonly available) additional data on demographics to construct an empirical price distribution from which the researcher can obtain simulated draws of consumer-specific prices. Our approach is similar in spirit to the traditional approach employed to identify unobserved individual product heterogeneity. Monte-Carlo results show that our method is an improvement over standard techniques and yields substitution patterns that reflect heterogeneity in prices across individuals. We take our approach to data from illicit markets where we know very little about the price paid.

[0] *Grigolon is at the the University of Mannheim, Department of Economics and CEPR, lgrigolo@mail.uni-mannheim.de; Jacobi is at the University of Melbourne, Department of Economics, ljacobi@unimelb.edu.au; Michelle Sovinsky is at the the University of Mannheim, Department of Economics and CEPR, michelle.sovinsky@gmail.com. Sovinsky acknowledges support from the European Research Council Grant #725081 FORENSICS. Grigolon and Sovinsky acknowledge support from the CRC Transregio Grant 224.

1 Introduction

Often empirical researchers face data constraints when estimating models of demand. This paper examines one problem that arises frequently - incomplete, missing, or insufficient data on individual transaction price. Individual pricing data may be lacking for a multitude of reasons: prices are aggregated (hence the data are imprecise), the purchase is illegal or the sales price is private information (hence the data are incomplete), or the individual did not make a purchase (hence the data are missing). In standard workhorse models, a common assumption is that all consumers (those who purchase or choose not to purchase) face the same product characteristics when making their purchase decisions, where one of these characteristics is the price. It is well-known that this practice can lead to measurement error bias (see Berry (1994)) as not all consumers face the same price. This motivates the instrumental variables approach to correct for price endogeneity which can arise from correlation between the price and unobserved (product-specific) demand characteristics.

However, there may be individual-specific heterogeneity in the transaction price that arises from non-random differences across consumers. To the extent this is the case, the price will be correlated with individual-specific unobserved characteristics. The instrumental variables approach proposed in Berry et al. (2004) may not be sufficient to overcome this individual source of endogeneity leading to inconsistent parameter estimates and biased policy recommendations.

We propose a method to remedy this problem by incorporating unobserved individual prices. Our method shows how to use variation in (commonly available) micro-level data combined with pricing data in a novel way. The main insight is to combine data on demographics together with aggregate pricing data to construct an empirical price distribution from which the researcher can obtain simulated draws of consumer-specific prices. Our approach is similar in spirit to the traditional approach in the literature employed to identify unobserved individual product heterogeneity. Monte-Carlo results show that our method is an improvement over traditional approaches and yields substitution patterns that reflect (true) heterogeneity in prices across individuals.

Individuals may face different transaction prices for the same products in many markets - take for example automobiles and housing. In both markets the final price is the outcome of a negotiation process, and, to the extent that there is variation in the ability of consumers to bargain or obtain information, the transaction price will reflect this non-random variation. Indeed, bargaining over final goods prices can be seen in the markets for health insurance, medical devices, capital assets, financial products, as well as in business-to-business transactions.[1] The literature has found that differences in negotiated prices from list prices can be substantial. For example, Chandra et al. (2017) show that consumers characteristics can explain about 20% of the differences in final transaction prices for identical new cars.

Another common source of non-random unobserved price variation arises from privately-known price discounts. For example, pharmaceutical companies typically offer discounts to large buyers, such as governments, hospitals or insurance companies, and these discounts are not public information. Unobserved discounts occur in numerous situations such as between: content providers and cable companies (Crawford (2012)), computer manufacturers and microprocessor producers (Eizenberg et al. (2019)), book publishers and online retailers, car manufacturers and dealerships (Huang (2017)), and advertisers and content providers.[2]

Finally, adequate pricing data are especially difficult to obtain for illicit products due to the illegal nature of the market. In fact, in many empirical studies of illicit markets researchers often have to resort to using prices from police drug busts (Williams (2004)), which are not always complete, suffer from selection issues, and certainly not individual-specific.

The literature has provided some direction to address inadequate pricing data. One such study is Miller and Osborne (2014) who recover optimal transaction prices in the demand for cement when only average prices are observed by using equilibrium conditions.

[1] For example, see Goldberg (1996), Scott-Morton et al. (2001), and Busse et al. (2006), D'Haultfœuille et al. (2018) (automobiles); Allen et al. (2014), Allen et al. (2019), Robles-Garcia (2019) (housing); Dafny (2010) (health insurance); Grennan (2013) (medical devices), Gavazza (2016) (capital assets); Hastings et al. (2017) (financial products); and Town and Vistnes (2001) (2001) business-to-business transactions.

[2] For example, esimates of the demand for advertising content incorporate an ad price which may not be correct due to (unobserved) advertising terms of trade with content providers. These discounts on listed ad media prices are rarely observed but are likely to be non-random.

Similarly, D'Haultfœuille et al. (2018) develop a method to estimate price discrimination (based on transaction prices) in the market for new cars when only aggregate list prices are available. They develop a method to identify transaction prices by taking advantage of supply side conditions. In addition, unobserved price heterogeneity can be partially incorporated by estimating random coefficients on the pricing term. However, Griffith et al. (2018) has shown that random coefficients specifications may have strong implications for the estimated rate of pass through (i.e., the change in prices resulting from a cost shock) (see also Weyl and Fabinger (2013)). This can influence findings related to, for example, the welfare effects of price-discrimination (Aguirre et al. (2010)), the impact of mergers (Jaffe and Weyl (2013)), and the quantification of cartel damages: Verboven and van Dijk (2009).

The method we propose has many advantages. First, it allows the researcher to incorporate price heterogeneity when the researcher can't (as in the case of illicit markets) or doesn't want to model the supply side. Second, it generates an implied price faced by purchasers and non-purchasers in a symmetric way (which is relevant for computing counterfactuals). Third, it allows researchers to obtain an individual price (that is not driven by random coefficients) while not having access to micro-level pricing data. Finally, the econometric methodology properly addresses the issue of unobserved individual prices by integration, which is critical for unbiased policy recommendations.

The paper proceeds as follows. In the next section we present the model and econometric methodology. In section 3 we present the results from Monte-Carlo experiments and compare our results to those from traditional estimation methods. We present the results from an application to the market for marijuana in section 4. We then conclude.

2 The Model

We propose a method to incorporate individual prices in the econometric model when the researcher observes aggregated pricing and demographic information. We first describe how empiricists incorporate unobserved consumer attributes when they don't observe individual purchase decisions in a random coefficients logit model for aggregate demand data. We then discuss a method for incorporating individual prices that is similar in spirit.

The random coefficients logit model for aggregate demand data Consider a model with T markets, $t = 1, ..., T$. In each market t we have I_t potential consumers. Each consumer i may choose one of $J + 1$ differentiated products: $j = 0, ..., J$, where $j = 0$ denotes the outside good (good 0). Consumer i's conditional indirect utility is:

$$u_{ijt} = x_{jt}\beta_i + \alpha_i p_{jt} + \xi_{jt} + \varepsilon_{ijt}, \tag{1}$$

where x_{jt} denotes a $1 \times K$ vector of observed product characteristics, p_{jt} is the price of product j, β_i denotes a $K \times 1$ vector of random coefficients capturing individual-specific preferences for the product characteristics, and ξ_{jt} is the unobserved product characteristic. Note that all consumers are assumed to face the same product characteristics, in particular the same price. The indirect utility from the outside good is, as standard practice, normalized to zero: $u_{i0t} = \varepsilon_{i0t}$.

The distribution of consumers' preferences can be modelled as:

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \overline{\alpha} \\ \overline{\beta} \end{pmatrix} + \Pi D_i + \Sigma \nu_i \tag{2}$$

$$D_i \sim P_D(D) \tag{3}$$

$$\nu_i \sim P_v(\nu) \tag{4}$$

where D_i is a $H \times 1$ vector of demographic variables, $P_D(D)$ an empirical distribution, Π a $(K+1) \times H$ matrix of coefficients that measures how preferences vary with demographics, ν_i is a $(K+1) \times 1$ vector of unobserved consumer valuations, which are often drawn from parametric distributions such as a standard normal: $\nu_i \sim N(0, I_{K+1})$.

Equation (1) is typically rewritten as the sum of three terms:

$$\begin{aligned} u_{ijt} &= \delta_{jt}(\, x_{jt}, p_{jt}, \xi_{jt}) + \mu_{ijt}(x_{jt}, D_i, \nu_i) + \varepsilon_{ijt}, \\ \delta_{jt} &= x_{jt}\beta + \xi_{jt} \\ \mu_{ij} &= x_{jt}(\Pi D_i + \Sigma \nu_i), \end{aligned}$$

where the mean utility δ_{jt} does not vary across consumers and is a function of product characteristics $(x_{jt}, p_{jt}, \xi_{jt})$. Consumers exhibit heterogeneity in purchase decisions which is expressed by $\mu_{ij} + \varepsilon_{ijt}$. In particular, this term is a function of the demographics of the individual (D_i): in particular, it allows different consumer types (based on demographics and unobservable heterogeneity) to have different tastes for product characteristics, as captured by the parameter matrix Π and Σ. This framework allows us to capture household level variation in purchase decisions even though we do not observe the choices made by a particular consumer.

To estimate this model, the econometrician "draws" consumers from the empirical distribution $P_D(D)$, which is usually a household survey, and a parametric distribution $P_v(\nu)$. Given that the market shares of consumers are simulated, this approach does not yield a closed form solution for the market shares. Market share is generated by integrating over the empirical distribution of individuals, $P(D, \nu)$:

$$s_{jt} = \int \frac{\exp(\delta_{jt} + \mu_{ijt})}{1 + \Sigma_{rt} \exp(\delta_{rt} + \mu_{irt})} dP(D, \nu).$$

Unobserved Individual Prices We propose a method for incorporating individual prices that is similar in spirit. Consider a situation where the researcher observes information about the individual (D_i) but does not observe the price that the individual paid. Our framework incorporates this unobserved price by drawing a price for each individual from an empirical price distribution (which may be observed or generated from data). In estimation, the method involves integrating out over this empirical price distribution when computing the market share. We exploit two sources of information: market level information on prices, characteristics, and manufacturers' prices; and individual-level information on prices and choices by demographic group. Such information is often present in surveys. For example, Simmons Survey is an individual-level dataset that contains information on demographics of individuals and computer purchases of these same individuals. However, these data are not sufficient to estimate a model of PC demand because they only contain information on the manufacturer of the PC that was purchased. In the above notation, product j would correspond to the manufacturer and the price would be the price paid by the individual for PCj. Another example is the CAMIP survey data in Berry et al. (2004) that contains micro data on households' choices, characteristics and transaction prices. Differently from us, the authors use the modal vehicle price to construct the vehicle prices, so that all individuals face the same price.

We first generate an "empirical" price distribution based on the average and standard deviation of prices by demographics D_t, using the information income quartiles and types. In short, instead of using list prices, we *draw* a price $\widehat{p}_{ijt}$ for each individual and product from an empirical *simulated* price distribution $\widehat{P}_t(D_{it}, p_{jt})$, which is generated to reflect the entire distribution of product prices. That is, the empirical price distribution does not exist, but is itself formed by combining information from data on consumer characteristics (within a certain market) and linking these to the price distribution (in the same markets). To construct this "empirical" distribution we can use the average and standard deviation of the market-level prices for each product $j = 1, 2, ..., J$ and summarized in vector $\overline{p}_t = \{\overline{p}_{jt} : j = 1, 2, ..., J\}$. Further we can leverage on individual-level data, for example from a consumer survey, on products purchased by some (perhaps aggregated) characteristics. Our aim is to exploit these observed quantities to construct an empirical distribution for the price that an individual faces, $p_{it} \sim \widehat{P}_t(p_{it})$, taking into account the consumption of products and price differences across products.

Distributions of prices for each product in the market, denoted $F_p(p_{ijt})$, can be specified as either a nonparametric distribution, or a parametric distribution with the parameters estimated from the consumer-level data. After prices are drawn, we specify the indirect utility of consumer i from buying product j in year t as follows:

$$u_{ijt} = x_{jt}\beta_i + \alpha_i \widehat{p}_{ijt} + \xi_{jt} + \varepsilon_{ijt}, \tag{5}$$

where $\widehat{p}_{ijt}$ is now individual-specific. The model allows consumer preferences to vary according to individual characteristics, which can be observed demographics D_i, or unobserved, ν_i. Such heterogeneity can be modeled as above in equation (2). As we will see in the next sections, the way we model unobserved heterogeneity in prices has strong implications for the results, both in terms of parameter estimates and substitution patterns. The literature has been very clear on the advantages of letting the taste parameters vary with the observed demographics, as it allows to include additional information on the demographics in the modelling framework, and to reduce the reliance on parametric assumptions (Nevo (2001)). We will show that drawing the unobserved individual prices from a distribution relying on the observed demographics achieves the same advantages in terms of flexibility.

All these assumptions generate individual choices and market shares. Given that ε_{ijt} is i.i.d. extreme value, the individual choice probability is a random coefficient logit:

$$\widehat{s}_{ijt} = \int \frac{\exp(\delta_{jt}(x_{jt}, \xi_{jt}) + \alpha \widehat{p}_{ijt} + \mu_{ijt})}{1 + \Sigma_{rt} \exp(\delta_{rt}(x_{rt}, \xi_{rt}) + \alpha \widehat{p}_{irt} + \mu_{irt})} dP(D, \nu), \tag{6}$$

where $P(\cdot)$ denotes the population distribution function.

We can compute the integral in the equation (6) and predict the individual market shares for each product. After, we can choose the parameters to minimize the distance between the predicted market shares and the individual choices

3 Monte Carlo experiment

3.1 The data generating process

We consider a set of experiments in which we generate data according to a random coefficients logit model with individual-level data. For each experiment, we generate 100 datasets with $T = 10$ markets, $J = 25$ products per market, and $I = 500$ consumers per market. Each dataset consists of a constant; an observed product-specific attribute x_{jt}^1 drawn from a uniform distribution; an unobserved (by the econometrician) product-specific attribute ξ_{jt} drawn from a normal and uncorrelated with prices p_{ijt} : we abstract from the issue of endogeneity of prices, since we want to focus on comparing the performance between using individual prices and aggregated list prices. Nevo (2000) and Berry (1994) note that if different consumers face different prices, using either a list or average transaction price will lead to measurement error bias: that is yet an additional reason to use instrumental variables as prices may be correlated with the error term. However, the authors also note that instrumentation as proposed by Berry (1994) and Berry et al. (1995) can deal with measurement error only if the variable measured with error enters in a restrictive way, namely, the part of utility that is common to all consumers (δ_{jt} as defined below). This will not be the case in the random coefficient logit model in which we allow for heterogeneity in price sensitivity and prices that vary across individuals, hence instruments will not be able to address the issue of insufficient pricing data.[3] Finally, we draw an i.i.d. individual and product-specific unobservable ε_{ijt} from a Type-1 extreme value distribution. We set the mean valuation of the product characteristics $x_{jt} = (1, x_{jt}^1)$ equal to $(-3; 2)$.

In contrast with the usual assumptions, consumers are offered different prices, which in turn depend on their socio-demographic characteristics. We construct individual and product-specific prices p_{ijt} as follows: (i) we draw income and a 1-0 "type" (examples are gender, age category, ethnicity, education level) characteristics for each individual: those characteristics do not vary across simulations; (ii) for each simulation, we draw product-specific list prices (for instance, manufacturer's suggested retail price) p_{jt}^L from a normal

[3] In future work, we will explicitly consider the correlation between ξ_{jt} and p_{ijt} and check the performance of instruments in dealing with measurement error.

distribution $N(2, 0.5)$ to ensure positive realizations of the draws; (iii) we relate the list prices with the simulated socio-demographic characteristics to generate individual prices p_{ijt}. In particular we treat the dummy variable for type S_{it} as the realization of a latent continuous variable S^*_{it} and we draw income (M_{it}) and type (S^*_{it}) from a multivariate lognormal distribution as follows:

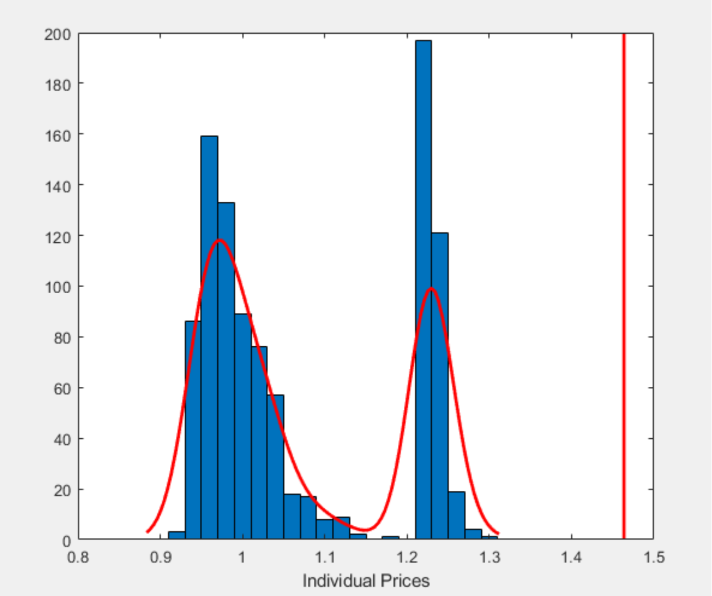
$$\begin{pmatrix}\ln M_{it}\\ S^*_{it}\end{pmatrix} \sim N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}1 & \varsigma_{MS}\\ \varsigma_{MS} & 1\end{bmatrix}\right),$$

and $S_{it} = 1_{\{S^*_{it} > \gamma\}}$. In the Monte Carlo design, we set $\varsigma_{MF} = 0.9; \gamma = -0.4$: these settings ensure that type 1 is more numerous than type 0 and enjoys, on average, a higher income. Individual prices are constructed as:

$$p_{ijt} = p^L_{jt} - \zeta_G F_{it} - \zeta_M(\overline{M}_{it} - M_{it}),$$

where ζ_G denotes a discount given to type 1 and $\zeta_M(\overline{M}_{it} - M_{it})$ a discount decreasing in individual income, with $\overline{M}_{it}$ the highest income draw: we set $\zeta_G = 0.2; \zeta_m = 0.03$. These assumptions generate a bimodal distribution of individual prices: figure 1 shows an example of individual price distribution for a particular product/market/simulation draw.

Figure 1: Distribution of Individual Prices and List Price



The figure reports the histogram and kernel density overlay for individual prices of product 1 in market 1. The vertical bar (on the right) represents the list price for the same product.

Similarly to equation (5), the indirect utility of consumer i from buying product j in year t is given by:

$$u_{ijt} = \underbrace{x_{jt}\beta + \xi_{jt}}_{\delta_{jt}} + \alpha_i p_{ijt} + \varepsilon_{ijt}, \tag{7}$$

where δ_{jt} includes all the product-specific characteristics. We model α_i as follows:

$$\alpha_i = \overline{\alpha} + \Pi D_i + \Sigma \nu_i, \tag{8}$$

where D_i is a vector of draws from the empirical distribution of income and ν_i is a vector of unobserved standard normal consumer valuations; parameter $\overline{\alpha}$ captures the mean valuations of price; Π is a 1×1 matrix describing how the valuations for price vary with income; and Σ is a 1×1 scaling matrix capturing unobserved heterogeneity in the valuations for price. We specify a DGP in which $\overline{\alpha} = -1.5; \Pi = 0.5; \Sigma = 0.25$. To reduce the number of parameters to be estimated, we assume that the distributions $P(D)$ and $P(\nu)$ are independent.

For each design and its associated 100 datasets, we estimate: (i) the correctly specified model; (ii) a product-level logit, with and without random coefficients, using list prices; (iii) a product-level logit, with and without random coefficients, using list prices with micro moments based on demographics, in the spirit of Berry et al. (2004); (iv) a model in which we generate the empirical price distribution drawing from a distribution matching the mean and standard deviation by demographic characteristics. The random coefficient logit models (ii-iii) are the benchmark models for empiricists.

For each design, we use GMM to estimate both the correctly specified model and the other, misspecified models. We now specify the GMM estimator in each model.

Correctly specified model We use two sets of moment conditions. The first set is generated from the choice probabilities as:

$$g^1_{ijt} = (d_{ijt} - \widehat{s}_{ijt}(\theta))z_{ijt},$$

where d_{ijt} is the dependent variable (equal to 1 if consumer i in market t chooses product j and 0 otherwise); z_{ijt} is a vector of instruments that vary over products and markets as well as over consumers in each market: as individual prices p_{ijt} are not endogenous in our specifications, prices are adequate instruments.

The second set of moment conditions consists of product-level moment conditions for the identification of the parameters of the constants for each product, δ_{jt} :

$$g^2_{jt} = \xi_{jt}(\theta)w_{jt} = (\delta_{jt}(\theta) - \beta x_{jt})z_{jt}, \tag{9}$$

where z_{jt} varies over products and markets: we use the vector of product characteristics x_{jt}. Given the dimensionality of δ (25×10), we estimate it using the contraction mapping procedure (Berry et al. (1995)) within the iterative process for the other parameters.

We stack the two sets of moments into one vector: $g_{ijt} = \binom{g^1_{ijt}}{g^2_{jt}}$, with the second set of moments repeated for each consumer i in market t. The moment conditions can then be written as $g = \sum_t \sum_j \sum_i g_{ijt} = 0$. The GMM estimator is the parameter value that minimizes the quadratic form $g' \cdot A \cdot g$.[4] We also estimate the model using the Maximum Simulated Likelihood estimator with and without the contraction. As expected, results are identical for a mode without random coefficients (the micro logit model), as the moment conditions are the first-order conditions for maximum likelihood and predicted shares equal sample shares. For the random coefficient logit model, these two features do not hold, but results are very close to the ones obtained using the GMM estimator.

Random coefficient logit The aggregate moments consist of the usual set of moment conditions proposed by Berry et al. (1995) to estimate aggregate product differentiated demand systems as in equation (9). As standard practice, we use list prices to calculate the market shares by product and market.

Random coefficient logit with micro-moments Following Berry et al. (2004), we add a second set of moment conditions that makes use of additional micro-level information. We assume that we observe not only the aggregate market shares of products by year and market, but also the market shares by demographic groups. In particular, we assume that we observe market shares for each car and consumer demographics (income and type), so we can calculate the average characteristics of consumers who have purchased a certain product: for example, the average number of consumers of type 1 who have purchased a product. We match this information with the sales by income and by type predicted by

[4] We use an approximation to the weighting matrix given an intial guess of the parameters, defined as follows:

$$A = \left(\sum_t \sum_j \sum_i g_{ijt}(\theta) g_{ijt}(\theta)' \right)^{-1}.$$

the model. This amounts to the following sample moment conditions:

$$g^{3}_{Djt} = I_t(s^{obs}_{ijt} - \widehat{s}_{ijt}(\theta))D_t \tag{10}$$

When satisfied, this moment conditions imply that the observed mean of the demographic for the choice alternative $\sum_{j} s^{obs}_{Djt} D_t / I_t$ is equal to the mean predicted by the model $\sum_{j} \widehat{s}_{Djt} D_t / I_t$. Subsequently we stack the set of moments defined in equation (9) and equation (10) into one vector and use the GMM estimator to recover the parameters.

The empirical price distribution As explained above, we generate an "empirical" price distribution: instead of using list prices, we *draw* a price $\widehat{p}_{ijt}$ for each individual and product from an empirical *simulated* price distribution $\widehat{P}_t(D_{it}, p^{D}_{jt})$, which is generated to reflect the entire distribution of product prices. In our setting, to construct this "empirical" distribution we use the average and standard deviation of the market-level prices by demographic, type and income quartile, for each product $j = 1, 2, ..., J$ and summarized in vector $\overline{p}_t = \{\overline{p}^{D}_{jt} : j = 1, 2, ..., J\}$. Further we assume that we observe individual-level data on product choices.

In our simulations, we specify distributions of prices for each product in the market, denoted $F_p(p_{ijt})$ as normals with the means set at the observed market averages by demographic and variances set using information by market:

$$p_{ijt} \sim F_p(p_{ijt}) \ , \ F_p(p_{ijt}) = N(\overline{p}_{jt}, \Omega^{p,D}_{jt}) \ \text{ for } \ j = 1, ..., J.$$

After prices are drawn, we use them to calculate the individual market shares by product and market and form the moment conditions following the same steps outlined above for the correctly specified model.

3.2 Parameter estimates

We compare the performance of the three estimation methods against the correct model. Table 1 reports the estimation results for the data generating process in which heterogeneity is modelled as described in equation (2). The parameter estimates for the correctly specified model are, as expected, very close to the true parameters. This confirms that our estimation procedure works well in practice. Next, consider the parameters of the misspecified models. The price parameter is estimated with a downward bias (in magnitude) in the misspecified logit and random coefficient logit model using the incorrect list prices p_{jt}^{L}: consumers avoid the higher prices less than they would thanks to the discount. In contrast, in the logit model with additional micro moments using list prices, the estimated price parameter presents a strong upward bias (in magnitude). Finally, in our proposed empirical price distribution which constructs the individual prices on the basis of the demographics, the estimated parameters are close to the true ones.

Table 1: Monte Carlo results: Parameter Estimates

Parameters	True	Micro Logit	Aggr Logit	Aggr RC Logit	Micro RC Logit	Micro Logit
		Ind price	List price	List price	List price	Emp. price
β	2.00	2.08	2.07	2.08	2.08	2.08
		(0.08)	(0.09)	(0.10)	(0.10)	(0.08)
α	-1.50	-1.50	-0.60	-0.90	-1.31	-1.51
		(0.04)	(0.06)	(0.77)	(0.59)	(0.06)
Π	0.50	0.50	n/a	0.10	0.47	0.51
		(0.08)		(0.58)	(0.42)	(0.10)
Σ	0.25	0.25	n/a	0.20	0.02	0.25
		(0.01)		(0.74)	(0.25)	(0.01)

The table reports the empirical means and standard deviations (in parentheses) of selected parameters. The estimates are based on a 100 random samples of 10 markets and 25 products. The true models is a model with individual prices.

3.3 Substitution patterns

It is most interesting to investigate the substitution patterns implied by our estimates. The logit model ignores heterogeneity across consumers and is clearly inferior with respect to all the other models. The random coefficient logit model yields to substitution patterns that appear to be more biased compared to our method, although the use of additional micro moments is very helpful in reducing the bias. Moreover, because of the use of list prices, the substitution patterns across individuals are more susceptible to be driven by functional form assumptions and tend to be particularly biased by demographic.

Table 2: Monte Carlo results: Substitution Patterns

	Own elasticity Product 1				
	Mean	St.Dev	10th	Median	90th
Micro Logit - True	-1.309	0.777	-2.257	-1.372	-0.302
Aggr Logit - List price	-1.231	0.000	-1.231	-1.231	-1.231
BLP RC Logit - List price	-1.340	0.479	-1.874	-1.422	-0.712
Micro RC Logit - List price	-1.382	0.809	-2.093	-1.701	-0.273
Micro Logit - Emp price	-1.299	0.790	-2.263	-1.359	-0.264
	Cross elasticity Product 1-2				
Micro Logit - True	0.055	0.030	0.015	0.061	0.086
Aggr Logit - List price	0.060	0.000	0.060	0.060	0.060
BLP RC Logit - List price	0.057	0.023	0.029	0.061	0.083
Micro RC Logit - List price	0.057	0.036	0.010	0.072	0.087
Micro Logit - Emp price	0.054	0.031	0.013	0.060	0.086

The table reports means and standard deviations of the own-elasticity of product 1 and the cross-elasticity between product 1 and product 2, as well as the 10th, 50th, and 90th percentiles across consumers ($I = 500$). The estimates are based on a 100 random samples of 10 markets and 25 products. The true models is a model with individual prices.

4 Application

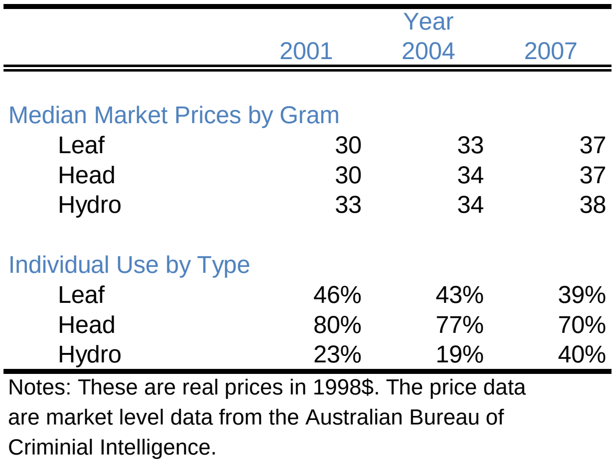
It is easiest to understand the approach in the context of an example. Our example is derived from and follows closely Jacobi and Sovinsky (2016) (hereafter JS). That paper focuses on predicting the demand for marijuana. JS observe individual consumption of marijuana but not how much the individual paid. They construct an empirical price distribution for marijuana, which they use to generate an implied price faced by users and non-users. This allows JS to estimate a model with individual prices while not observing these in the data. We first discuss the data and then the framework.

The data in Jacobi and Sovinsky (2016) are from two sources. The first are individual-level cross-section data from the Australian National Drug Strategy Household Survey (NDSHS). The NDSHS was designed to determine the extent of drug use among the non-institutionalized civilian Australian population aged 14 and older. These data are particularly useful as they contain demographic, market, and illicit drug use information. The second are market-level pricing data collected from drug seizures by the Australian Bureau of Criminal Intelligence. These data consist of prices for all drug busts made in that region time period.

The major psychoactive chemical compound in marijuana is delta-9-tetrahydrocannabinol (or THC). The amount of THC absorbed by marijuana users differs according to the part of the plant that is used (e.g., leaf, head), the way the plant is cultivated (e.g., hydro), and the method used to imbibe the drug. On average marijuana contains about 5% THC, where the flowering tops contain the highest concentration followed by the leaves (Adams and Martin, 1996). Marijuana that is grown hydroponically (hydro), indoors under artificial light with nutrient baths, typically has higher concentrations of THC relative to naturally grown leaf and head (Poulsen and Sutherland, 2000).

The NDSHS survey contains information about which form of marijuana the user uses (leaf, head or hydro). Table 1 presents median market prices (across the country) and individual percentage of use per type by year.[5] Given the higher amount of THC present in hydro it demands a higher price.

[5] It is common to use types in combination (i.e., a bag might contain leaf and head), hence the percentages do not sum to one.



		Year	
	2001	2004	2007
Median Market Prices by Gram			
Leaf	30	33	37
Head	30	34	37
Hydro	33	34	38
Individual Use by Type			
Leaf	46%	43%	39%
Head	80%	77%	70%
Hydro	23%	19%	40%

Notes: These are real prices in 1998$. The price data are market level data from the Australian Bureau of Criminial Intelligence.

Table 1: Prices and Use by Type (source Jacobi and Sovinsky, 2016)

An individual i chooses whether or not to consume marijuana in market m (which is a state-year combination). The indirect utility is given by

$$U_{im} = p_{im}\alpha + f(d_i, x_m, L_{im}) + \varepsilon_{im}, \qquad p_{im} \sim \widehat{P}_m(p_{im})$$

which depends on a function of demographic characteristics d_i (which we observe), market specific variables x_m, variables related to the legal status L_{im}, and an idiosyncratic error term ϵ_{im}.[6]

The indirect utility also depends on the price the individual pays (p_{im}). However, we do not observe these prices in the data. The common approach is to assign each consumer an average price for the product in that market. This approach has a few drawbacks. First, by using the average across markets the strategy precludes price variation within the market. Second, some consumers may prefer different (quality) types of products and hence face systematically different prices.

An alternative approach is to use additional data on the price distribution (for each product type) and draw a price for each consumer from this distribution. As we mentioned

[6] Individuals have utility from not using marijuana, which we model as $U_{im0} = \alpha_0 + \epsilon_{im0}$,where all non stochastic terms are normalized to zero, because we cannot identify relative utility levels.

above, we have information on the distribution of the prices from the data as well as what types each consumer uses. We construct an "empirical" price distribution ($\widehat{P}_m(p_{im})$) by exploiting prevalences on the type of marijuana used and market-level price data. In short, instead of using a weighted average product price, we *draw* a price for each individual from an empirical *simulated* price distribution, which is generated to reflect the entire distribution of product prices. That is, the empirical price distribution does not exist, but is itself formed by combining information from data on consumer characteristics (within a certain market) and linking these to price distributions (in the same markets).

To construct this "empirical" distribution we use the average market-level marijuana prices ($\overline{p}_{mt}$) for each type $t = 1, 2, 3$ (leaf, head, hydro) summarized in the vector $\overline{p}_m = \{\overline{p}_{m,leaf}, \overline{p}_{m,head}, \overline{p}_{m,hydro}\} = \{\overline{p}_{mt} : t = 1, 2, 3\}$. These are based on the prices reported by the Australian Bureau of Criminal Intelligence. Further we observe which type of marijuana an individual uses (from NDSHS). Using these data we construct market level probabilities of using a type, $\overline{\pi}_m = \{\overline{\pi}_{m,leaf}, \overline{\pi}_{m,head}, \overline{\pi}_{m,hydro}\} = \{\overline{\pi}_{mt} : t = 1, 2, 3\}$.[7]

Our aim is to exploit these observed quantities to construct an empirical price distribution that an individual faces, $p_{im} \sim \widehat{P}_m(p_{im})$, taking into account the consumption of the three types and price differences across types. We specify distributions of prices and probabilities of use for each type by market, denoted $F_p(p_{imt})$ and $F_\pi(\pi_{imt})$, respectively

[7] A different example: Consider a market (that consists of consumers with different demographics) where we have some average market-level prices by product (characteristic) type $t = 1, 2, ..., T$ and summarized in vector $\overline{p}_m = \{\overline{p}_{mt} : t = 1, 2, ..., T\}$. Further we often observe individual-level data from a consumer survey on products purchased by characteristic. Based on these responses for all individuals in a market, we can construct market level probabilities of buying each product type in each market, $\overline{\pi}_m = \{\overline{\pi}_{mt} : t = 1, 2, ..., T\}$. Our aim is to exploit these observed quantities to construct an empirical distribution for the price that an individual faces, $p_{im} \sim \widehat{P}_m(p_{im})$, taking into account the consumption of the product types and price differences across types. For example, Simmons Survey is an individual-level dataset that contains information on demographics of individuals and computer purchases of these same individuals. However, these data are not sufficient to estimate a model of PC demand because they only contain information on the manufacturer which made the PC that was purchased. In the above notation the "type" of product would correspond to the manufacturer and the price would be the price paid by the individual for that PC of type t.

as truncated normals, where

$$\begin{array}{rcl} p_{imt} & \sim & F_p(p_{imt}) \ , \ F_p(p_{imt}) = TN_{(0,\infty)}(\overline{p}_{mt}, \Omega^p_{mt}) \ \text{ for } \ t = 1,2,3 \\ \pi_{imt} & \sim & F_\pi(\pi_{imt}) \ , \ F_\pi(\pi_{imt}), = TN_{(0,\infty)}(\overline{\pi}_{mt}, \Omega^\pi_{mt}) \quad s.t. \ \sum_t \pi_{imt} = 1. \end{array}$$

with the means set at the observed market averages and variances set using information across all markets. Assuming that the “average” price (p_{im}) an individual faces depends on the relative use of each type we then define this price as an average of the prices over the three different types weighted by their respective use probabilities

$$p_{im}|\pi_{imt}, p_{imt} = \sum_{t=1}^{3}(\pi_{imt} * p_{imt}).$$

The price p_{im} reflects the average price faced by individual i in market m based on draws from the market and type specific distributions of price and the probability of use. The implied marginal empirical distribution of price for individuals in a market is given by

$$\widehat{P}_m(p_{im}) = \int \sum_{t=1}^{3}(\pi_{imt} * p_{imt}) \ dF_p(p_{imt}) \ dF_\pi(\pi_{imt})$$

assuming independence in the distributions across types and across prices.

Assuming the individual has access to marijuana, the probability i chooses to use marijuana in market m (the individual market share) is given by

$$\begin{array}{rcl} S_{im} & = & \int_{R_{im}} dF_{\epsilon,p}(\epsilon, p) \\ & = & \int_{R_{im}} dF_\epsilon(\epsilon) d\widehat{P}_m(p_{im}), \end{array}$$

where R_{im} is the set of variables that results in consumption of marijuana given the parameters of the model, $F(\cdot)$ denotes a distribution function, and $\widehat{P}_m(p_{im})$ represents the market-specific empirical price distribution. The latter equality follows from independence assumptions,

This method of generating individual prices from an empirical distribution improves upon the typical approach in the literature that uses average market prices as those do not vary within a market neither by type used nor probability of use of each type, whereas this method generates a distribution of prices in each market. Importantly, this approach also allows the researcher to obtain the implied price faced by users and non-users in a symmetric way and to properly address the econometric issue of unobserved individual prices in estimation by integration.

4.1 Results

Ongoing

5 Conclusions

Missing or incomplete data is a common problem faced by empirical researchers. Empirical economists have addressed it using a variety of techniques including alternative datasets and/or modeling. We propose a framework for overcoming inadequate information about individual transaction prices.

Using the framework we propose, the researcher can include price heterogeneity from individual prices while not observing these in the data. This is relevant in many situations as many markets are characterized by heterogeneity in transaction price, but if it is not possible to incorporate it due to insufficient data, then the estimates can lead to biased price elasticities and incorrect policy conclusions.

We show, via Monte Carlo experiments, that our method of generating prices from a simulated empirical distribution improves upon the typical approach that uses average prices (over some dimension) for the same individual. A major benefit to this approach is that it properly address the econometric issue of unobserved individual prices by integration. In addition, it allows the researcher to generate an implied price faced by purchasers and non-purchasers in a symmetric way (which is relevant for computing counterfactuals).

References

Aguirre, Iñaki, Simon Cowan, and John Vickers, "Monopoly Price Discrimination and Demand Curvature," *American Economic Review*, sep 2010, *100* (4), 1601–1615.

Allen, Jason, Robert Clark, and Jean-François Houde, "The Effect of Mergers in Search Markets: Evidence from the Canadian Mortgage Industry," *American Economic Review*, oct 2014, *104* (10), 3365–3396.

_ , _ , and _ , "Search Frictions and Market Power in Negotiated-Price Markets," *Journal of Political Economy*, aug 2019, *127* (4), 1550–1598.

Berry, Steven T., "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 1994, *25* (2), 242–262.

_ , James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, *63* (4), 841–890.

_ , _ , and _ , "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, February 2004, *112* (1), 68–105.

Busse, Meghan, Jorge Silva-Risso, and Florian Zettelmeyer, "1,000 Cash Back: The Pass-Through of Auto Manufacturer Promotions," *American Economic Review*, 2006, *96* (4), 1253–1270.

Chandra, Ambarish, Sumeet Gulati, and James M. Sallee, "Who Loses when Prices are Negotiated? An Analysis of the New Car Market," *The Journal of Industrial Economics*, jun 2017, *65* (2), 235–274.

Crawford, Gregory S., "Endogenous product choice: A progress report," *International Journal of Industrial Organization*, 2012, *30* (3), 315–320.

Dafny, Leemore S, "Are Health Insurance Markets Competitive?," *American Economic Review*, sep 2010, *100* (4), 1399–1431.

D'Haultfœuille, Xavier, Isis Durrmeyer, and Philippe Février, "Automobile Prices in Market Equilibrium with Unobserved Price Discrimination," *The Review of Economic Studies*, oct 2018, *86* (5), 1973–1998.

Eizenberg, Alon, Andras Pechy, and Michelle Sovinsky, "The Dynamics of Technology Adoption and Vertical Restraints: an Empirical Analysis," 2019. 2019.

Gavazza, Alessandro, "An Empirical Equilibrium Model of a Decentralized Asset Market," *Econometrica*, 2016, *84* (5), 1755–1798.

Goldberg, Pinelopi Koujianou, "Dealer Price Discrimination in New Car Purchases: Evidence from the Consumer Expenditure Survey," *Journal of Political Economy*, jun 1996, *104* (3), 622–654.

Grennan, Matthew, "Price Discrimination and Bargaining: Empirical Evidence from Medical Devices," *American Economic Review*, feb 2013, *103* (1), 145–177.

Griffith, Rachel, Lars Nesheim, and Martin O'Connell, "Income effects and the welfare consequences of tax in differentiated product oligopoly," *Quantitative Economics*, mar 2018, *9* (1), 305–341.

Hastings, Justine, Ali Hortacsu, and Chad Syverson, "Sales Force and Competition in Financial Product Markets: The Case of Mexico's Social Security Privatization," *Econometrica*, 2017, *85* (6), 1723–1761.

Huang, Guofang, "When to Haggle, When to Hold Firm? Lessons from the Used Car Retail Market," *SSRN Electronic Journal*, 2017.

Jacobi, Liana and Michelle Sovinsky, "Marijuana on Main Street? Estimating Demand in Markets with Limited Access," *American Economic Review*, aug 2016, *106* (8), 2009–2045.

Jaffe, Sonia and E. Glen Weyl, "The First-Order Approach to Merger Analysis," *American Economic Journal: Microeconomics*, nov 2013, *5* (4), 188–218.

Miller, Nathan H. and Matthew Osborne, "Spatial differentiation and price discrimination in the cement industry: evidence from a structural model," *The RAND Journal of Economics*, may 2014, *45* (2), 221–247.

Nevo, Aviv, "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand," *Journal of Economics & Management Strategy*, December 2000, *9* (4), 513–548.

_ , "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, March 2001, *69* (2), 307–42.

Robles-Garcia, Claudia, "Competition and Incentives in Mortgage Markets: The Role of Brokers," Finance 3796, Stanford May 2019.

Scott-Morton, Fiona, Florian Zettelmeyer, and Jorge Silva-Risso, "Internet Car Retailing," *The Journal of Industrial Economics*, dec 2001, *49* (4), 501–519.

Town, Robert and Gregory Vistnes, "Hospital competition in HMO networks," *Journal of Health Economics*, sep 2001, *20* (5), 733–753.

Verboven, Frank and Theon van Dijk, "Cartel Damages Claims and the Passing-on Defense," *The Journal of Industrial Economics*, sep 2009, *57* (3), 457–491.

Weyl, E. Glen and Michal Fabinger, "Pass-Through as an Economic Tool: Principles of Incidence under Imperfect Competition," *Journal of Political Economy*, jun 2013, *121* (3), 528–583.

Williams, J., "The effects of price and policy on marijuana use: what can be learned from the Australian experience?," *Health Economics*, jan 2004, *13* (2), 123–137.