

Structural Workshop Paper

Data Selection and Procurement

Carl F. Mela

Fuqua School of Business Administration, Duke University, Durham, North Carolina 27708,
mela@duke.edu

In this note I overview the data selection and procurement process in the context of structural models. Data selection for structural models presents unique challenges because data and structure often substitute and because it is imperative to consider what information identifies causal effects of interest.

I further discuss three types of field data on which to build empirical models: (i) data that are proprietary to firms, (ii) data that can come from the public domain, or (iii) data that can be purchased from private research firms, and I discuss the benefits and limits of each. I then detail a process for obtaining proprietary data and the potential pitfalls inherent in the process.

Key words: structural models; data

History: Received: November 16, 2010; accepted: March 18, 2011; Eric Bradlow served as the editor-in-chief and Jean-Pierre Dubé served as associate editor for this article. Published online in *Articles in Advance* June 6, 2011.

Now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it. (Varian 2009)

1. Overview

Equipped with nothing but a camera and a Berry et al. (1995) model in hand, Raphael Thomadsen photographed fast-food menu pricing and inferred how prices affect demand in geographically differentiated industries, even in the absence of sales data (Thomadsen 2005, 2007). This example illustrates the interplay between structural models and the data used to estimate them. Because structure interacts with data, structural models require thoughtful consideration of the phenomenon modeled and how the information used to estimate the model is sufficient to uncover the theoretical parameters of interest.

As defined in Reiss (2011) and Reiss and Wolak (2007), structural models use economic or behavioral theory to develop a mathematical relationship between endogenous outcomes and explanatory variables (both observable and unobservable). A structural model might proceed by delineating the agents whose behavior is being considered, their objective functions (e.g., maximize utility, revenues, or profits), their information sets (e.g., how much they know about competitors' decisions) and how these evolve over time, and the rules of the game they play (e.g., whether decisions are made simultaneously or sequentially). These factors would suggest a set of best decisions for the agents (or decision rules in

dynamic games) conditioned on their objectives and the environment (states).

Estimation proceeds by choosing parameters to match the agents' predicted behaviors obtained by the underlying economic theory with those observed in the data. In the Thomadsen (2005) example, firms set prices conditioned on demand. When customers are price sensitive, firms lower prices. In this fashion, observed firm prices are informative about the nature of the unobserved underlying consumer preferences. By coupling economic theory with data, Thomadsen (2005) can therefore infer price response even with no data on consumer choices. This case highlights one way in which structural models are unique in the context of data acquisition and procurement; structure can be used to infer missing information. This illustration also highlights the tension between data and structural models. With data on demand, Thomadsen (2005) might not need the supply-side structural assumptions to infer it.¹

Reasons for using a structural approach include (i) enabling or improving the efficiency of inference of model primitives, (ii) creating empirical tests of alternative theories, and (iii) counterfactual analyses to

¹ Raphael collected the data for Thomadsen (2005) while a doctoral student at Stanford in 1999; thus those collecting data for structural modeling and estimation should be aware of the long latency between data collection and publication. Also of interest, the data and model in Thomadsen (2005) were foundational to Thomadsen (2007), a John D. C. Little Award finalist at *Marketing Science* who inspired Duan and Mela (2009).

improve firm decision making, consumer welfare, or public policy (Chintagunta et al. 2006, Mazzeo 2006). Each of these three goals and their respective implications for data are discussed next.

First, one can use economic structure to infer or identify missing primitives such as customer utility on the demand side, as exemplified in Matzkin (1993) and Matzkin (1994), or on the supply side, as in the inference of firms' costs by Nevo (2001). In the former case, economic theory restricts the choice response function to identify the utility function non-parametrically. In the supply-side example, Bertrand–Nash pricing implies that higher costs induce higher prices. Hence, the observed price variation is informative about the unobserved costs. The data ramifications of this structural modeling research objective are apparent. If the primitive is observed, structural assumptions are not necessary to infer them. For example, Chintagunta et al. (2003) observe both wholesale and retail prices. Because retailer markups are purely a function of the demand parameters and retailer costs (wholesale prices), wholesale prices can therefore be used to instrument for retail prices when estimating demand parameters.² Accordingly, the additional wholesale cost information obviates the need to assume a retailer pricing rule. In light of recent findings regarding mixed evidence for Stackelberg pricing (Draganska et al. 2010), avoiding assumptions about the retailer markup rule is a distinct advantage.³ Not only can additional data reduce reliance on assumptions but they can also improve the efficiency of the model's estimates. In this vein, Berry et al. (2004) augment the automobile market share data used in Berry et al. (1995) with additional moments developed from consumer survey data. Specifically, they collect data on consumers' second choices, which are informative about substitution patterns in demand. The additional moments mitigate the need to create supply-side moments as developed in Berry et al. (1995). Similarly, Petrin (2002) collects

data linking consumer demographics to the characteristics of goods these demographics purchase, thereby increasing the information available to infer differences in tastes across groups. As another example, Albuquerque and Bronnenberg (2009) aggregate information across individual-level purchases in the form of the distribution of purchase set sizes and brand penetration to augment the moments in the Berry et al. (1995) demand system. These moments are informative about the dispersion of brand preferences across individuals, and they therefore substantially improve the efficiency of individual-level preference dispersion. Though preferable in some instances to estimate preferences directly from the individual-level data, such information is often expensive to obtain relative to the aggregate statistics used by Albuquerque and Bronnenberg (2009). In totality, these examples suggest the collection of new data can be a contribution in its own right.⁴

Second, because structural models stem from theory, it is possible to discriminate between competing theories by ascertaining which model structure affords a better fit to the data (Bresnahan 1982). For example, Draganska et al. (2010) test various theories about manufacturer–retailer price interactions. Whereas Stackelberg pricing implies one set of pricing rules, Nash bargaining implies another. By assessing which set of pricing rules are more consistent with the data, Draganska et al. (2010) find Nash bargaining outcomes to be more consistent with observed manufacturer–retailer pricing interactions than Stackelberg pricing.⁵ In these cases, structure is used to infer institutional details (decision contexts) that are otherwise not observed. However, even when the goal

² A structural approach might not be necessary if the research goal is to solely infer price response. Fong et al. (2011) directly address pricing endogeneity by conducting a field experiment where price is manipulated exogenously. Interestingly, they find that price response estimates using wholesale prices as instruments as in Chintagunta et al. (2003) are nearly identical to those obtained using experimental methods, thereby enhancing the validity of these instruments. The example further illustrates the link between research goals, approach, and data and how the collection of additional data can be useful to identify the phenomenon of interest or validate model assumptions and instruments.

³ As will be discussed in §2.1, the instrumental variable approach toward inference, although useful for estimating primitives, is difficult to use in counterfactual analysis because there is no theoretical accounting of the data-generating process for the instrumented variable. In this case, the full information approach to estimation is often necessary.

⁴ A similar logic extends to dynamic structural models. Rust (1994) shows that preferences, discounts, and beliefs are not separately identified in dynamic models estimated with panel data. Hence, researchers have typically set discount rates to estimate the remaining two factors. Recently, researchers have used additional data to disentangle these effects. Dubé et al. (2011) use experiments to manipulate beliefs in order to identify discounts and preferences. Yao et al. (2011) exploit a field study wherein consumer decisions are made in a static context in one period (with linear pricing) and then a dynamic context in another (with a three-part tariff). This enables them to identify preferences under the static setting so that they can then estimate discounts and beliefs in the dynamic context. The authors find discount rates lower than commonly assumed in the literature and that this can lead to suboptimal pricing policies in practice. Both examples point to the utility of collecting additional information in dynamic contexts.

⁵ Taking a Bayesian approach, Otter et al. (2011) develop tests to compare the conditional marginal density of demand estimates given alternative supply-side models (including no supply-side restrictions on marketing allocations). Considering a case wherein a service firm allocates branches and promotions across regions, the authors find supply-side restrictions improve the marginal density of the demand. The result implies a pricing strategy consistent with optimal firm behavior.

of a structural model is to test theory, more information regarding institutional details can prove useful. Interviews with managers, assuming they were forthright about their internal pricing, might validate the nature of strategic conduct. As noted by Reiss (2011), tests of competing theories are conditioned on the assumptions that underpin the model being tested, such as linear demand or constant marginal costs. Discussions with managers can help gauge the plausibility of these assumptions. Moreover, insights regarding the institutional details can provide additional insights into model identification. Dubé et al. (2005), for instance, consider advertising dynamics in the context of consumer packaged goods. One implication of the theoretical model is that there should be no small levels of advertising observed. The lack of advertising variation in the lower portion on the demand curve leads to a conundrum; it is difficult to measure a threshold effect for advertising when (i) the implied optimal policy precludes advertising over the lower range of the advertising response function, and (ii) the imputed threshold depends on estimating advertising response in this region of low advertising. However, the authors note the network policy of “make goods” makes identification feasible: when networks fail to meet gross rating points targets for a given campaign, they will credit advertisers with make-good advertising in subsequent periods. This leads to variation in the lower advertising portion of the demand curve.

Third, structural models can be useful in policy simulation, often allowing firms to forecast the effects of their policies beyond the range of the observed data. In this instance, structural models are used to infer outcomes (decisions) that could otherwise not be observed. Even when this is the goal of the research, there exists the alternative to find better data. For example, a field experiment might mitigate the need for policy simulations in new regions of the decision space. That said, field experiments are often costly and limited inasmuch as the outcomes are once again constrained to the manipulation. Hence, the problem of forecasting to other novel states again becomes manifest. One illustration of this consideration is afforded by Duflo et al. (2010), who conduct a field experiment on the effect of incentives on teacher absence in rural India. Although the experiment enables Duflo et al. to measure the effect of a particular set of incentives on absenteeism, it did not enable them to forecast how alternative incentive structures might affect attendance. Hence, they augmented the field experiment with a structural dynamic labor supply model to explore the effect of alternative compensation plans on absenteeism and student outcomes.

As a general concern, the foregoing discussion suggests that structure is often an incomplete substitute for data, meaning that it becomes especially desirable to obtain as many primitives from the data as are feasible and garner as much knowledge about the institutional setting as possible to (i) improve the reliability of the other primitives to be estimated or inferences regarding conduct, (ii) reduce or validate the assumptions necessary to estimate the model, and (iii) afford more leeway in specifying structure on other aspects of interest. Regarding validation, one might argue that the supply-side model developed in Berry et al. (1995) for automobiles does not validate well in the context of consumer packaged goods. Although used often in marketing, the static Bertrand–Nash pricing game implies constant pricing on the part of retailers and manufacturers. However, discussions with store managers and manufacturers as well as casual inspection of pricing series clearly indicate that pricing is not static; rather, discounts are prevalent for most goods. The example illustrates why it is important to validate model predictions against the data and why additional data can also be useful for validation.⁶

In the remainder of this paper, I will outline the data selection and procurement process for structural models. This process, which as noted above should be conditioned on a clear conception of the research motivation for using structural models, consists of six steps. First, one must determine the ideal data to address the research problem. Equipped with the ideal hypothetical data, the second step involves finding the right source for that data. To the degree this involves proprietary data sources, the third step is pitching the research topic to agents who have the data that the researcher needs. Assuming the data source agrees, the fourth step involves negotiating the appropriate disclosure agreement: without permission to publish, years of effort can be wasted. The fifth step involves the data transfer and checking. Last, should the data be complete, the sixth

⁶ Sometimes data used in model estimation can be augmented with supplementary information to validate model assumptions. For instance, Wilbur (2008) infers television show advertisements (network promotions) estimated from a structural model of two-sided networks in the television industry. He then validates the model by taping a sample of television shows to ensure the imputed advertisements are consistent with the estimated advertisements—and he finds the correlation to be about 90%. Duan and Mela (2009) develop a supply-side model of apartment rental pricing to infer the marginal costs of apartment units. Using self-reported marginal costs of \$204 obtained from a survey of apartment managers, the authors find their estimated costs of \$233 to be quite close. Musalem et al. (2010) collect data on out of stocks to show that models ignoring this problem generate downwardly biased estimates of demand (i.e., models incorrectly assume zero sales reflects lack of demand, not availability).

step involves keeping the communication lines open with the firm while the research project proceeds and reporting/implementing the results when the project is done. Although each step is not unique to structural models, its careful emphasis on theory and causality require additional consideration when it comes to specifying data. Structure forces researchers to consider which aspects of the research problem are evident in the data and which require structure or additional information to impute; this trade-off often guides the data choices. In the next section, I address these concerns and other aspects of specifying the appropriate data.

2. Determine the Necessary Data

Determining the data one needs involves two steps: choosing an impactful topic to guide the choice of data and ensuring the data that are chosen can address the issue of interest. The research question and objectives are primal. Without a well-defined research question, data selection and procurement become moot. Moreover, as noted in the previous section, the trade-offs between structure and data further mandate the research problem be well defined before determining the data one needs. It is beyond the scope of this paper to address what constitutes an interesting research question; an excellent resource to address that issue is Varian (1997). Suffice it to say, novel topics typically require novel data. Accordingly, there exists a strong incentive to pursue novel data sources and refrain from limiting one's research by relying on data that are readily accessible, amplifying the data hurdles in structural modeling research.

2.1. Data Components

Data contain four components: (i) a dependent variable of interest one seeks to explain (the interesting research problem), (ii) covariates that drive the effect of interest (treatment effects, or the marketing decisions available to the firm), (iii) a potential set of instruments (or natural experiment) that identifies the effects of the covariates of interest, and (iv) a set of institutional factors that underpin the data-generating process. Regarding points (i)–(iii), it is ideal that the treatment effects are randomized to the outcomes. When this is the case, the covariates and instruments are the same. However, in many instances, this is not the case, and the estimation strategy should reflect this lack of randomization. Under an instrumental variable estimation strategy, this involves finding appropriate instruments for the endogenous regressors. In an illustrative example, Hofstetter et al. (2010) consider the effect of content generation on social ties on a windsurfing website. Website users who create content might receive more friend invitations. To the extent firms can induce their

users to post (especially if friending also induces posting), the interaction can enhance website engagement. Yet unobservable factors can influence both outcomes, making it hard to assess the causal relationship between the two; for example, friending and posting might both increase with Internet penetration. Accordingly, Hofstetter et al. (2010) collect exogenous instruments that vary with posting but not with friending. In particular, they consider wind speed, reasoning that an increase in wind will lead to more posting but that its effect on friending is negligible because most friends are established in an off-line environment. The instrument is ideal in many regards, because it is hard to argue wind is endogenous (unless one has Aeolus as a friend).

Although instruments can be helpful to reduce the modeling assumptions regarding the distribution of unobservables as noted above, they can be an inadequate substitute for structure in many contexts; instrumenting for endogenous variables is of limited utility in conducting counterfactual analyses because this approach does not detail the underlying economic process that generates the instrumented data. If the goal of the model is to engage a counterfactual analysis, a full information estimation approach to account for the endogenous variable might be preferred. Consider Tucker (2008), who explores network effects in video technology adoption at an investment bank. In this context, a joint adoption by those who communicate might reflect a network externality or might reflect omitted factors such as a firm's policy to increase adoption. Given that the adoption decision could not be randomized to disentangle these effects, Tucker (2008) instead relies on quasi-random variation across regions that differ across adopting pairs to exploit variation in adoption propensities. Although these instruments are helpful in isolating and measuring network externalities, they are not as informative about the underlying processes that generate adoption in the first place. Thus, the instrumental variable approach offers little insight regarding how to induce changes in adoption strategy. Hence, Ryan and Tucker (2011) develop a structural model of forward-looking behavior, where beliefs about the current and future technology adoptions of others affect the likelihood of a given user adopting in the current period. This model is then used to explore how to seed lead users to accelerate technology adoption. The structural model also enables them to allow for heterogeneity in adoption effects; whereas a reduced-form approach would need 64 instruments (one for each type of caller in their data) to model each type's adoption, the structural model relies on ex post adoption calling behavior of the users to infer the utility of various types. Of course, as noted in §1, the full information approach, by incorporating all the theoretical

restrictions on the endogenous variables, also requires more assumptions. To the extent that these assumptions are not accurate, inference about agent behaviors can be biased. Hence, one trade-off between the use of full information and instrumental variable estimation approaches is that the former invokes assumptions but is also more useful for counterfactual analysis.

Regarding (iv), organizations involved in the game are an integral source of contextual data. Counterfactual analyses regarding strategy, for example, are more persuasive when one can document that the agent is not already behaving optimally. Working with a firm to obtain data is sometimes necessary to the appropriate specification of these models. Whenever possible, it is sensible to contact agents in the industry prior to modeling it. Although this insight is not estimation data, *per se*, the institutional data are no less critical for structural model estimation. Consider Gordon and Hartmann (2010), who explore the role of political advertising. Prevailing advertising rates are readily available from Nielsen's Campaign Media Analysis Group (CMAG). However, instead of taking these rates as given, the authors interviewed CMAG four times, eventually clarifying rates with its president. They learned advertisers are obligated to charge the lowest advertising spot rates for campaigns, and these rates often deviate from published market rates. Moreover, they learned that market rates are more prevalent for large campaigns. The key point is that, without consulting the agents in the game, they could have used the wrong rates and made the wrong inferences regarding advertising campaigns in political campaigns. Given that structural models are hard to falsify, as the outcomes follow from the assumptions, there is an especially strong onus to be placed on researchers to ensure that they have a complete and correct perspective of the considered context and problem and that their assumptions are valid.⁷

2.2. Data Types

With a research question of interest in hand, and a sense of which exogenous variables might identify the effects of interest, several data sources exist to address the research question, including (i) firm proprietary data, (ii) free public data, and (iii) commercially available market research data.

⁷ An anonymous reviewer presents another example concerning price competition between two national U.S. retailers. Interviews with industry members revealed that retailers evidenced little regional variation in pricing in their industry and that pricing decisions were made at a corporate level. Accordingly, a regional pricing model would be misspecified, leading to poor inference and invalid counterfactual analyses. This further indicates the importance of institutional data.

Proprietary data are typically sourced directly from firms. Because these data are developed in conjunction with a firm to address a research problem, they are free, flexible, and often tailored directly to the research problem. Moreover, because firms are engaged with customers, proprietary data are often emblematic of important research problems. Furthermore, by working directly with an agent involved in decision making, one can often obtain the necessary insights into the rules of the game, information states, and agent objective functions needed to properly specify the structural model. The uniqueness of the data also increases the potential to make a unique contribution. However, the time to procure such data is often lengthy because of the resources a firm needs to commit to obtain them. Other limitations inherent in proprietary data are worth noting. The data are subject to the predilections and vagaries of the sponsoring firm, and hence the data are risky to collect. Moreover, because these data are often difficult to share, the potential impact of the work can be limited if others have difficulty building on the research or generalizing the findings. Finally, the institutional details of the data can be so pathological as to limit the utility of the research beyond a narrow context. An example of proprietary data is Yao and Mela (2008), who worked with an auction house to obtain data on the bidding history of bidders and the listing history of sellers. The research explored how auction house pricing affected revenue. By obtaining data directly from the firm, they were able to observe all bidders' bids across all auctions; in contrast, auction papers often collect data by parsing the website for a single auction, meaning that only the leading bids are observed. Hence, Yao and Mela (2008) could obtain a more complete sense of the distribution of customer valuations and control for unobserved heterogeneity from the repeated observations. Working with the firm, therefore, enabled the authors to more clearly differentiate their research relative to the existing work on auctions.

Public data, such as websites, are free and can often be collected more expediently than proprietary data, yet these data are harder to customize to the research application. Illustrative of public domain data research, Forman et al. (2009) explore competition between local and electronic markets by collecting public records pertaining to local bookstore openings and scraping Internet websites to determine Amazon.com book sales in that region. Another example is afforded by Kim et al. (2010), who explore how consumers search across alternatives. Ideally, this exploration would involve collecting details regarding the specific set of goods searched by a consumer. However, these data are often not available from Internet retailers owing to privacy concerns.

Thus Kim et al. (2010) parsed the Amazon.com website to scrape the set of alternative products viewed by the collection of visitors to a particular product page as well as the rank ordering of the viewership overlap. Together with the site's rules for determining these ordering, Kim et al. (2010) were able to obtain estimates of the precise number of joint visits to competing products' pages. This enabled the authors to form inferences about search. Facilitating the task of scraping Internet data, tools such as Mozenda (<http://www.mozenda.com>) have become commercially available. Government websites are another major source of publicly available data; Du and Kamakura (2008) develop a model of consumers seeking to maximize their utility over their life cycle by selecting various goods ranging from education to housing. Collecting such a long stream of data would be a formidable task, but Du and Kamakura were able to exploit the National Bureau of Economic Research Consumer Expenditure Survey, which spans 22 years. Archival research of public data can often be time consuming to the extent that it involves finding and integrating data sources. Dataferret (<http://dataferret.census.gov>) is one tool for navigating and extracting U.S. Census data, and similar products exist for other public and international data; Swedish consumer prices and household expenditures across a wide range of categories are available at <http://www.scb.se>. Of course, not all records are available from the Internet: in these instances, data collection can become especially challenging. Bronnenberg et al. (2009) explore order of entry effects in spatial market shares to obtain product entry timing, the authors had to visit the National Museum of American History in Washington, DC.

Finally, there exist a litany of commercial market research firms that specialize in collecting data and selling this information to marketers and researchers. These include Impact RX in pharmaceuticals, Nielsen and IRI for Internet and shopping panels, J.D. Power and Associates for automotive, NPD for point-of-sale data on technical goods, and so forth. These commercial data are immediate to obtain and often expensive, though grants often defray the cost. Sometimes, the university library will purchase the data so that they will be available to all researchers; for example, Duke subscribes to Taylor Nelson Sofres's Ad Spender data, and both economists and marketers have used the data. Another advantage of commercial market research data is that they are quick to obtain and are formatted in a fashion that makes them easy to use. A key limitation is that they are limited in the sense that there is no flexibility in the information they collect. Illustrative of this style of research, Dong et al. (2009) explore the effect of a firm's strategic targeting of detailing on estimates of physician response

to detailing. Using data from Impact RX, Dong et al. (2009) show that failure to account for endogeneity arising from the supply side leads to an underestimation of the benefits of targeting. Dubé et al. (2010) obtained data from NPD Techworld regarding the price and sales of video-game consoles as well as game availability. This enabled them to infer the role of network effects in how a market tips to a standard. Goldfarb and Xiao (2011) integrate two types of data sources: commercial research data and public data. To explore the role of managerial ability to iterate to equilibrium in an entry game, Goldfarb and Xiao purchased firm entry data from a reseller, used public government data to ascertain local entry conditions, and conducted an Internet search to ascertain publically available managerial experience of the considered firms.⁸

In the remainder of this paper, I will focus on a systematic approach to data acquisition for proprietary data. Acquisition of public data (from commercial market research firms and public sources) is more formulaic though it can sometimes be time consuming. Public and commercial data involve standard protocols for their use that can be easily followed. Because proprietary data involve the consent of a counterparty, there are additional considerations in obtaining data that I address next.

3. Find the Right Contact

Partnering with a firm often involves working with a visionary and prominent agent within the firm. Individuals with the willingness to take risks on academic projects, allocate time to the endeavor, and have the initiative to sell the potential to the organization are rare. Several domains exist in which to find such research partners; these include corporate engagements, personal interactions, and collegial (academic) connections.

Corporate contacts involve interactions made in the process of professional engagements. For example, consulting relationships can lead to access to data. Although rarely will firms allow consulting data from a current project to be used for research, they will

⁸ Although the public government census data and commercial entry data were relatively easy to source, public data on managerial ability were collected only after a laborious search of news sources to obtain information on managerial experience and characteristics. Even then, reviewers asked for additional information on managerial characteristics, much of which was not available at the time of this paper's inception. Fortunately, innovations like LinkedIn and online white pages appeared that enabled Goldfarb and Xiao (2011) to obtain almost all of the additional information requested (though involving another lengthy search). All totaled, the data collection process required more than six months of calendar time. A key point from this case is that a great deal of creativity and persistence is often needed to find appropriate data.

often allow academics to use the data after a period of a few years or guide researchers to other data in that firm. Similarly, advisory board positions can be used to obtain data. One of my key reasons for participating on the analytic advisory board of IRI was to develop an academic data set (see Bronnenberg et al. 2008). Consulting and advisory board opportunities are a function of the relevance of the work, suggesting the potential for a feedback loop between data and research: more relevant data lead to more relevant research which, in turn, yields more data. The key theme is to work on cultivating corporate contacts through active research. Former work colleagues are often instrumental in supporting research. The data for Yao and Mela (2008) arose through my interactions with a former colleague.

In addition to working on applied problems, promoting relevant research plays a role in facilitating corporate interactions. Conferences such as the Advertising Research Foundation (ARF) annual conference, the American Marketing Association Advanced Research Techniques (ART) Forum, the Institute of International Research conferences, the Direct Marketing Education Foundation annual meeting, and the Word of Mouth Marketing Association annual conference all represent singular opportunities to be involved. Kempe et al. (2011) use cable TV set-top box viewership data from a major cable company that were obtained after the second author presented related research at an ARF conference. Another colleague of mine, Wagner Kamakura, presented a tutorial at the ART Forum, which has since opened up a number of data opportunities. Organizing research presentations at and with local companies is another approach that is quick and can yield high dividends. Another means for increasing the managerial exposure of one's research is to publish a summary of the work in a managerial outlet such as the *Harvard Business Review*. Writing books or creating software for general distribution also opens doors. The key point is that one's research must be highly relevant and must be strongly promoted to facilitate managerial access for new data. When all else fails, a more prominent professional profile can also enhance the likelihood of success when cold-calling for data.

It perhaps goes without saying that family or college contacts can be instrumental in research. In my case, one example of a paper that materialized through a friend of a family connection is Yao and Mela (2011), who develop a dynamic structural model of advertiser bidding and consumer search in the context of keyword search.

The university environment also facilitates research connections on a number of levels. One can invite speakers to class that serve to both enhance the relevance of the class and also open discussions

for new data. An illustration of this approach is Bronnenberg et al. (2010), who contrast the purchase behavior of households who received a DVR to those who did not and find that DVRs have no effect on the sales of advertised goods. The idea for the research was conceived in conjunction with a visit from IRI and TiVo to my MBA class, and the data followed from additional discussions with the speakers. Second, alumni and students can also source data and provide research insights. Chintagunta et al. (2010) assess the effect of user reviews on sales; the data for this project came from a student of the third author. Along these lines, it is helpful to read the résumés of students and learn to use the alumni database systems; many will have deep connections into firms that capture data to address research problems of interest. Third, the university itself sometimes has data. Grubb and Osborne (2010) model plan choice and usage in the cell phone industry wherein consumers learn about their usage needs with time. A distinguishing feature of their data that enable them to identify learning effects is the call detail records at the user level from the adoption of a plan. These were obtained from a university that proffered these plans. Fourth, colleagues or advisors can also be a rich source of data. Hartmann and Nair (2010), in a dynamic model of demand, consider the problem of tied goods. The model is calibrated using razors and blades; they augmented their original data with advance data from the IRI data set provided by Bronnenberg et al. (2008). My colleagues and I had to obtain special permission to provide it, but many colleagues will make a concerted effort to help peers in this regard. Fifth, more organized exemplars of this cooperative collegial data effort include the Marketing Science Institute (MSI) (<http://www.msi.org>), the Wharton Customer Analytics Initiative (<http://wharton.upenn.edu/wcai>), and the University of Chicago Booth School's Kilts Center for Marketing (<http://research.chicagobooth.edu/marketing/index.aspx>). MSI has a roster of blue-chip marketing firms and will make introductions to facilitate data acquisition and research grants. WIMI has been especially active in generating interesting data; recently, they have partnered with Expedia to offer data from the hospitality industry. Iyengar et al. (2007) consider a structural model of consumer learning about service quality: the research stems from data that were obtained from the Teradata Center for Customer Relationship Management at Duke. Sixth, several journals now have databases available online. A good example is the online database page on the *Marketing Science* website (<http://www.informs.org/pubs/MktSci/Online-Databases>).

Finally, it is worth considering the level of the organization with which to work. Ideally, it is sufficiently

senior that there is budgetary discretion to authorize the research and obtain approval for the data but not so senior that the project is of modest relevance.

4. Make the Pitch

Once one has established contact with a firm whose data are aligned with the research proposition, the next step is to pitch the project. Two points are critical. First, the pitch should address what the firm is to gain from the work. Second, it should address the costs.

Regarding the former, one needs to address specifically how the research will affect which decisions made by whom and the resulting impact on profit or revenues. One should also indicate why the firm or contact needs him or her. Another way of thinking about these questions is whether the work will help the contact in his or her annual review, and then one should frame one's pitch from the perspective of why it would help the contact and the firm. The firm cares only about whether one's work is an asset to it; else, it has no incentive to provide data. If the research considers pricing, for example, one must consider how prices are currently set, how and why they should be set differently, and the implications for the firm's revenues. Yao and Mela (2008) consider pricing in online auctions. One approach to pricing might be to regress demand against price and use this demand curve to establish the optimal price. Aside from endogeneity concerns, however, this approach is not feasible in online auctions because price changes are infrequent, and the growth rates are so substantial that older data points are not so relevant. The authors instead imputed sellers' reservation values for listing based on past listing behavior and used these reservation values to generate a demand curve. Thus, the research enabled the firm to consider how to price in the absence of a regression-based approach and was of key interest to the firm in setting pricing policy. The bottom line, of course, is that the firm cares little about publications, sophisticated modeling, or tenure aspirations. It wants to know the return on its investment from sponsoring one's work, so this had better be stated clearly in the pitch.

Costs are integral to firms' willingness to share data. It is sensible to determine the most "raw" form of the data that can be pulled from a firm's servers with the minimal level of effort on the part of the firm and then do the data cleaning and aggregating by one's self. Likewise, to the extent one can develop his or her code and analysis in a way that a firm can easily port that analysis to other years or products, it is more likely to embrace the research.

As a pragmatic matter, it is often helpful to provide relevant exemplars of papers and historic research

findings to make the benefits and costs of the research more concrete. This approach makes the project more tangible, especially when the researcher makes it clear that the end product of the partnership will yield a similar output.

5. Negotiate the Nondisclosure Agreement

Assuming one is sufficiently fortunate that the firm has promised to send the ideal data set one's way, the firm will often ask that individual to sign a nondisclosure agreement (NDA). The purpose of this agreement is to protect the firm from potentially adverse outcomes. This might occur, for example, if one were to forecast the firm's growth is to slow, and the analysts reading this information begin to issue sell recommendations against the firm. Firms can also be concerned about the release of private information that may advantage a competitor or create concerns with regulators (such as their cost data). An example of the problems firms face when sharing data is afforded by Netflix, who sought to leverage the academic community by providing disguised account-level rental data to researchers; the goal was to determine which movies should be recommended to whom. The research resulting from the \$1 million prize improved the Netflix's recommendation and forecasting systems and generated considerable positive press. However, Narayanan and Shmatikov (2008) integrate the Netflix data with the Internet Movie Database (<http://www.imdb.com>), and the authors imputed the identities of the disguised panelists as well as their rental histories. Netflix was subsequently contacted by the Federal Trade Commission and faced a class action lawsuit; Netflix also canceled its next prize competition.

The firm has little interest in seeing one's work published given its primary focus on revenue generation, and it fears that competitors might benefit as much—if not more—from that work. Hence, it will seek to protect all data, often asking an individual to sign the same NDA as its consultants. Were one to do this and still publish his or her work, he or she can be sued for breach of contract and damages. To add insult to injury, the agreements will often ask that individual to pay for the legal costs. It is preferable to abort the research project than sign an NDA that precludes publication or mandates sponsor review prior to publication. Two years into an individual's research is too late to discover that he or she is not allowed to publish his or her work; that time cannot be recovered. In light of the time often needed to estimate a structural model, the time cost of this discovery to the researcher can be exceptionally high when the firm objects to publication. Another standard practice is for firms to

indicate that any work one does with them becomes their property. This means that an individual cannot ever use the tools he or she develops in the future with other organizations.

Ideally, one would like no restrictions whatsoever on what he or she publishes. In practice, this is rarely amenable, and an individual can agree to disguise the firm's name and/or aggregate the data (meaning one can publish aggregate statistics or parameter value but not release individual-level observations). Subject to these caveats, most firms will be amenable to signing an agreement.

If the firm amends the NDA to one's liking, it is prudent to have the university lawyers review the document before it is signed. They are often experienced in this domain. An e-mail to one's faculty asking for prototypes of NDAs is also a sensible exercise because others have worked through this issue. Note that an NDA protects not only the firm but one's right to publish. Hence, it is inadvisable to proceed without one. One should also check with one's human subjects committee: though not experimental data, if person-specific data that are not commercially available are obtained, it is sometimes necessary to obtain university approval for these data.

6. Data Delivery and Data Checking

Although a signed NDA is a key hurdle, additional pitfalls await in the data handoff. Generally, there is a short window over which data are collected and transferred. A bad time to learn about incomplete data is two years into the research, when reviewers ask for a robustness check and one finds a relevant field has not been populated. It is not uncommon at that stage for the research contact to have moved to another position or for the firm's information system to render the potential for supplementing the data impossible. This problem is exacerbated in the context of structural models because owing to their inherent complexity, the time needed to specify and estimate them—and to navigate the review process—can often be exceptionally long.

Data are often massive, so it is wise to start with a smaller set of metadata (for example, first receive only a few households of data over a short period). This can be used to check basic statistics, the degree of missing observations, and odd values in the data. Simple regressions and correlations can be used to consider the relationships between key variables of interest to ensure that there is sufficient information to explore these. Check to see whether the variables in the data let one address the considered questions, as noted in §2.1. Outliers observed at this stage can be informative; when outliers exist, they should be explained. In grocery data, for example, I have

observed unusually high weekly sales as a result of institutional, nonconsumer purchases. Such purchases should not be included in a model of consumer demand.

It is also important to understand the data structure completely at this phase: an exhaustive data dictionary should list, for each file, the variable definition the unit of observation and, ideally, the variables that intersect data files and enable one to join them. The Teradata cell phone data set (Iyengar et al. 2007) included disparate files for call detail records, stores, handsets, demographics, promotional targeting, and billing statements. It is imperative to note, for example, that there is a handset indicator in the call detail records that enables one to splice the calls made with the handsets owned. There are additional lookup tables that link numeric codes for promotions to the details of the promotions. These relational data structures can be enormously complex and should be completely mapped out. Likewise, obtain definitions for each variable. Would you know, for example, what that variable "user_behavior_id_skc_url" means?

Assuming the data are in great shape and one is ready to consider the data transfer, the next step is to determine how much data can be handled. Yao and Mela (2008) focus on coin auctions because there is relatively little cross-bidding behavior with other categories, especially for antique coins. Although it was conceivable to obtain information beyond coins, the total data would have been difficult to store and process. Hence, it was not feasible to obtain information regarding all auctioned goods. Once the scope of the data are defined, data transfer becomes increasingly easy as the price of external hard drives falls. A company can easily record a terabyte of data onto a hard drive and ship it, assuming one has the processing power to analyze it. Sometimes it can be sensible to visit the company and meet with the IT department to help expedite the process.

7. Project Management

Academics move at a glacial pace. In contrast, businesses are driven by quarterly returns. This mismatch in expectations creates problems if mismanaged. There exist four project phases to consider: project initialization, the analysis phase, project completion, and project implementation.

In the earliest phase, it is imperative to recheck the full data to ensure that they match the metadata. Often, data are collected going forward; this was the case with Yao and Mela (2011), where the sponsoring firm collected a couple of months of log file data around the project's data requirements. This is an active phase of collaboration where the data suggest more questions about the institutional context of

the project and possibly the need for more data. In the middle phase, model estimation, validation, and paper writing are ongoing. In this phase, checking in with one's sponsoring firm every couple of months is especially important to ensure the firm's continued interest.

Once the project is complete, an on-site or phone conference is required to (a) ensure insights from the project are realized and (b) present research to the firm to ensure the project is not missing any key institutional details. In conjunction with one's primary contact, it is desirable to invite as many persons as possible from the firm; this will maximize feedback and possibly open the door for new projects.

The presentation should remain conceptual, with few equations (unless it is to a modeling group in the firm). I recall one example where I computed derivatives on a profit function to recommend an optimal price: the firm to whom I presented the analysis thought I was out of touch. On the next occasion, I simply plotted the profit surface and showed its maximum, and the response was far more positive. Obviously, structural models are far more complex, meaning that greater effort is needed in explaining their benefits in a coherent and concise fashion, such as why they are so well suited for counterfactual analyses. In addition to overviewing the project in readily accessible terms, one must be clear about what the firm should do differently as a result of his or her research. This affords an opportunity to learn more about implementation concerns.

Finally, should this meeting go well, the occasion sometimes presents itself to implement the research. A superlative example of this potential is presented in Misra and Nair (2011), who study salesforce compensation plans. Using an agency-theoretic model to underpin a dynamic structural model, Misra and Nair (2011) are able to infer salesperson effort functions. Conditional on these functions, they explore counterfactual compensation schemes to improve firm profits. Interactions with the firm quickly ruled out some that were infeasible, but the authors persuaded the firm to try an alternative approach. Early results suggest their research resulted in a \$1,000,000 per month (9%) improvement to the firm's bottom line. Their paper exhibits the potential for structural models and data to interact in a fashion that can generate impressive improvements in business practice.

A key caveat in the implementation of structural models sometimes involves their scalability, especially in the context of dynamic models. The large number of observations in practice, the large array of state and control variables, and the frequency of decision making can render the application of structural models infeasible. For example, the pricing of

perishable goods is in real time. Should a structural model lead to an optimal policy function that requires numerical computation, it will be impossible to implement because states (e.g., ticket availability) will change by the time the optimal decision is computed.⁹ Whereas Misra and Nair (2011) are careful to specify their dynamic model in a way that is useful to the firm, other examples abound where models do not scale, limiting the counterfactual potential of the structural model. In that case, researchers should proactively consider collecting additional data, computational innovations, or scaling down the research problem to enhance the utility of the research. An example of a recent innovation to ease the effect of dimensionality of the state space is the application of Bayesian methods to the estimation of dynamic discrete models (Imai et al. 2009).

8. Conclusion

Structural models, like all empirical research, are predicated on finding the right data. In many regards, the hurdles for this task are more challenging in the context of structural models because of their emphasis on causal attribution and because data and structure are sometimes used interchangeably. Hence, the objectives of the structural research interact with the data. When primitives are missing in the data, structure is useful to infer them. When institutional and environmental details are missing, structure can be used to test them. When outcomes of interest extend beyond the range of the data, structure can enable one to predict them. In each case, richer data enable more reliable inferences and deeper insights.

In this paper, therefore, I detailed a process for data selection and procurement in the context of structural models by exploring this interaction between data and structure and by proposing an approach for obtaining data. Although publicly available data and data from market research firms are often sufficient, the paths to obtain these data are more incumbent on researcher initiative than that of a private market partner. Hence, I outlined a series of steps in obtaining proprietary data, including making a contact with the firm, pitching the project, negotiating an NDA, transferring the data, managing the project, and reporting the results. It is my hope that these tips will be useful to students and others seeking to estimate structural models. Finally, it is also my hope that, to the extent these endeavors are successful, researchers will take steps to share these data in a collective effort to advance the field.

⁹ Although the scalability concerns might be especially substantial with structural models (especially in dynamic contexts), they can also apply to simulation-based methods in marketing such as Bayesian models (Rossi and Allenby 2003).

Acknowledgments

The author thanks the editor, area editor, reviewers, Dae-Yong Ahn, Peter Fader, Avi Goldfarb, Mike Hanssens, John Hauser, Wagner Kamakura, Sanjog Misra, Andres Musalem, Jason Roos, Peter Rossi, Gerard Tellis, Catherine Tucker, and Ken Wilbur for helping to shape the ideas in this paper. The author also thanks the conference organizers, Brett Gordon, Wes Hartmann, Rick Staelin, and Raphael Thomadsen, for their feedback, and the attendees of the 2010 Structural Modeling Workshop at Duke University.

References

- Albuquerque, P., B. J. Bronnenberg. 2009. Estimating demand heterogeneity using aggregated data: An application to the frozen pizza category. *Marketing Sci.* **28**(2) 356–372.
- Berry, S., J. Levinsohn, A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* **63**(4) 841–890.
- Berry, S., J. Levinsohn, A. Pakes. 2004. Differentiated products demand systems from a combination of micro and macro data: The new car market. *J. Political Econom.* **112**(1) 68–105.
- Bresnahan, T. F. 1982. The oligopoly solution concept is identified. *Econom. Lett* **10**(1–2) 87–92.
- Bronnenberg, B. J., S. K. Dhar, J.-P. H. Dubé. 2009. Brand history, geography, and the persistence of brand shares. *J. Political Econom.* **117**(1) 87–115.
- Bronnenberg, B. J., M. W. Kruger, C. F. Mela. 2008. Database paper—The IRI marketing data set. *Marketing Sci.* **27**(4) 745–748.
- Bronnenberg, B. J., J.-P. Dubé, C. F. Mela. 2010. Do digital video recorders influence sales? *J. Marketing Res.* **47**(6) 998–1010.
- Chintagunta, P. K., J.-P. Dubé, V. Singh. 2003. Balancing profitability and customer welfare in a supermarket chain. *Quant. Marketing Econom.* **1**(1) 111–147.
- Chintagunta, P. K., S. Gopinath, S. Venkataraman. 2010. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* **29**(5) 944–957.
- Chintagunta, P., T. Erdem, P. E. Rossi, M. Wedel. 2006. Structural modeling in marketing: Review and assessment. *Marketing Sci.* **25**(6) 604–616.
- Dong, X., P. Manchanda, P. K. Chintagunta. 2009. Quantifying the benefits of individual-level targeting in the presence of firm strategic behavior. *J. Marketing Res.* **46**(2) 207–221.
- Draganska, M., D. Klapper, S. B. Villas-Boas. 2010. A larger slice or a larger pie? An empirical investigation of bargaining power in the distribution channel. *Marketing Sci.* **29**(1) 57–74.
- Du, R. Y., W. A. Kamakura. 2008. Where did all that money go? Understanding how consumers allocate their consumption budget. *J. Marketing* **72**(6) 109–131.
- Duan, J., C. F. Mela. 2009. The role of spatial demand on outlet location and pricing. *J. Marketing Res.* **46**(2) 260–278.
- Dubé, J.-P. H., G. J. Hitsch, P. K. Chintagunta. 2010. Tipping and concentration in markets with indirect network effects. *Marketing Sci.* **29**(2) 216–249.
- Dubé, J.-P., G. J. Hitsch, P. Jindal. 2011. Estimating durable goods adoption decisions from stated preference data. Working paper, University of Chicago, Chicago.
- Dubé, J.-P., G. J. Hitsch, P. Manchanda. 2005. An empirical model of advertising dynamics. *Quant. Marketing Econom.* **3**(2) 107–144.
- Duflo, E., H. Rema, S. P. Ryan. 2010. Incentives work: Getting teachers to come to school. Working paper, MIT, Cambridge, MA.
- Fong, N. M., D. I. Simester, E. T. Anderson. 2011. Private label vs. national brand price sensitivity: Evaluating non-experimental identification strategies. Working paper, MIT, Cambridge, MA.
- Forman, C., A. Ghose, A. Goldfarb. 2009. Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management Sci.* **55**(1) 47–57.
- Goldfarb, A., M. Xiao. 2011. Who thinks about the competition? Managerial ability and strategic entry in U.S. local telephone markets. *Amer. Econom. Rev.* Forthcoming.
- Gordon, B., W. R. Hartmann. 2010. Structural analysis of political advertising. Working paper, Stanford University, Stanford, CA.
- Grubb, M. D., M. Osborne. 2010. Cellular service demand. Working paper, Stanford University, Stanford, CA.
- Hartmann, W. R., H. S. Nair. 2010. Retail competition and the dynamics of demand for tied goods. *Marketing Sci.* **29**(2) 366–386.
- Hofstetter, R., S. K. Shriver, H. Nair. 2010. Social ties and user generated content: Evidence from an online social network. NET Institute Working Paper 09-28, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1520110.
- Imai, S., N. Jain, A. Ching. 2009. Bayesian estimation of dynamic discrete choice models. *Econometrica* **77**(6) 1865–1899.
- Iyengar, R., A. Ansari, S. Gupta. 2007. A model of consumer learning for service quality and usage. *J. Marketing Res.* **44**(4) 529–544.
- Kempe, D., K. C. Wilbur, L. Xu. 2011. How television networks can sell audiences instead of time. Working paper, Duke University, Durham, NC.
- Kim, J. B., P. Albuquerque, B. J. Bronnenberg. 2010. Online demand under limited consumer search. *Marketing Sci.* **29**(6) 1001–1023.
- Matzkin, R. L. 1993. Nonparametric identification and estimation of polychotomous choice models. *J. Econometrics* **58**(1–2) 137–168.
- Matzkin, R. L. 1994. Restrictions of economic theory in nonparametric methods. R. F. Engle, D. McFadden, eds. *Handbook of Econometrics*, Vol. 4. Elsevier, Amsterdam, 2523–2558.
- Mazzeo, M. J. 2006. Marketing structural models: “Keep it real.” *Marketing Sci.* **25**(6) 617–619.
- Misra, S., H. S. Nair. 2011. A structural model of sales-force compensation dynamics: Estimation and field implementation. *Quant. Marketing Econom.* **9**(3) 211–257.
- Musalem, A., M. Olivares, E. T. Bradlow, C. Terwiesch, D. Corsten. 2010. Structural estimation of the effect of out-of-stocks. *Management Sci.* **56**(7) 1180–1197.
- Narayanan, A., V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets. *IEEE 2008 Sympos. Security Privacy*, IEEE Computer Society, Washington, DC, 111–125.
- Nevo, A. 2001. Measuring market power in the ready-to-eat cereal market. *Econometrica* **69**(2) 307–342.
- Otter, T., T. J. Gilbride, G. M. Allenby. 2011. Testing models of strategic behavior characterized by conditional likelihoods. *Marketing Sci.* **30**(4) 686–701.
- Petrin, A. 2002. Quantifying the benefits of new products: The case of the minivan. *J. Political Econom.* **110**(4) 705–729.
- Reiss, P. C. 2011. Descriptive, structural and experimental methods in marketing research. Working paper, Stanford University, Stanford, CA.
- Reiss, P. C., F. A. Wolak. 2007. Structural econometric modeling: Rationales and examples from industrial organization. J. J. Heckman, E. E. Leamer, eds. *Handbook of Econometrics*, Vol. 6A. Elsevier, Amsterdam, 4277–4415.
- Rossi, P. E., G. M. Allenby. 2003. Bayesian statistics and marketing. *Marketing Sci.* **22**(3) 304–328.
- Rust, J. 1994. *Structural Estimation of Markov Decision Processes*. Elsevier Science, Amsterdam.
- Ryan, S., C. Tucker. 2011. Heterogeneity and the dynamics of technology adoption. *Quant. Marketing Econom.*, ePub ahead of print August 11.
- Thomadsen, R. 2005. The effect of ownership structure on prices in geographically differentiated industries. *RAND J. Econom.* **36**(4) 908–929.
- Thomadsen, R. 2007. Product positioning and competition: The role of location in the fast food industry. *Marketing Sci.* **26**(6) 792–804.
- Tucker, C. 2008. Identifying formal and informal influence in technology adoption with network externalities. *Management Sci.* **54**(12) 2024–2038.

- Varian, H. 2009. Hal Varian on how the Web challenges managers. *McKinsey Quart.* (January), http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286.
- Varian, H. R. 1997. How to build an economic model in your spare time. *Amer. Economist* 41(2) 3–11.
- Wilbur, K. C. 2008. A two-sided, empirical model of television advertising and viewing markets. *Marketing Sci.* 27(3) 356–378.
- Yao, S., C. F. Mela. 2008. Online auction demand. *Marketing Sci.* 27(5) 861–885.
- Yao, S., C. F. Mela. 2011. A dynamic model of sponsored search advertising. *Marketing Sci.* 30(3) 447–468.
- Yao, S., C. F. Mela, J. Chiang, Y. Chen. 2011. Determining consumers' discount rates with field studies. Working paper, Northwestern University, Evanston, IL.