Forensic Econometrics: Demand Estimation when Data are Missing

Julian Hidalgo and Michelle Sovinsky*

September 18, 2018

Abstract

Often empirical researchers face many data constraints when estimating models of demand. These constraints can sometimes prevent adequate evaluation of policies. In this article, we discuss two such missing data problems that arise frequently: missing data on prices and missing information on the size of the potential market. We present some ways to overcome these limitations in the context of two recent research projects. Liana and Sovinsky (2018) which addresses how to incorporate unobserved price heterogeneity and Hidalgo and Sovinsky (2018) which focuses on how to use modeling techniques to estimate missing market size. Our aim is to provide a starting point for thinking about ways to overcome common data issues.

1 Introduction

Empirical economists often encounter issues of inadequate, partially available, or missing data. This confounds efforts to provide useful policy recommendations. An insightful exercise, to understand the magnitude of the problem, is to consider what data an empiricist interesting in assessing demand would use in an ideal world. First, he or she would use information about the product for sale, which (at a minimum) would include: the price, the description of the product (i.e., all observed characteristics), the (unobserved) quality of the product, and the costs of production. Second, he or she would use information about the consumer, such as: how much (or whether) the consumer would buy the product at each price point, the demographics of the consumer, how well the consumer knows the product, the marketing the consumer has seen about the product, and the other products the consumer is considering to buy. Finally, if the supply side matters for the policy analysis, the researcher would need information on the market that details: the other firms selling this product, how many other (and types of) products each firm sells, which products are available when, and how much the firms would sell at each price.

[0] *Hidalgo is at KU Leuven. Sovinsky is at University of Mannheim and CEPR. This project was supported by the European Research Council Grant #725081 FORENSICS (Sovinsky).

Even this, admittedly partial, "wish list" is unattainable. So, any empirical economist estimates demand models subject to a variety of data complications. Some of these issues have been addressed in the literature and solutions are readily available. For example, we often wish to predict consumer demand but have access only to aggregated sales data (e.g., on market shares). It is well known (e.g., see Berry (1994), Berry, Levinsohn, Pakes, 1995 (BLP), and Nevo (2000)) that product substitution patterns are likely to depend on household characteristics. As a consequence, the literature presents ways to incorporate unobserved consumer demographics into the demand framework. In many markets additional data on consumer characteristics are available (e.g., Consumer Population Survey in the US, Eurostat in the EU). Following the seminal paper of BLP, researchers take advantage of such data to incorporate interactions of product characteristics with individual characteristics, the latter of which are empirical draws from a household survey.

In addition, studies in the marketing literature (e.g. Chiang, et. al. (1999), Mehta et. al. (2003), Nierop et. al. (2010), among others) as well as those in the economics literature (e.g., Sovinsky Goeree (2008), among others) have developed methods to incorporate limited information on consumers choice sets into the demand framework. Sovinsky Goeree (2008) shows that this is particularly important in markets characterized by rapid change, as it is probable that consumers know only a subset of all available products, but we often don't observe which products they know. These models may rely on additional data to estimate the probability that a consumer is aware of a product, which may depend on marketing undertaken by the firm.

In this article, we discuss two other missing data problems that arise frequently: missing data on prices and missing information on the size of the potential market. Individual pricing data may be lacking because: purchase was illegal (hence the data are incomplete), the individual did not make a purchase (hence the data are necessarily missing) or, only aggregate pricing data are available (hence the data are imprecise). The second data problem is missing information on the size of the potential market. The standard approach is to use (a function of) population or information from household surveys. However, these survey data are not available for illicit markets (such as drugs or counterfeit markets) or are not reliable in legal markets in developing nations as it is more difficult to obtain timely population data.

We present alternatives that can be useful to overcome data hurdles related to missing

prices and missing market size. In section 2, we present an example based on Jacobi and Sovinsky (2016) to incorporate unobserved price heterogeneity while not observing purchase price. In section 3, we present ongoing work from Hidalgo and Sovinsky (2018) where we estimate the impact of a pricing subsidy on internet adoption while we do not observe the size of the at-risk population. In the last section we provide concluding thoughts.

2 Missing Data on Prices

In this section we present a method to incorporate individual prices in the econometric model when the researcher only has access to more aggregated pricing information. Our framework is related to the method empiricists use to incorporate unobserved consumer attributes without observing individual purchase decisions. For example, consider the simple model where the indirect utility consumer i receives from product j is

$$u_{ij} = \delta_j + \mu_{ij} + \epsilon_{ij},$$

with mean utility δ_j that is the same for all consumers and typically modeled as a function of prices and product characteristics (x_j). Consumers express heterogeneity in purchase decisions which is encompassed by $\mu_{ij} + \epsilon_{ij}$. The latter term is an idiosyncratic error term and in this example is represented by

$$\mu_{ij} = x_j^{'}(\Omega D_i).$$

Notice that this heterogeneity term is a function of the demographics of the individual (D_i) in particular it allows different types of consumers (based on demographics) to have different tastes for product characteristics (as captured by the parameter matrix Ω). Even though we don't observe the choices made by a particular consumer, the framework captures household level variation in purchase decisions.

In order to estimate this model the econometrician "draws" consumers from a empirical distribution (usually a household survey). This approach does not yield a closed form solution for the market shares as the consumers are simulated. The resulting market share is formed by integrating over the empirical distribution of individuals, $G_D(D)$,

$$s_j = \int \frac{\exp(\delta_j + \mu_{ij})}{1 + \Sigma_r \exp(\delta_r + \mu_{ir})} dG_D(D).$$

We propose a method for incorporating individual prices that is in a similar spirit. In our data we observe information about the individual (so we observe D_i) but we don't 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. In estimation we integrate out over this empirical price distribution when computing the market share.

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). In that paper, we are interested in predicting the demand for marijuana. We observe individual consumption but not prices. We construct an empirical price distribution for marijuana, which we use to generate an implied price faced by users and non-users. This allows us to estimate a model with individual prices while not observing these in the data. We first discuss the data and then the framework. More details on the method are ongoing in Jacobi and Sovinsky (2018).

The data we used 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.

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 market prices and the individual percentage of use per type by year. Given the higher amount of THC present in hydro it demands a higher price.

	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}.[1]

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 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,

[1] 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.

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, $\overline{\pi}_m = \{\overline{\pi}_{m,leaf}, \overline{\pi}_{m,head}, \overline{\pi}_{m,hydro}\} = \{\overline{\pi}_{mt} : t = 1, 2, 3\}$.

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 as truncated normals, where

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

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

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

where $F(\cdot)$ denotes a distribution function, the latter equality follows from independence assumptions, and $\widehat{P}_m(p_{im})$ represents the market-specific empirical price distribution.

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. Note that while the analytical form of the distribution is unknown, it can be easily approximated within a Bayesian estimation framework by a simple extension of the MCMC algorithm for the model estimation, essentially expanding the parameter space to include the vector of prices.

3 Missing Data on Market Size

Another data limitation may regard the size of the targeted population. Consider, for example, a setting in which the researcher wishes to evaluate the impact of a policy on demand. In order to do so the researcher would need to know the size of the underlying population. Usually this is easy information to obtain as some rough idea of the number of households is recorded in census data or in other household surveys. However, in some instances these data are not available - particularly in illegal markets or in developing markets.

To overcome this data limitation some papers have used existing data in a creative way. For example, Parey and Rasul (2017) wish to measure the size of the market for cannabis, but as this is an illicit market these data are not readily available. Instead they note that they can use data from a licit market (the market for smoking papers and tobacco) to approximate demand for the illicit market of cannabis. In developing nations market size data may be

unavailable as household surveys are less frequent. However, there are satellite images that provide some information on the location of residences (Sutton, 1997) and can be used to proxy population.

We provide another alternative in which we use both additional data and modeling to determine market size. In Hidalgo and Sovinsky (2018), we are interested in examining the impact of a subsidy for internet access available to low-income consumers in Colombia. The ultimate goal of the subsidy policy in Colombia is to close the digital divide and stimulate the adoption of residential Internet services among low income households. Hence, the price decrease seeks to change households' decisions and thereby to make no-adoption (the outside option) less appealing. To this end we are interested in the impact of the policy on reducing the share of the outside good. To evaluate this policy it is critical that we have a good measure of the relevant population (i.e., the market size).

As is more common in developing nations, it is not straightforward to get data on the number of individuals in narrow geographic regions by income.[2] However, in order to gauge the impact of the policy on take-up among low income households it is crucial to know the number of low income individuals who are impacted by the policy. So we face a missing data problem in that we don't observe the market size. This is further complicated by the nature of the program in that the internet service providers provide services that may be different even though the individuals live in the same municipality. That is, the plans are targeted according to income status. Hence, we need a measure of market size that varies by income and geographic unit (municipality in this case).

We note that the population of interest are those households that have not adopted internet services. Therefore the number of consumers depends on how internet access diffuses across populations. Based on this observation, in our approach we use models of innovation diffusion to estimate the size of the effected market. This yields an approximation of the market size for each municipality/income strata.

More specifically, we follow the literature (e.g., Geroski (2000) and Gruber and Verboven (2001)) and estimate a Griliches (1957) logistic model given by

[2] Typically, one could find information on the population size in a geographic region from a population survey, which often contains information on income as well. This can be used to generate a region specific measure of the number of households by income ranges. However, data surveys of this type are more difficult to come by in developing nations and, as we discussed in Hidalgo and Sovinsky (2018), Colombia is no exception.

$$y_{mt} = \frac{\mathcal{M}_m}{1 + \exp(-a_{mt} - b_{mt}t)}$$

where y_{mt} denotes the number of subscribers at the market level m in period t. The model is a function of time t and market m and has variables related to the number of potential adopters (the saturation level) ($\mathcal{M}_m$), the timing of diffusion (a_{mt}) and speed of diffusion (b_{mt}). The estimated variable $\mathcal{M}_m$ is then the measure of market size. In our setting we also observe information on number of households that have a computer or a fixed telephone line (denoted w_{mt}) so we estimate the market size as $\widehat{\mathcal{M}_m} = \widehat{\pi}_m w_{mt}$.

This approach allows us to learn about unobserved market size in a small geographic region by income. As a result, we are able to evaluate the impact of the pricing subsidy on closing the digital divide prevalent in developing nations.

4 Conclusions

Missing or incomplete data is a common problem faced by empirical researchers. We (empirical economists) have addressed it using a variety of techniques including alternative datasets and/or modeling. In this note, we examined two instances of missing data - on prices and market size.

Using the methodologies we outline here, 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. In our ongoing work we are examining our conjecture that this 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 generates an implied price faced by purchasers and non-purchasers in a symmetric way (which is relevant for computing counterfactuals) and it properly address the econometric issue of unobserved individual prices by integration.

We also provide some insights into dealing with missing data on market size. Some ways to address it include looking to other (related or complementary) markets to obtain additional data (as in Parey and Rasul, 2017) or incorporating a way to estimate the market

size. We provide one example from Hidalgo and Sovinsky (2018) in which we use both additional data and modeling techniques. In this ongoing work we plan to examine the impact of using less precise measures of market size in demand estimation.

Ultimately whether these methods apply depend on the particular market in which the research is focused . However, we hope that this article has provided a starting point for thinking about ways to overcome the unfortunately common data limitations present in our non-ideal world.

References

Adams, B. and B. Martin. 1996. "Cannabis: Pharmacology and Toxicology in Animals and Humans." *Addiction.* 91:1585-1614.

Australian Crime Commission Illicit Drug Data Report. 2000. Report Number 2000-01, Australian Crime Commission (ACC) URL: http://www.crimecommission.gov.au/publications/illicit-drug-data-report.

Australian Crime Commission Illicit Drug Data Report. 2003. Report Number 2003-04, Australian Crime Commission (ACC) URL: http://www.crimecommission.gov.au/publications/illicit-drug-data-report.

Australian Crime Commission Illicit Drug Data Report. 2006. Report Number 2006-07, Australian Crime Commission (ACC) URL: http://www.crimecommission.gov.au/publications/illicit-drug-data-report.

Berry, S. 1994. "Estimating Discrete Choice Models of Product Differentiation." *Rand Journal of Economics* 25(2): 242-262.

Berry, S., J.Levinsohn and A. Pakes 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63(4): 841-890.

Chiang, J., S. Chib and C. Narasimhan 1999. "Markov Chain Monte Carlo and Models of Consideration Set and Parameter Heterogeneity." *Journal of Econometrics* 89: 223-248.

Geroski, P., 2000. "Models of technology diffusion." *Research Policy* 29 (4–5): 603–625.

Griliches, Z. 1957. "Hybrid Corn: An exploration in the economics of technical change." *Econometrica* 25: 501-522.

Gruber, H and F. Verboven. 2001. "The diffusion of mobile telecommunications services in the European Union." *European Economic Review* 45, 577-588.

Hidalgo, J. and M. Sovinsky. 2018. "Internet (Power) to the People: The Impact of Demand-side Subsidies in Colombia." Working Paper, University of Mannheim.

Jacobi, L. and M. Sovinsky. 2016 "Marijuana on Main Street: Estimating Demand in Markets with Limited Access," *American Economic Review* 106(8): 2009-45

Jacobi, L. and M. Sovinsky. 2018. "Incorporating (Unobserved) Price Heterogeneity." Mimeo, University of Mannheim.

Mehta, N., S. Rajiv and K. Srinivasan. 2003. "Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation." *Marketing Science* 22(1): 58-84

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

Nierop, E., R. Paap, B.Bronnenberg, P. Franses, and M. Wedel. 2010. "Retreiving Unobserved Consideration Sets from Household Panel Data." *Journal of Marketing Research* 47(1): 63-74.

Parey, M. and Rasul, I. (2017) 'Measuring the Market Size for Cannabis: a New Approach Using Forensic Economics.' *Economica* ISSN 0013-0427 (In Press)

Petrin, A. 2002. "Quantifying the Benefits of New Products: The Case of the Minivan " *Journal of Political Economy* 110(4):705-729.

Poulsen, H. and G. Sutherland. 2000. "The Potency of Cannabis in New Zealand from 1976 to 1996." *Science and Justice* 40:171-176.

Sovinsky Goeree, M. 2008. "Limited Information and Advertising in the US Personal Computer Industry." *Econometrica* 76(5): 1017-1074.

Sutton, P. 1997. "Modeling population density with night-time satellite imagery and GIS" *Computers, Environment and Urban Systems* 21(3–4): 227-244.