CHAPTER

Inference for marketing decisions

2

Greg M. Allenby[a], **Peter E. Rossi**[b,*]

[a]*Fisher School of Business, Ohio State University, Columbus, OH, United States*
[b]*Anderson School of Management, University of California Los Angeles, Los Angeles, CA, United States*
[*]*Corresponding author: e-mail address:* perossichi@gmail.com

Contents

Handbook of the Economics of Marketing, Volume 1, ISSN 2452-2619, https://doi.org/10.1016/bs.hem.2019.04.007

1 Introduction

Much has been written on the virtues of various inference frameworks or paradigms. The ultimate judgment regarding the usefulness of a given inference framework is dependent upon the nature of the inference challenges presented by a field of application. In this chapter, we discuss important challenges for inference presented by both the nature of the problems dealt with in marketing as well as the nature of the data available. Given that much of the current work in quantitative marketing is influenced by economics, we will also contrast the prevalent view in economics regarding inference with what we have found useful in marketing applications.

One important goal of quantitative marketing is to devise marketing policies which will help firms to optimize their choice of marketing actions. For example, a firm might seek to improve profitability by better measurement of demand function for its products. Ultimately, we would like to maximize profitability over the space of policy functions which determine the levels and combinations of marketing actions. This goal imposes a high bar for inference and modeling, requiring a measurement of the entire surface which relates marketing actions to sales consequences, not just a derivative at a point or an average derivative. In addition to these problems in response surface estimation, some marketing decisions take on a discrete nature such as which person or sub-group to target for advertising exposure or which ad creative is best. For these actions, the problem is how to evaluate a large number of discrete combinations of marketing actions.

To help solve the demanding problem of optimizing firm actions, researchers in marketing have access to an unprecedented amount of highly detailed data. Increasingly it is possible to observe highly disaggregate data on an increasing number of consumer attributes, purchase history, search behavior, and interaction with firms. Aggregation occurs over consumers, time, and products. At its most granular level, marketing data involves observing individual consumers through the process of consideration and purchase of specific products. In any given time period, most consumers purchase only a tiny fraction of the products available to them. Thus, the most common observation in purchase data is a "0." In addition, products are only available in discrete quantities with the most commonly observed quantity of "1." This puts a premium on models of demand which generate corner solutions as well as econometric models of discrete or limited dependent variables (see Chapter 1 for discussion of demand models which admit corner solutions). Consumer panel data features not only a very large number of variables which characterize consumer history but also very large number of consumers observed over a relatively short period of time.

In the past, marketing researchers used only highly aggregate data where the aggregation is made over consumers and products. Typically, information about the consideration of products or product search was not available. Today, in the digital area at least, we observe search behavior from browsing history. This allows for the possibility that we can infer directly regarding consumer preferences before the point of purchase. In the past, only demographic or geo-demographic[1] consumer characteristics were observed. Now we observe self-generated and other social media content which can help the researcher infer preferences. We also observe the social network of many if not most potential customers, opening up new possibilities for targeted marketing activities. This explosion of data holds out great promise for improved "data-based" marketing decisions, while at the same time posing substantial challenges to traditional estimation methods. For example, in pure predictive tasks we are faced with a huge number (more than 1 billion in some cases) of potential explanatory variables.

Firms have been quick to use new sources of consumer data as the basis for marketing actions. The principle way this new data has been used is to target messages

[1] Demographics inferred from the location of the residence of the consumer. Here the assumption is that consumers in a given geographic area are similar in demographic characteristics.

and advertisements in a much more customized and, hopefully, effective way. If, for example, which ad is displayed to a customer is a function of the customer's preferences, this creates a new set of challenges for statistical methods which are based on the assumption that explanatory variables are chosen exogenously or as though they are a result of a process independent of the outcome variable. Highly customized and targeted marketing activities make these assumptions untenable and put a premium on finding and exploiting sources of exogenous (random-like) variation.

Some would argue that the only true solution to the "endogeneity" problem created by targeted actions is true random variation of the sort that randomized experimentation is thought to deliver. Many in economics have come to a view that randomized experiments are one of the few ways to obtain a valid estimate of an effect of an economic policy. However, our goal in marketing is not just to estimate the effect of a specific marketing action (such as exposure to a given ad) but to find a policy which can help firms optimize marketing actions. It is not at all clear that optimization purely via randomized experimentation is feasible in the marketing context. Conventional randomized experiments can only be used to evaluate (without a model) discrete actions. With more than one marketing variable and many possibilities for each variable, the number of possible experiments required in a purely experimental approach becomes prohibitively large.[2]

In Section 2, we consider various frameworks for inference and their suitability given the desiderata of marketing applications. Given the historic emphasis on predictive validation in marketing and the renewed emphasis spurred by adoption of Machine Learning methods, it is important to review methods for evaluation models and inferences procedures. Pervasive heterogeneity in marketing applications has spurred a number of important methodological developments which we review in Section 3. We discuss the role of causal inference in marketing applications, discussing the advantages and disadvantages of experimental and non-experimental methods in Section 4. Finally, we consider the endogeneity problem and various IV approaches in Section 5.

2 Frameworks for inference

Researchers in marketing have been remarkably open to many different points of view in statistical inference, taking a decidedly practical view that new statistical methods and procedures might be useful in marketing applications. Currently, marketing researchers are busy investigating the potential for Machine Learning techniques to be useful in marketing problems. Prior to Machine Learning, Bayesian methods made considerable inroads into both industry and academic practice. Again, Bayesian methods were welcomed into marketing as long as they proved to be worthwhile in

[2] Not withstanding developments that use approximate multi-arm bandit solutions to reduce the cost of experimentation with many alternatives (Scott, 2014).

the sense of compensating for the costs of using the methods with improved inferences and predictions. In recent years, a number of researchers in marketing as well as allied areas of economics have brought perspectives from their training in economics to bear on marketing problems. In this section, we will review the various paradigms for inferences and provide our perspective on the relative advantages and disadvantages of each approach to inference.

To prove a concrete example for discussion, consider a simple marketing response model which links sales to marketing inputs.

$$y = f(x|\theta)$$

Here y is sales and x is a vector of inputs such as prices or promotional/advertising variables. The goal of this modeling exercise is to estimate the response surface for the purpose of making predictions regarding the level of sales expected for over the input space. These predictions can then be used to maximize profits and provide guidance to improve firm policies regarding the setting of these inputs. In the case of price inputs, the demand theory discussed in Chapter 1 of this volume can be used to select the functional form for $f\,()$. However, most researchers would want to explicitly represent the fact that sales data is not a deterministic function of the input variables – and represent deviations of y from that predicted from $f\,()$ as corresponding to draws from error distributions. It is common to introduce an additive error term into this model.

$$y = f\,(x|\theta) + \varepsilon \tag{1}$$

There are several ways to view ε. One way is to view the error term as arising from functional form approximation error. In this view, the model parameters can be estimated via pure projection methods such as non-linear least squares. Since the estimator is a projection, the error term is, by construction, orthogonal to f.

Another interpretation of the error term is as arising from omitted variables (such as variables which describe the environment in which the product is sold but are not observed or included in the x). These could also include unobservable demand shocks. In random utility models of demand, the error terms are introduced out of convenience to "rationalize" or allow for the fact that when markets or individuals are faced with the same value of x, they don't always demand the same quantity. For these situations, some further assumptions are required in order to perform inference. If we assume that the error terms are independent of x,[3] then we can interpret the projection estimator as arising from the moment condition

$$E_{x,\varepsilon}\left[\nabla f\,(x|\theta)\,\varepsilon\right] = 0.$$

Here ∇f is the gradient of the response surface with respect to θ. This moment condition can be used to rationalize the non-linear least squares projection in the

[3] In linear models, this assumption is usually given as ε is mean independent of x, $E\,[\varepsilon|x] = 0$. Since we are allowing for a general functional form in f, we must make a stronger assumption of full independence.

sense that we are choosing parameter values, θ, so that sample moment is held as close as possible to the population moment and the sample moment condition is the first order condition for non-linear least squares.

The interpretation of least squares as a method of moments estimator based on assumptions about the independence of the error term and the gradient of the response surface provides a "distribution-free" basis for estimation. This is the sense that we do not have to specify a particular distribution for the error term. The independence assumption assumes that the x variables are chosen independently of the error term or shocks to aggregate demand (conditional on x). In settings where the firm chooses x with some partial knowledge of the demand shocks, the independence assumption is violated and we must resort to other methods of estimation. In Section 5.3 below, we consider these methods.

2.1 A brief review of statistical properties of estimators

This chapter is not designed to be a reference on inference methods in general, but, instead, to discuss how features of marketing applications make particular demands of inference and what methods have shown promise in the marketing literature. However, to fix notation and to facilitate discussion, we provide a brief review of statistical properties of estimation procedures.

Prior to obtaining or analyzing a dataset, it is entirely reasonable to choose an estimation procedure for model parameters on the basis of the general properties of the procedure. Statistical properties for an estimation procedure are deduced by regarding the procedure as specifying a function of the data and studying the sampling properties of the estimator by considering the distribution of the estimator over repeated samples from a specific data generation mechanism.

That is, we have an estimator, $\hat{\theta} = g(D)$, where D represents that data. The estimator is specified by the function $g()$. Viewed in this fashion, the estimator is a random variable whose distribution comes from the distribution of the data via the summary function, g. The performance of the estimator must be gauged by specifying a loss function and examining the distribution of loss. A common loss function is the squared error loss function, $\ell(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^t A(\hat{\theta} - \theta)$, where A is a positive-definite weighting matrix. We can evaluate alternative estimation procedures by the comparing their distribution of loss. Clearly, we would prefer estimators with loss massed near zero. For convenience, many use $MSE \equiv E_D[\ell(\hat{\theta}(D), \theta)]$ as a measure of the distribution of squared error and look for estimators that offer the lowest possible value of MSE.

Unfortunately, there is no general solution to problem of finding the minimal MSE procedure for all possible values of θ even conditional on a specific family of data generating mechanisms/distributions. This problem is well illustrated by shrinkage estimation methods. Shrinkage methods modify an existing estimation procedure by "shrinking" the point estimates toward some point in the parameter space (typically 0). Thus, shrinkage methods will have lower MSE than base estimation procedure for values of θ near the shrinkage point but higher MSE far from the point.

As a practical matter, we do not expect arbitrarily large values of θ and this gives shrinkage methods an advantage in most applications. But the point is that shrinkage methods dominate the base method because they attempt to improve upon for only certain parts of the parameter space at the expense of inferior performance elsewhere. The best we can say is that we would like to use estimators in the admissible class of estimators – estimators that can't be improved upon everywhere in the parameter space.

Another way to understanding the problem is to observe that MSE can be re-expressed as relating to bias and the sampling variance. For a scalar estimator,

$$MSE = \left(E[\hat{\theta}] - \theta\right)^2 + \mathbb{V}(\hat{\theta}) = \text{Bias}^2 + \text{Variance}$$

Clearly, if one can find an estimation procedure that reduces both bias and sampling variance, this procedure will improve MSE. However, at some point in the pursuit of efficient estimators, there may well be a trade-off between these two terms. That is, we can reduce overall MSE by a favorable trade-off of somewhat larger bias for an even larger reduction in variance. Many of the modern shrinkage and variable selection procedures exploit this trade-off by finding a favorable point on the bias-variance trade-off frontier.

A further complication in the evaluation of estimation procedures is that the sampling distribution (and MSE) can be extremely difficult to derive and there may be no closed-form expression for MSE. This has led many statisticians and econometricians to resort to large sample or asymptotic approximations.[4] A large sample approximation is the result of an imaginary sampling experiment in which the sample size is allowed to grow indefinitely. While we may not be able to derive the sampling distribution of an estimator for a fixed N, we may be able to approximate its distribution for arbitrarily large N or in a limiting sense. This approach consists of two parts: (1) a demonstration that the distribution of an estimator is massed arbitrarily close to the true value for large enough N and (2) the use of some variant of the Central Limit Theorem to provide a normal large sample or asymptotic approximation to the sampling distribution.

In a large sample framework, we consider infinite increases in the sample size. Clearly, any reasonable procedure (with enough independence in the data) should "learn" about the true parameter value with access to an unlimited amount of data. A procedure that does not learn at all from large and larger datasets is clearly a fundamentally broken method. Thus, a very minimal requirement of all estimation procedures is that as N grows to infinity, we should see the sampling distribution massed closer and closer to the true value of θ with smaller and smaller MSE. This property is called *consistency* and is usually defined by what is called a probability limit or plim. The idea of a plim is very simple – if the mass of the sampling distribution becomes concentrated closer and closer to the true value then we should be able

[4] We note that, contrary to popular belief, the bootstrap does not provide finite sample inference and can only be justified by appeal to large sample arguments.

to find a sample size sufficiently large that for any sample of that size or larger we can make the probability that the estimator lies near the true value as large as possible.

It should be emphasized that the consistency property is a very minimal property. Among the set of consistent estimation procedures, there can be some procedures with lower MSE than others. As a simple example, consider estimation of the mean. The Law of Large numbers tells us that, under very minimal assumptions, the sample mean converges to the true mean. This will be true whether we use all observations in our sample or only every 10th observation. However, if we estimate the mean using only 1/10 of the data, this procedure will be consistent but inefficient relative to the standard procedure. If an estimation procedure has a finite sample bias, this bias will be reduced to zero in the asymptotic experiment or else the estimator would be inconsistent. For this reason, consistent estimators are sometimes referred to as "asymptotically unbiased." In other words, in large sample MSE converges to just sampling variance. Thus, asymptotic evaluation of estimators is entirely about comparison of sampling variance. Statistical efficiency in large samples is measured by sampling variance.

2.2 Distributional assumptions

To complete the econometric specification of the sales response model in (1), some would argue that we have to make an assumption regarding the joint distribution of the vector of error terms. Typically, we assume that the error terms are independent which is not really a restrictive assumption in a cross-section of markets. There is a prominent school of thought in econometrics that one should not make any more distributional assumptions than minimally necessary to identify the model parameters. This view is associated with the Generalized Method of Moments (see Hansen, 1982) and applications in macro/finance. In these applications, the moment conditions are suggested by economic theory and, therefore, are well-motivated. In our setting, the moment restriction is motivated by assumptions regarding the independence of the error term from the response variables which is not motivated by appeal to any particular economic theory of firm behavior. For these reasons, we believe that the argument for not making additional distributional assumptions is less forceful in marketing applications.

Another related argument from the econometrics literature is what some call the "consistency-efficiency" trade-off. Namely, that we might be willing to forgo the improved statistical efficiency afforded by methods which employ more modeling assumptions in exchange for a more "robust" estimator which gives up some efficiency in exchange for providing consistent estimates over a wider range of possible model specifications. In the case of a simple marketing mix response model and additive error terms with a continuous sales response, changes in the distribution model for the errors will typically not result in inconsistent parameter estimators. However, various estimators of the sampling variance of the estimators may not be consistent in the presence of heteroskedastic error terms. For this reason, various authors (see, for example, Angrist and Pischke, 2009) endorse the use of Eicker-White style heteroskedastic consistent estimators of the variance. If a linear approximation to an underlying non-linear response function is used, then we might expect to see heteroskedastic errors as the error term would include functional form approximation error. In marketing applications, the use of a linear regression function can be much less appealing than in economics. We seek to optimize firm behavior and linear approximations to sales response models will not permit computation of global optima.

For these reasons, we believe it is far less controversial, in marketing as opposed to economics, to complete the econometric specification with a specific distributional assumption for basic sales response and demand models. Moreover, in demand models which permit discrete outcomes as discussed in Chapter 1, then the specification of the error term is really part of the model and few are interested in model inferences that are free from a specific distributional choice for the model error terms. For example, the pervasive use of logit and other closely related models is based on the use of extreme value error terms which can be interpreted as marginal utility errors. While one might want to change the assumed distribution of error terms to create a different model specification, we do not believe that "distribution-free" inference has much value in models with discrete outcomes.

2.3 Likelihood and the MLE

Invariably, we will want to make inferences either for the generic response surface models as in (1) or in demand models which are derived from an assumed direct utility function along with a distribution of marginal utility errors. The generic demand model can be written as

$$y = g\left(p, E, \varepsilon | \theta\right) \tag{2}$$

where y is a vector of quantities demanded, p is a vector of prices, E is expenditure allocated to the group of products in the demand group, and ε is a vector of marginal utility errors. In either the case of (1) or (2), if we postulate a distribution of the error terms, this will induce a distribution on the response variable in (1) or the vector of demanded quantities given above.

As is well known, this will allow us to compute the likelihood of the observed data. In the case of the generic sales response model, the likelihood is the joint distribution of the observed data.

$$p\left(y, x | \theta, \psi\right) = p\left(y | x, \theta\right) p\left(x | \psi\right) \tag{3}$$

Here $p\left(y|x,\theta\right)$, the conditional distribution $y|x$ is derived from (1) along with an assumed distribution of the error term. With additive error terms, the Jacobian from ε to y is 1. In the demand model application, x consists of (p, E) and the Jacobian may be more complicated due (see Chapter 1 for many examples). In the case where the error terms are independent of the right hand side variables, the marginal distribution of the right hand variables, e.g. $p\left(x|\psi\right)$ above, will not depend on the parameters that govern the dependence of the left hand side variable on the right hand side variables. Under these assumptions, the likelihood for θ is proportional to the conditional distribution.

$$\ell(\theta) \propto p(y|x,\theta) \tag{4}$$

A very powerful principle is the *Likelihood Principle* which states that all sample-based information is reflected in the likelihood function. Or, put another way, the likelihood is sufficient for the data. Two datasets with the same likelihood function are informationally equivalent with respect to inferences about θ even though the datasets do not have to have identical observations.[5] Another consequence of the likelihood principle is that approaches which are not based on likelihood function are based on potentially inferior information and, therefore, may not be efficient.

Some statisticians believe that one must base inferences procedures on more than just the likelihood and do not endorse the likelihood principle. Typically, this is argued via special and somewhat pathological examples where strict adherence to the likelihood principle produces non-sensical estimators. We are not aware of any practical example in which it is necessary to take into account more than the likelihood to make sensible inferences.

Exactly how the likelihood is used to create estimators does not follow directly from the likelihood principle. The method of maximum likelihood creates an estimator using the maximum of the likelihood function.

$$\hat{\theta}_{MLE} = \arg\min \ell(\theta) = f(y,x) \tag{5}$$

The MLE appears to obey the likelihood principle in the sense that the MLE depends only on the likelihood function. However, the MLE only uses one feature of the likelihood function (the max). The analysis of the MLE depends only on the local behavior of the likelihood function in the vicinity of the MLE. The properties of the MLE can only be deduced via large sample analysis. Under moderately weak conditions, the MLE is consistent and asymptotically normal. The most attractive aspect of the MLE is that it attains the Cramer-Rao lower bound for sampling variance in large samples. In other words, there is no other consistent estimator which can have a smaller sampling variance in large samples. However, the finite sample properties of the MLE are not especially compelling. There are examples where the MLE is inadmissible. The analysis of the MLE strongly suggests if an estimator is to be considered useful it should be asymptotically equivalent to the MLE. However, in finite samples, an estimator may differ appreciably from the MLE and have superior sampling properties.

From a practical point of view, to use the MLE the likelihood must be evaluated at low cost and maximization must be feasible. In order for asymptotic inference to be conducted, the likelihood must be differentiable at least in a neighborhood of the

[5] In cases where the sampling mechanism does not depend on θ, then the likelihood principle states that inference should ignore the sampling mechanism. The classic example is the coin toss experiment. It does not matter whether the observed data of m heads in n coin tosses was acquired by tossing a coin n times or tossing a coin until m heads appear. This binomial versus negative binomial example has spurred a great deal of debate.

maximum. Typically, researchers used non-linear programming techniques to maximize the likelihood and most of these methods are gradient-based. There are other methods such as simulated annealing and simplex methods which do not require the gradient vector; however, these methods are slow and often impractical with a large number of parameters. In addition, a maximum of the likelihood is not useful without a method for conducting inference which requires computation of gradients and or second derivatives.

In many settings, we assume that economic agents act with a larger information set than is available to the researcher. In these situations, the likelihood function for the observed data must be evaluated by integrating over the distribution of unobservable variables. In many cases, this integration must be accomplished by numerical means. Integration by simulation-based methods creates a potentially non-smooth likelihood and can create problems for those who use gradient-based non-linear programming algorithms to maximize the likelihood.

Inference for the MLE is accomplished by reference to the standard asymptotic result:

$$\sqrt{N}\left(\hat{\theta}_{MLE} - \theta\right) \dot{\sim} N\left(0, \mathbb{I}^{-1}\right)$$

where $\mathbb{I}$ is the information matrix, $\mathbb{I} = -E\left[\frac{\partial^2 \ln \ell}{\partial\theta\partial\theta^t}\right]$. For any but the most trivial models, the expected information matrix must be approximated, typically by using an estimate of the Hessian evaluated at the MLE. Most optimizers provide such an estimate on convergence. The quality of Hessian estimates particularly where numerical gradients are used can be very low and it is advisable to expend additional computational power at the optimum to obtain the highest quality Hessian estimate possible. Typically, numerical Hessians should be symmetrized.[6] In some cases, numerical Hessians can be close to singularity (ill-conditioned) and cannot be used to estimate the asymptotic variance-covariance matrix of the MLE without further "regularization."[7] The average outer product of the log-likelihood function can also be used to approximate the information matrix.

2.4 Bayesian approaches

As mentioned above, there is a view by some econometricians that any likelihood-based approach requires additional assumptions that may not be supported by economic theory and that can be viewed as somewhat arbitrary. For this reason, non-likelihood approaches which make minimal assumptions are viewed favorably. In our view, the fact that one has to specify a complete model is a benefit rather than a cost of a likelihood-based approach to inference. That is to say, it is easy to see that some data has zero likelihood even though non-likelihood based methods might be used to

[6] $A^* = .5A + .5A^t$.

[7] Typically achieved by adding a small amount to each of the diagonal elements of the Hessian estimate.

make inferences. For example, if we postulate a discrete choice model, we must eliminate any data that shows consumers purchasing more than one product in the demand group at a single time. This would violate the mutually exclusive assumption of the choice model. However, we could use moment conditions to estimate a choice model which has zero likelihood for a given dataset. To guard against mis-specification, our view is that a more fruitful approach is to develop models specifically designed to accommodate the important features of marketing datasets and to be able to easily change the specifications to perform a sensitivity analysis.

Given that the MLE has only desirable large sample properties, there is a need for a inference framework which adheres to the likelihood principle but offers better finite sample properties. It is well known than Bayesian procedures are asymptotically similar to the MLE but have desirable finite sample properties. Bayes estimators can be shown to be admissible and it can also be shown that all admissible estimators are Bayes. Bayesian procedures are particularly useful in very high dimensional settings and in highly structured multi-level models. Bayesian approaches have many favorable properties including shrinkage and adaptive shrinkage (see Section 3.2 for discussion of adaptive shrinkage) in which the shrinkage adapts to the information in the data. In the machine learning literature, these properties are called "regularization" which simply means reducing sampling error by avoiding outlandish or absurd estimates and by various forms of shrinkage. Many popular regularization methods such as the Lasso can be given a Bayesian interpretation (Park and Casella, 2008).

There are many treatments of the Bayesian approach (see, for example, Gelman et al., 2004 and Rossi et al., 2005). We briefly review the Bayesian approach. Bayesians take the point of view that any unknown quantity (including model parameters, but also including predictions and which model governs the data) should be described by a probability distribution which represents our current state of information. Given the limited information available to any researcher, no quantity that cannot be directly measured is known with certainty. Therefore, it is natural to use the machinery of probability theory to characterize the degree of uncertainty regarding any unknown quantity. Information arises from two sources: (1) prior information and (2) from the data via the likelihood. Prior information can either come from other datasets or from economic theory (such as monotonicity of demand) or from "structure" such as I believe there is a super-population from which a given dataset is drawn from. Bayes theorem provides the way in which prior and sample information is brought together to make "after the data" or a posteriori inferences. In the case of our simple response model example, Bayes theorem states

$$p(\theta|y, X) = \frac{p(y, \theta|X)}{p(y|X)} = \frac{p(y|X, \theta)\, p(\theta)}{p(y|X)} \tag{6}$$

Thus, given a prior distribution on the response parameters, Bayes theorem tells us how to combine this prior with the likelihood to obtain an expression for the posterior distribution. The practical value of Eq. (6) is that the posterior distribution of the model parameters is proportional to the likelihood times the prior density.

$$p(\theta|y, X) \propto \ell(\theta|y, X)\, p(\theta) \tag{7}$$

All features of the posterior distribution are inherited from the likelihood and the prior. Of course, this equation does not define an estimator without some further assumptions. Under squared error loss, the posterior mean is the Bayes estimator, $\hat{\theta}_{Bayes} = \int \theta p\,(\theta|y, X)\,d\theta$. However, Bayesians do not think of the Bayesian apparatus as simply a way to obtain an estimator but rather as a complete, and different, method of inference. The posterior distribution provides information about what we have learned from the data and the prior and expresses this information as a probability distribution. The degree of uncertainty can be expressed by the posterior probability of various subsets of the parameter space (for example, the posterior probability that a price elasticity parameter is less than -1.0). The marginal distributions of individual parameters in the θ vector are often used to characterize uncertainty though the computation of the Bayesian analogues of the familiar standard errors and confidence intervals. The posterior standard deviation for an element of the θ vector is analogous to the standard error and the quantiles of the posterior distribution can be used to construct a Bayesian "credibility" interval (the analogue of the confidence interval). Of course, the joint posterior offers much more information than the marginals – unfortunately this information is rarely explored or provided.

It is also important to note that Bayes estimators use both information from the prior and the likelihood. While the MLE is based only the likelihood, an informative prior serves to modify location of posterior and influences the Bayes estimators. All Bayes estimators with informative priors can be interpreted as a form of shrinkage estimator. The likelihood is centered at the MLE which (within the confines of the particular parametric model) is a "greedy" estimator which tries to fit the data according the likelihood criterion. The prior serves to "shrink" or pull the Bayes estimator into a sort of compromise between the prior and the likelihood. In simplified conjugate cases, the posterior mean can be written as an average of the prior mean and the MLE where the weights in the average depend on the relative informativeness of the prior relative to the likelihood. Often the prior mean is set to zero and the Bayes estimator will shrink the MLE toward zero which improves sampling properties by exploiting the bias-variance tradeoff. However, as the sample size increases, the weight accorded the likelihood increases relative to the prior, reducing shrinkage and allowing the Bayes estimator to achieve consistency.

Most Bayesian statisticians are quick to point out that there are very fundamental differences between the Bayesian measure of uncertainty and the sampling theoretic ones. Bayesian inference procedures conditional on the observed data which differs dramatically from sampling theoretic approaches that consider imaginary experiments in which new datasets are generated from the same model. Bayesians argue very convincingly that sampling properties can be useful to select a method of inference but that the applied researcher is interested in the information content of a given dataset. Moreover, Bayesian procedures do not depend on asymptotic experiments which are even of more questionable relevance for a researcher who wishes to summarize the information in one finite sample.

The Bayesian approach is appealing due to superior sampling properties coupled with the appropriate inference statements that conditional on a specific dataset. The problem is that the Bayesian approach appears to exact higher costs than a standard MLE or method of moments approach. The cost is twofold: (1) a prior distribution must be provided and (2) some practical method must be provided to compute the many integrals that are used to summarize the posterior distribution.

2.4.1 The prior

Recognizing that the prior is an additional "cost" which many busy researchers might not be interested in providing, the Bayesian statistics literature pursued the development of various "reference" priors.[8] The idea is that the "reference" priors might be agreed upon by researchers as providing modest or minimal prior information and that the "reference" priors can be assessed at low cost. It is our view that the literature on reference priors is largely a failure in the sense that there can't really be one prior or form of prior that is satisfactory in all situations. Our view is that informative priors are useful and that even a modest amount of prior information can be exceptionally useful in the sense of eliminating absurd or highly improbable parameter estimates. Priors and prior structure become progressively more important as the parameterization of models becomes increasingly complex and high dimensional.

Bayesian methods have become almost universally adopted in the analysis of conjoint survey data[9] and in the fitting of marketing mix models. These successes in adoption of Bayesian methods come from the regularization and shrinkage properties afforded Bayesian estimators, particularly in what is called the hierarchical setting. In Section 3, we explore Bayesian approaches to the analysis of panel data which provide an element of what is termed adaptive shrinkage – namely, a multilevel prior structure which can be inferred partly on the basis of data from other units in the panel. This notion, called "borrowing strength," is a key attribute of Bayesian procedures with highly structured and informative priors.

In summary, the assessment of an informative prior is a requirement of the Bayesian approach beyond likelihood. In low dimensional settings, such as a linear regression with a small number of potential regressors and only one cross-section or time series, any in a range of moderately informative priors will produce similar results and, therefore, the prior is not important or particularly burdensome. However, in high dimensional settings such as flexible non-parametric models or in the case of panel data where there are many units and relatively few observations, the Bayesian approach provides a practical procedure where the prior is important and confers important sampling benefits. Even the simple regression example becomes a convincing argument for the Bayesian approach when the number of potential regressors becomes extremely large. Many of the variable selection Machine Learning techniques can be interpreted as Bayesian procedures. In the background, there is an informative prior which is assessed indirectly, typically through cross-validation.

[8] See, for example, Bernardo and Smith (1994), Sections 5.4 and 5.62.

[9] For example, Sawtooth Software implements a Bayesian hierarchical model for choice-based conjoint. Sawtooth is the marketshare leader. SAS has several procedures for fitting systems of regression equations using Bayesian approaches and these are widely applied in the analysis of aggregate market data.

Thus, Bayesian methods have become useful even in the simple linear regression problem. For example, Lasso and Ridge regression are Bayesian methods with particular prior forms (see, for example, Park and Casella, 2008).

2.4.2 Bayesian computation

Historically, Bayesian methods were not used much in applications due to the problems with summarizing the posterior via computation of various integrals. For example, the minimal requirement of many investigators is to provide a point estimate and a measure of uncertainty. For the Bayesian, this involves computing the posterior mean and posterior standard deviation both of which are integrals involving the marginal posterior distribution of a given parameter. That is, inference regarding θ_i requires computation of the marginal posterior distribution,

$$p_i\left(\theta_i|y, X\right) = \int_{\theta_{-i}} p\left(\theta|y, X\right) d\theta_{-i}$$

Here θ_{-i} is all elements of the θ vector except the ith element. In addition, we must compute the normalizing constant for the posterior since the likelihood times the prior is only proportional to the posterior. Clearly, these integrals are available as closed-form solutions for only very special cases. Numerical integration methods such as quadrature-based methods are only effective for very low dimensional integrals.

The current popularity of Bayesian methods has been make possible by various simulation-based approaches to posterior computation. Obviously if we could simulate from the posterior at low computational cost and only require knowledge of the posterior up to a normalizing constant, this would provide a practical solution. While iid samplers from arbitrary multivariate distributions are not practical, various Markov Chain methods have been devised that can effectively simulate from the posterior at low cost. These MCMC (Markov Chain Monte Carlo) methods (see the classic treatment in Robert and Casella (2004) for a complete treatment and Rossi et al. (2005) for many applications to models of interest to marketing) create a continuous state space Markov Chain whose invariant distribution is the posterior. The accuracy of this method is determined by-the ability to simulate large number of draws from the Markov Chain at low cost as well as our ability to construct Markov Chains with limited dependence. MCMC-based Bayes procedures use simulations or draws from the Markov Chain to compute the posterior expectation of any arbitrary function of the model parameters. For example, if we have R draws from the chain, then a simulation-based estimate of the posterior expectation of any function can be obtained by simply forming an average over the function evaluated at each of these R draws. Typically, we use extremely large numbers of draws (typically greater than 10,000) to ensure that the error in the simulation approximation is small. Note that the number of draws used in under the control of the investigator (contrast to the fixed sample size of the data).

$$\widehat{E_{\theta|y,X}\left[g\left(\theta\right)\right]} = \frac{1}{R}\sum_{r=1}^{R} g\left(\theta_r\right) \quad (8)$$

Thus, our simulation-based estimates are averages of draws from a Markov Chain constructed with an invariant distribution equal to the posterior. While there is a very general theory that assures ergodicity (convergence of these ensemble averages to posterior expectations), in practice, the draws from the Markov Chain can be highly correlated, requiring a very large number of draws. For the past 25 years, Bayesian statisticians and econometricians have enlarged the class of models which can be treated successfully by MCMC methods. Advancement in computation also has allowed applications to data sets and models that were previously thought to be impossible to analyze. We routinely analyze highly nonlinear choice models with thousands of cross-sectional units and tens of thousands of observations. Multivariate mixtures of normals in high dimensions and with a large number of components can be implemented on laptop computing equipment using an MCMC approach.

Given the availability of MCMC methods (particularly the Gibbs Sampler), statisticians have realized that many models whose likelihoods involve difficult integrals can be analyzed with Bayesian methods using the principle of data augmentation. For example, models with latent random variables that must be integrated out to form the likelihood can be analyzed from in a Bayesian approach by augmenting the parameter space with these latent variables and defining a Markov Chain on this "augmented" state space. Marginalizing (integrating) out the latent can be achieved trivially by simply discarding draws of the latent variable. These ideas have been applied very successfully to random coefficient models as well as models like the multinomial and multivariate Probit.

In summary, the Bayesian approach was previously thought to be impractical given lack of a computational strategy for computing various integrals of the non-normalized posterior. MCMC methods have not only eliminated this drawback but, with the advent of data augmentation, have made Bayesian procedures the only practical methods for models with a form of likelihood that involves integration. The current challenge to Bayesian methods is to apply to truly enormous data sets generated by millions of consumers and a vast number of potential explanatory variables. As currently implemented, MCMC methods are fundamentally sequential in nature (the rth simulate of the Markov Chain depends on the value of the $(r-1)$st simulate). The vast computing power currently available is obtained not by the speed of any one given processor but the ability to break the computing task into pieces and farm this out to large array of processors. Sequential computations do not naturally lend themselves to anything other than very tightly coupled computer architectures. This is a current area of research which awaits innovation in our methods as well as possible changes in the computing environment. In Section 3, we review some of the approaches to "scaling" MCMC methods to truly huge panel data sets.

2.5 Inference based on stochastic search vs. gradient-based optimization

Up to this point, we have viewed MCMC methods as a method for indirectly sampling from the posterior distribution. The theory of MCMC methods says that, if Markov

Chain is allowed to run long enough, the procedure will visit any "appreciable" set with frequency proportional to the posterior probability of that set. Clearly, then a MCMC sampler will visit high probability areas of the posterior much more often low probability areas. In some cases, the set of possible parameter values is so large that, as a practical matter, the MCMC will only visit a part of the parameter space. For example, consider the application of Bayesian methods to the problem of selecting regression models from the space of all possible regression models (if there are k possible regressions there are 2^k possible models). For $k > 30$, any of the MCMC methods (see, for example, George and McCulloch, 1997) will visit only a subset of the possible models and one would not necessarily want to use the frequency of model visits as an estimate of the posterior probability of a model.

Thus, one way of looking at the MCMC method is as a method of stochastic search. All MCMC methods are designed to draw points from the parameter space with some degree of randomness. Standard methods such as the Gibbs Sampler or random walk MCMC methods do not rely on any gradient information regarding the posterior (note: there are variational and stochastic gradient methods that do use gradient information). This means that the parameter space does not have to be continuous (as in the case of variable selection) nor does the likelihood have to be smooth. There are some problems which give rise to a likelihood function with discrete jumps (see, for example, Gilbride and Allenby, 2004). Gradient-based MLE methods simply cannot be applied in such situations. However, an MCMC method does not require even continuity of the likelihood function.

In other situations, the likelihood function requires an integral. The method of simulated maximum likelihood simply replaces that integral with a simulation-based estimate of the integral. For example, the integral might be taken over a normal distribution of consumer heterogeneity. Given the simplicity and low cost of normal draws, a simulated MLE seems to be a natural choice to evaluate the likelihood function numerically. However, given that only a finite number of draws are used to approximate the integral, the likelihood function is now non-differentiable and gradient-based maximization methods can easily fail. A Bayesian has a choice of whether to use a random walk or similar MCMC method directly on the likelihood function evaluated by simulation-based integration or to augment the problem with latent variables. In either case, the Bayesian using stochastic search methods is not dependent on any smoothness in the likelihood function.

2.6 Decision theory

Most of the recent Bayesian literature in marketing emphasizes the value of the Bayesian approach to inference, particularly in situations with limited information. Bayesian inference is only a special case of the more general Bayesian decision-theoretic approach. Bayesian Decision Theory has two critical and separate components: (1) a loss function and (2) the posterior distribution. The loss function associates a loss with a state of nature and an action, $L(a, \theta)$, where a is the action and θ is the state of nature. The optimal decision maker chooses the action so as to minimize expected loss where the expectation is taken with respect to the posterior distribution.

$$\min_{a} \bar{L}(a) = \int L(a,\theta)\, p(\theta|\text{Data})\, d\theta$$

Inference about θ can be viewed as a special case of decision theory where the "action" is to choose an estimate based on the data. Model choice can also be thought of as a special case of decision theory. If the loss function associated with model choice is takes on the value of 1 if the model is correct and 0 if not, then the solution which minimizes expected loss is to select the model (from a set of models) with highest posterior probability (for examples and further details see Chapter 6 of Rossi et al., 2005).

2.6.1 Firms profits as a loss function

In the Bayesian statistical literature, decision theory has languished as there are few compelling loss functions, only those chosen for mathematical convenience. The loss function must come from the subject area of application and is independent of the model. That is to say, a principle message of decision theory is that we use the posterior distribution to summarize the information in the data (via likelihood) and prior and that decisions are made as a function of that information and a loss function. In marketing, we have a natural loss function, namely, the profit function of the firm. Strictly speaking, the profit function is not a loss function which we seek to minimize. We maximize profits or minimize the negative of profits.

To take the simple sales response model presented here, the profit function here would be

$$\pi(x|\theta) = E\left[y|x,\theta\right](p - c(x)) = f(x|\theta)(p - c(x)) \tag{9}$$

where $y = f(x|\theta) + \varepsilon$ and $c(x)$ is the cost of providing the vector of marketing inputs.[10] Optimal decision theory prescribes that we should make decisions so as to maximize the posterior expectation of the profit function in (9).

$$x^* = \operatorname{argmax} \bar{\pi}(x)$$
$$\bar{\pi}(x) = \int \pi(x|\theta)\, p(\theta|\text{Data})\, d\theta$$

The important message is that we act to optimize profits based on the posterior expectation of profits rather than inserting our "best guess" of the response parameters (the plug-in approach) and proceeding as though this estimate is the truth. The "plug-in" approach can be thought of as expressing overconfidence in the parameter estimates. If the profit function is non-linear in θ then the plug-in and full decision theoretic

[10] It is a simple matter to include covariates out of the control of the firm in the sales response surface. Optimal decision could either be done conditional on this vector of covariates or the covariates could be integrated out according to some predictive distribution.

approaches will yield different solutions and the plug-in approach will typically overstate potential profits.

In commercial applications of marketing research, many firms offer what are termed marketing mix models. These models are built to help advise firms how to allocate their budgets over many possible marketing activities including pricing, trade promotions, and advertising of many kinds including TV, print, and various forms of digital advertising. The options in digital advertising have exploded and now include sponsored search, web-site banner ads, product placement ads, and social advertising. The marketing mix model is designed to attack the daunting task of estimating the return on each of these activities and making predictions regarding the consequence of possible reallocation of resources on firm profits. As indicated above, the preferred choice for estimation in the marketing mix applications are Bayesian methods applied to sets of regression models. However, in making recommendations to clients, the marketing mix modeler simply "plugs-in" the Bayes estimates and is guilty of overconfidence. The problem with this approach is that, if taken literally, the conclusion is often to put all advertising resources in only one "bucket" or type of advertising. The full decision theoretic approach avoids these problems created by overconfidence in parameter estimates.

2.6.2 Valuation of information sets

An important problem in modern marketing is the valuation of information. Firms have an increasing extensive array of possible information sets on which to base decisions. Moreover, acquisition of information is considered a major part of strategic decisions. For example, Amazon's recent acquisition of Whole Foods as opening of brick and mortar stores has been thought to be motivated by the rich set of off-line information which can be accumulated by observing Amazon customers in these store environments. In China, retailing giants Alibaba and JD.com have built what some term "data ecosystems" that can link customers across many different activities including web-browsing, social media, and on-line retail. On-line ad networks and programmatic ad platforms offer unprecedented targeting opportunities based on information regarding consumer preferences and behavior. The assumption behind all of these developments is that new sources of information are extremely valuable.

The valuation of information is clearly an important part of marketing. One way of valuing information is in the ability of this new information to provide improved estimates of consumer preferences and, therefore, more precise predictions of consumer response to various marketing activities. However, statistically motivated estimation and prediction criteria such as mean squared error do not place a direct monetary valuation on information. This can only be obtained in a specific decision context and a valid loss function such as firm profits. To make this clear, consider two information sets, A and B, regarding sales response parameters. We can value these information sets by solving the decision theoretic problem and comparing the attainable expected profits for the two information sets. That is, we can compute

$$\Pi_k = \max_x \int \pi\,(x|\theta)\, p_k\,(\theta)\, d\theta$$
$$k = A, B$$

where $p_k\,(\theta)$ is the posterior based on information set k.

In situations where decisions can be made at the consumer level rather than at the aggregate level, information set valuation can be achieved within a hierarchical model via different predictive distributions of consumer preferences based on alternative information sets (in Section 3, we will provide a full development of this idea).

2.7 Non-likelihood-based approaches

2.7.1 Method of moments approaches

As indicated above, the original appeal of the Generalized Method of Moments (GMM) methods is that they use only a minimal set of assumptions consistent with the predictions of economic theory. However, when applied in marketing and demand applications, the methods of moments is often used as a convenient way of estimating the parameters of a demand model. Given a parametric demand model, there are a set of moment conditions which can identify the parameters of the model. These moment conditions can be used to define a method of moments estimator even for a fully parametric model. The method of moments approach is often chosen to avoid deriving the likelihood of the data and associated Jacobians. While this is true, the method of moments approach does not specify which set of moments should be used and there is often an infinite number of possible sets of moments conditions, any one of which is sufficient to identify and estimate model parameters. The problem, of course, is that method of moments provides little guidance as to which set of moments to use. The most efficient procedure for method of moments is to use the score function (gradient of the expected log-likelihood) to define the moment conditions. This means, of course, that one can only approach the asymptotic efficiency of the maximum likelihood estimator.

In other situations, the method of moments approach is used to estimate a model which is only partially specified. That is, most parts of the model have a specific parametric form and specific distributional assumptions, but investigators purport to be reluctant to fully specify other parts of the model. We do not understand why it is defensible to make full parametric assumptions about part of a model but not others when there is no economic theory underpinning any of the parametric assumptions made. GMM advocates would argue that fewer assumptions is always better than more assumptions. The problem, then, is which parts of the model are designated for specific parametric assumptions and which parts of the model are not? The utility of the GMM approach must be judged relative to the arguments made by the investigator in defense of a particular choice of which part of the model is left unspecified. Frequently, we see no arguments of this sort and, therefore, we conclude that the method of moments procedure was chosen primarily for reasons of convenience.

The aggregate share model of Berry et al. (1995) provides a good example of this approach. The starting point for the BLP approach is to devise a model for aggregate share data that is consistent with valid demand models postulated at the individual level. For example, it is possible to take the standard multinomial logit model as the model governing consumer choice. In a market with a very large number of consumers, the market shares are the expected probabilities of purchase which would be derived by integrating the individual model over the distribution of heterogeneity. The problem is that, with a continuum of consumers, all of the choice model randomness would be averaged out and the market shares would be a deterministic function of the included choice model covariates. To overcome this problem, Berry et al. (1995) introduced an additional error term into consumer level utility which reflects a market-wide unobservable. For their model, the utility of brand j for consumer i and time period t is given by

$$U_{ijt} = X_{jt}\theta_j^i + \eta_{jt} + \varepsilon_{ijt} \tag{10}$$

where X_{jt} is a vector of brand attributes, θ_j^i is a $k \times 1$ vector of coefficients, η_{jt}, is an unobservable common to all consumers, and ε_{ijt} is the standard idiosyncratic shock (i.i.d. extreme value type I). If we normalize the utility of the outside good to zero, then market shares (denoted by s_{jt}) are obtained by integrating the multinomial logit model over a distribution of consumer parameters, $f\left(\theta^i|\delta\right)$, $\theta^i = \left[\theta_1^i, \ldots, \theta_J^i\right]$. δ is the vector of hyper-parameters which govern the distribution of heterogeneity.

$$\begin{aligned} s_{jt} &= \int \frac{\exp\left(X_{jt}\theta_j^i + \eta_{jt}\right)}{1 + \sum_{k=1}^{J} \exp\left(X_{kt}\theta_k^i + \eta_{kt}\right)} f\left(\theta^i|\delta\right) d\theta^i \\ &= \int s_{ijt}\left(\theta^i|X_t, \eta_t\right) f\left(\theta^i|\delta\right) d\theta^i \end{aligned}$$

While it is not necessary to assume that consumer parameters are normally distributed, most applications assume a normal distribution. In some cases, difficulties in estimating the parameters of the mixing distribution force investigators to further restrict the covariance matrix of the normal distribution to a diagonal matrix (see Jiang et al., 2009). Assume that $\theta^i \sim N\left(\bar{\theta}, \Sigma\right)$, then the aggregate shares can be expressed as a function of aggregate shocks and the preference distribution parameters.

$$s_{jt} = \int \frac{\exp\left(X_{jt}\theta^i + \eta_{jt}\right)}{1 + \sum_{k=1}^{J} \exp\left(X_{kt}\theta_k^i + \eta_{kt}\right)} \phi\left(\theta^i|\theta, \Sigma\right) d\theta^i = h\left(\eta_t|X_t, \bar{\theta}, \Sigma\right) \tag{11}$$

where η_t is the $J \times 1$ vector of common shocks.

If we make an additional distributional assumption regarding the aggregate shock, η_t, we can derive the likelihood. Given that we have already made specific assumptions regarding the form of the utility function, the distribution of the idiosyncratic choice errors, and the distribution of heterogeneity, this does not seem particularly restrictive. However, the recent literature on GMM methods for aggregate share models does emphasize the lack of distributional assumptions regarding the aggregate shock. In theory, the GMM estimator should be robust to autocorrelated and heteroskedastic errors of an unknown form. We will assume that the aggregate shock is i.i.d. across both products and time periods and follows a normal distribution, $\eta_{jt} \sim N\left(0, \tau^2\right)$. The normal distribution assumption is not critical to the derivation of the likelihood; however, as Bayesians we must make some specific parametric assumptions. Jiang et al. (2009) propose a Bayes estimator based on a normal likelihood and document that this estimator has excellent sampling properties even in the presence of misspecification and, in all cases considered, has better sampling properties than a GMM approach (see Chen and Yang, 2007 and Musalem et al., 2009 for other Bayesian approaches).

The joint density of shares at "time" t (in some applications of aggregate share models, shares are observed over time for one market and in other shares are observed for a cross-section of markets. In the latter case, the "t" index would index markets) can be obtained by using standard change of variable arguments.

$$\begin{aligned}\pi\left(s_{1t},\ldots,s_{Jt}|X,\bar{\theta},\Sigma,\tau^2\right) &= \phi\left(h^{-1}\left(s_{1t},\ldots,s_{Jt}|X,\bar{\theta},\Sigma\right)|0,\tau^2 I_J\right) J_{(\eta\rightarrow s)} \\ &= \phi\left(h^{-1}\left(s_{1t},\ldots,s_{Jt}|X,\bar{\theta},\Sigma\right)|0,\tau^2 I_J\right)\left(J_{(s\rightarrow\eta)}\right)^{-1}\end{aligned} \tag{12}$$

$\phi\left(\cdot\right)$ is the multivariate normal density. The Jacobian is given by

$$J_{(s\rightarrow\eta)} = \left\|\frac{\partial s_j}{\partial \eta_k}\right\| \tag{13}$$

$$\frac{\partial s_j}{\partial \eta_k} = \begin{cases} \int -s_{ij}\left(\theta^i\right) s_{ik}\left(\theta^i\right)\phi\left(\theta^i|\bar{\theta},\Sigma\right) & k \neq j \\ \int s_{ij}\left(\theta^i\right)\left(1-s_{ik}\left(\theta^i\right)\right)\phi\left(\theta^i|\bar{\theta},\Sigma\right) & k = j \end{cases} \tag{14}$$

It should be noted that, given the observed shares, the Jacobian is a function of Σ only (see Jiang et al., 2009 for details).

To evaluate the likelihood function based on (12), we must compute the h^{-1} function and evaluate the Jacobian. The share inversion function can be evaluated using the iterative method of BLP (see Berry et al., 1995). Both the Jacobian and the share inversion require a method for approximation of the integrals required to compute "expected share" as in (11). Typically, this is done by direct simulation; that is, averaging over draws from the normal distribution of consumer level parameters. It has been noted that the GMM methods can be sensitive to simulation error in the evaluation of the integral as well as errors in computing the share inversion. Since the number of integral estimates and share inversions is of the order of magnitude of the number of likelihood or GMM criterion evaluations, it would desirable, from a strictly numerical point of view, that the inference procedure exhibit little sensitivity to the number of iterations of the share inversion contraction or the number of simulation draws used in the integral estimates. Our experience is that the Bayesian methods that use stochastic search as opposed to optimization are far less sensitive to these numerical errors. For example, Jiang et al. (2009) show that the sampling properties of Bayes estimates are virtually identical when 50 or 200 simulation draws are used in the approximation of the share integrals; this is not true of GMM estimates.[11]

In summary, the method of moments approach has inferior sampling properties to a likelihood-based approach and the literature has not fully explored the efficiency losses of using method of moments procedures. The fundamental problem is that the set of moments conditions is arbitrary and there is little guidance as to how to choose the most efficient set of moment conditions.[12]

2.7.2 Ad hoc approaches

Any function of the data can be proposed as an estimator. However, unless that function is suggested or derived from a general approach to inference that has established properties, there is no guarantee that proposed estimator will have favorable properties. For example, any Bayesian estimator with non-dogmatic priors will be consistent as is true with MLE and Method of Moment estimators under relatively weak conditions. Estimators outside these classes (Bayes, MLE, and MM), we term "ad hoc" estimators as there are no general results establishing the validity of the estimator procedure. Therefore, the burden is on the investigator proposing an estimator not based on established principles to demonstrate, at a minimum, that the estimator is consistent. Unfortunately, pure simulation studies cannot establish consistency. It is well known that various biased estimators (for example, Bayes) can have very favorable finite sample properties by exploiting the bias-variance trade-off. However, as the amount of sample information becomes greater in large samples, all statistical procedures should reduce bias. The fact that a procedure can be constructed to do well by some criterion for a few parameter settings does not insure this. This is the value of theoretical analysis.

Unfortunately, there are examples in the marketing literature (see Chapter 3 on conjoint methods) of procedures that have been proposed that are not derived from an inference paradigm that insures consistency. That is not to say that these procedures are inconsistent, but that consistency should and has not been established. Our view is that consistency is a necessary condition which must be made in order to admit an estimation procedure for further evaluation. It may well be, that a procedure offers superior (or inferior) performance to existing methods but this must wait until this minimal property is established. Some of the suggestions provided in the marketing literature are based purely on optimization method without establishing that the criterion for optimization is derived from a valid probability model. Thus, establishing consistency for these procedures is apt to be difficult.

In proposing new procedures, investigators should be well aware of the complete class theorem which states that all admissible estimators are Bayes estimators. In

[11] It is possible to lessen the sensitivity of the method of moments approach to numerical inversion error (Dubé et al., 2012).

[12] The GMM literature does have results about the asymptotically optimal moment conditions but, for many models, it is impractical to derive the optimal set of moment conditions.

other words, it is not possible to dominate a Bayes estimator in finite samples for a specific and well-defined model (likelihood). Thus, procedures which are designed to be competitive with Bayes estimators must be based on robustness considerations (that is procedures that make fewer distributional assumptions in hopes of remaining consistent across a broader class of models).

2.8 Evaluating models

Given the wide variety of models as well as methods for estimation of models, a methodology for comparison of models is essential. One approach is to view model choice as a special case of Bayesian decision theory. This will lead the investigator to compute the posterior probability of a model.

$$p(M_i|y) = \frac{p(y|M_i)\,p(M_i)}{p(y)} \tag{15}$$

where M_i denotes model i. Thus, Bayesian model selection is based on the "marginal" likelihood, $p(y|M_i)$, and the prior model probability. We can simply select models with the highest value of the numerator of (15) or we can average predictions across models using these posterior probabilities. The practical difficulty with the Bayesian approach to model selection is that the marginal likelihood of the data must be computed.

$$p(y|M_i) = \int p(y|\theta, M_i)\,p(\theta|M_i)\,d\theta \tag{16}$$

This integral is difficult to evaluate using only the MCMC draws that are used to perform posterior inferences regarding the model parameters (see, for example, the discussion in Chapter 6 of Rossi et al., 2005). Not only are these integrals difficult to evaluate numerically, but the results are highly sensitive to the choice of prior. Note also that the proper priors would be required for unbounded likelihoods in order to obtain convergence of the integral which defines the marginal likelihood.

We can view the marginal likelihood as the expected value of the likelihood taken over the prior parameter distribution. If you have a very diffuse (or dispersed) prior, then the expected likelihood can be small. Thus, the relative diffusion of priors for different models must be considered when computing posterior model probabilities. While this can be done, it often involves a considerable amount of effort beyond that required to perform inferences conditional on a model specification. However, when properly done, the Bayesian approach to model selection does afford a natural penalty for over-parameterized models (the asymptotic approximation known as the Schwarz approximation shows the explicit dependence of the posterior probability on model size; however, the Schwarz approximation is notoriously inaccurate and, therefore, not of great practical value).

Given difficulties in implementing a true decision theoretic approach to model selection, investigators have searched for other methods which are more easily implemented while offering some of the benefits of the more formal approach. Investigators are aware that, given the "greedy" algorithms that dominate estimation, in-sample measures of model fit will understate true model error in prediction or classification. For this reason, various predictive validation exercises have become very popular in both marketing and the Machine Learning literatures. The standard predictive validation exercise involves dividing the sample into two data sets (typically by random splits): (1) estimation dataset and (2) validation dataset. Measures of model performance such as MSE are computed by fitting the model to the estimation dataset and predicting out-of-sample on the validation data. This procedure will work to remove the over-fitting biases of in-sample measures of MSE. Estimation procedures which have favorable bias-variance trade-offs will perform well by the criteria of predictive validation. The need to make arbitrary divisions of the data into estimation and validation datasets can be removed by using a k-fold cross-validation procedure. In k-fold cross-validation, the data is divided randomly into k "folds" or subsets. Each fold is reserved for validation and the model is fit on the other $k-1$ folds. This is averaged over many draws of the fold classification and can be shown to produce an unbiased estimate of the model prediction error criterion.

While these validation procedures are useful in discriminating between various estimation procedures and models, some caution should be exercised in their application to marketing problems. In marketing, our goal is to optimize policies for selection of marketing variables and we must consider models that are policy invariant. If our models are not policy invariant, then we may find models that perform very well in pure predictive validation exercises make poor predictions for optimal policy determination. In Section 4, we will consider the problem of true causal inference. The causal function linking market variables to outcomes such as sales can be policy invariant. The need for causal inference may also motivate us to consider other estimation procedures and these are explored in Section 5.

3 Heterogeneity

A fundamental premise of marketing is that customers differ both in preferences for product features as well as their sensitivities to marketing variables. Observable characteristics such as psycho-demographics can only be expected to explain a limited portion of the variation in tastes and responsiveness. Disaggregate data is required in order to measure customer heterogeneity.[13] Typically, disaggregate data are obtained for a relatively large number of cross-sectional units but with a relatively short history of activity. In the consumer packaged goods industry, store level panel data are common, especially for retailers. There is also increased availability of customer level purchase data from specialized panels of consumers or from detailed purchase

[13] Some argue that, with specific functional forms, the heterogeneity distribution can be determined from aggregate data. Fundamentally, the functional forms of response models and the distribution of heterogeneity are confounded in aggregate data.

histories assembled from firm records. As the level of aggregation decreases, discrete features of sales data become magnified. The short time span of panel data coupled with the comparatively sparse information in discrete data means that we are unlikely to have a great deal of sample information about any one cross-sectional unit. If inference about unit-level parameters is important, then Bayesian approaches will be important. Moreover, the prior will matter and there must be reasonable procedures for assessing informative priors.

Increasingly firms want to make decentralized marketing decisions that exploit more detailed disaggregate information. Examples include store or zone level pricing, targeted electronic couponing, and sales force activities in the pharmaceutical industry. All of these examples involve allocation of marketing resources across consumers or local markets and the creation of possibly customized marketing treatments for each unit. In digital advertising, the ability to target an advertising message at a very specific group of consumers, defined by both observable and behavioral measures, makes the modeling of heterogeneity even more important.

The demands of marketing applications contrast markedly with applications in micro-economics where the average response to a variable is often deemed more important. However, even the evaluation of policies which are uniform across some set of consumers will require information about the distribution of preferences in order to evaluate the effect on social welfare.

In this section, we will review approaches to modeling heterogeneity, particularly the Bayesian hierarchical modeling approach. We will also discuss some of the challenges that truly huge panel datasets offer for Bayesian approaches.

3.1 Fixed and random effects

A generic formulation of the panel data inference problem is that we observe a cross-section of H units over "time." The panel does not necessarily have to be balanced, namely each unit can have a different number of observations. For some units, the number of observations may be very small. In many marketing context, a new "unit" maybe "born" with 0 observations, but we still have to make predictions for this unit. For example, a pharmaceutical company has a very expensive direct sales force that calls on many key "accounts" which in this industry is defined as a prescribing physician. There may be some physicians with a long history of interaction with the company and others who are new "accounts" with no history of interaction. Any firm that acquires new customers over time faces the same problem. Many standard econometric methods simply have no answer to this problem.

Let $p\left(y_h|\theta_h\right)$ be the model we postulate at the unit level. If the units are consumers or households, this likelihood could be a basic demand model, for example. Our goal is to make inferences regarding the collection $\{\theta_h\}$. The "brute force" solution would be to conduct separate likelihood-based analyses for each cross-sectional unit (either Bayes or non-Bayes). The problem is that many units may have so little information (think singular X matrix or choice histories in which a unit did not purchase all of the choice alternatives) that the unit-level likelihood does not have a maximum. For this reason, separate analyses is not practical. Instead, the coefficients allowed to be unit-specific is limited.

The classic example is what is often called at "Fixed Effects" estimator which has its origins in a linear regression model with unit specific intercepts and common coefficients.

$$y_{ht} = \alpha_h + \beta' x_{ht} + \varepsilon_{ht} \tag{17}$$

Here the assumption is that there is a common "effect" or coefficients on the x variables but individual-specific intercepts and that there are unit-level intercept parameters. A common interpretation of this set-up is that there is some sort of unobservable variable(s) which influences the outcome, y, and which varies across the units, h. With panel data, we can just label the effect of these unobservables as time-invariant intercepts and estimate the intercepts with panel data. Sometimes econometricians will characterize this as solving the "selection on unobservables" problem via the introduction of fixed effects. Advocates for this approach will explain that no assumptions are made regarding the distribution of these unobservables across units nor are the unobservables required to be independent of the included x variables.

For a linear model, estimation of (17) is a simple matter of concentrating the $\{\alpha_h\}$ out of the likelihood function for the panel data set. This can be done by either subtracting the unit-level means of all variables or by differencing over time.

$$y_{ht} - \bar{y}_{h.} = \beta^t \left(x_{ht} - x_{h.}\right) + \varepsilon_{ht} - \varepsilon_{h.} \tag{18}$$

In most cases, there is no direct estimation of the intercepts but, instead, the demeaning operation removes a part of the variation of both the independent and dependent variables from the estimation of the β terms. It is very common for applied econometricians to use hundreds if not thousands of fixed effect terms in estimation of linear panel data models. The goal is to isolate or control for unobservables that might compromise clean estimation of the common β coefficients. Typically, a type of sensitivity analysis is done where groups of fixed effect terms are entered or removed from the specification and changes in the β estimates are noted.

Of course, this approach to heterogeneity does not help if the goal is to make predictions regarding a new unit with no data or a unit with so little data even the fixed effect intercept estimator is no defined. The reason is that there is no common structure assumed for the intercept terms. Each panel unit is unique and there is not source of commonality of similarity across units. This problem does not normally trouble econometricians who are concerned more with "effect estimation" or estimating β rather than prediction for new units or units with insufficient information. In marketing applications, we have no choice – we must make predictions for all units in the analysis.

The problems with the fixed effects approach to heterogeneity does not stop with prediction for units with insufficient data. The basic idea that unobservables have additive effects and simply change the intercept and the assumption of a linear mean function allows the econometrician to finesse the problem of estimating the fixed effects by concentrating them out of the likelihood function. That is, there is a transformation of the data so that the likelihood function can be factored into two terms – one term does not involve β and the other term only involves the data through the transformation. Thus, we lose no information by "demeaning" or differencing the data. This idea does not extend to non-linear models such as discrete choice or non-linear demand models. This problem is so acute that many applied econometricians fit "linear probability" models to discrete or binary data in order to use the convenience of the additive intercept fixed effects approach even thought they know that a linear probability model is very unlikely to fit their data well and must only be regarded as an approximation to the true conditional mean function.

This problem with the fixed effects approach does not apply to the random coefficients model. In the random coefficient model, we typically do not distinguish between the intercept and slope parameters and simply consider all model coefficients to be random (iid) draws from some "super-population" or distribution. That is, we assume the following two part model:

$$\begin{aligned} y &\sim p\left(y|\theta_h\right) \\ \theta_h &\sim p\left(\theta_h|\tau\right) \end{aligned} \tag{19}$$

Here the second equation is the random coefficient model. Almost without exception, the random coefficient model is taken to be a multivariate normal model, $\theta_h \sim N\left(\bar{\theta}, \Sigma\right)$. It is also possible to parameter the mean of the random coefficient model by observable characteristics of each cross sectional unit, i.e. $\bar{\theta} = \Delta z$ where z is a vector of unit characteristics. However, there is still the possibility that there are unobservable unit characteristics that influence x explanatory variables in each unit response model. The random coefficient model assumes independence between the random effects and the levels of unit x variables conditional on z. Many regard this assumption as an drawback to the random coefficient model and point out that a fixed effects specification does not require this assumption. Manchanda et al. (2004) explicitly model the joint distribution of random effects and unit level x variables as one possible approach to relaxing the conditional independence assumption used in standard random coefficient models.

If we start with a linear model, then the random coefficient model can be expressed as a linear regression model with a special structured covariance matrix as in

$$y_{ht} = \bar{\theta}^t x_{ht} + \varepsilon_{ht} + v_h^t x_{ht} \tag{20}$$

Here $\theta = \bar{\theta} + v$, $v \sim N\left(0, \Sigma\right)$. Thus, we can regard the regression model as estimating the mean of the random coefficient distribution and the covariance matrix is inferred from the likelihood of the correlated and heteroskedastic error terms. This error term structure motivates the use of "cluster" covariance matrix estimators. Of course, what we are doing by substituting the random coefficient model into the unit level regression is integrating out the $\{\theta_h\}$ parameters. In the more general non-linear setting, those who insist upon doing maximum likelihood would regard the random coefficient model as part of the model and integrate or "marginalize" out the unit-level parameters.

$$\ell(\tau) = \prod_{h=1}^{H} \int p(y_h|\theta_h)\, p(\theta_h|\tau)\, d\theta_h \tag{21}$$

Some call random coefficient models "mixture" models since the likelihood is a mixture of the unit level distribution over an assumed random coefficient distribution.

Maximum likelihood estimation of random coefficient models requires an approximation to the integral in the likelihood function over τ.[14] Investigators find that they must restrict the dimension of this integral (by assuming that only parts of the coefficient vector are random) in order to obtain reasonable results. As we will see below, Bayesian procedures finesse this problem via data augmentation. The set of random coefficients are "augmented" in the parameter space, exploiting the fact that given the random coefficients, inference regarding τ is often easily accomplished.

Mixed logit models

Econometricians are keenly aware of the limitations of the multinomial logit model to represent the demand for a set of products. The multinomial logit model has only one price coefficient and thus the entire matrix of cross-price elasticities must be generated with that one parameter. This problem is often known as the IIA property of logit models. As is shown in Chapter 1, one way of interpreting the logit model with linear prices is as corresponding to demand derived from linear utility with extreme value random utility errors. Linear utility assumes that all products are perfect substitutes. The addition of the random utility error means that choice alternatives are no longer exact perfect substitutes but the usual iid extreme value assumption means that all products have substitutability differences that can be expressed as a function of market share (or choice probability) alone.

However, applied choice modelers are quick to point out that if aggregate demand is formed as the integral of logits over a normal distribution of preferences, then this aggregate demand function no longer has the IIA property. Our own experience is that while this is certainly true as a mathematical statement that aggregate preferences often exhibit elasticity structures which are very close to those implied by IIA. Our experience is that high correlations in the mixing distribution are required to obtain large deviations form IIA in the aggregate demand system.

Many of the claims regarding the ability of mixed logit to approximate arbitrary aggregate demand systems stem from a misreading of McFadden and Train (2000). A superficial reading of this article might imply that mixed logits can approximate any demand structure but this is only true if explanatory variables such as price are allowed to enter the choice probabilities in arbitrary non-linear ways. In some sense, it must always be true that any demand model can be approximated by arbitrary

[14] This situation seems to be a clear case where simulated MLE might be used. The integral in (21) is approximated by a set of R draws from the normal distribution.

functions of price. One should not conclude that mixtures of logits with linear price terms can be used to approximate arbitrary demand structures.

3.2 Bayesian approach and hierarchical models

3.2.1 A generic hierarchical approach

Consider a cross-section of H units, each with a likelihood, $p\left(y_h|\theta_h\right), h = 1, \ldots, H$. θ_h is a $k \times 1$ vector. y_h generically represents the data on the hth unit and θ_h is a vector of unit-level parameters. While there is no restriction on the model for each unit, common examples include a multinomial logit or standard regression model at the unit level. The parameter space can be very large and consists of the collection of unit level parameters, $\{\theta_h, h = 1, \ldots, H\}$. Our goal will be to conduct a posterior analysis of these joint set of parameters. It is common to assume that units are independent conditional on θ_h. More generally, if the units are *exchangeable* (see Bernardo and Smith, 1994), then we require a prior distribution which is the same no matter what the ordering of the units are. In this case, we can write down the posterior for the panel data as

$$p\left(\theta_1, \ldots, \theta_H | y_1, \ldots, y_H\right) \propto \prod_{h=1}^{H} p\left(y_h|\theta_h\right) p(\theta_1, \ldots, \theta_H|\tau) \tag{22}$$

τ is a vector of prior parameters. The prior assessment problem posed by this model is daunting as it requires specifying a potentially very high dimensional joint distribution. One simplification would be to assume that the unit-level parameters are independent and identically distributed, *a priori*. In this case, the posterior factors and inference can be conducted independently for each of the H units.

$$p\left(\theta_1, \ldots, \theta_H | y_1, \ldots, y_H\right) \propto \prod_{h=1}^{H} p\left(y_h|\theta_h\right) p(\theta_h|\tau) \tag{23}$$

Given τ, the posterior in (23) is the Bayesian analogue of the classical *fixed effects* estimation approach. However, there are still advantages to the Bayesian approach in that an informative prior can be used. The informative prior will impart important shrinkage properties to Bayes estimators. In situations in which the unit-level likelihood may not be identified, a proper prior will regularize the problem and produce sensible inferences. The real problem is a practical one in that some guidance must be provided for assessing the prior parameters, τ.

The specification of the conditionally independent prior can be very important due to the scarcity of data for many of the cross-sectional units. Both the form of the prior and the values of hyper-parameters are important and can have pronounced effects on the unit-level inferences. For example, consider a normal prior, $\theta_h \sim N\left(\bar{\theta}, V_\theta\right)$. Just the use of a normal prior distribution is highly informative regardless of the value of hyper-parameters. The thin tails of the prior distribution will reduce the influence of the likelihood when the likelihood is centered far away from the prior. For this reason, the choice of the normal prior is far from innocuous. For many applications, the shrinkage of outliers is a desirable feature of the normal prior. The prior results in very stable estimates but at the same time this prior might mask or attenuate differences in consumers. It will, therefore, be important to consider more flexible priors. In other situations, the normal prior may be inappropriate. Consider the problem of random coefficient distribution of price coefficients in a demand model. Here we expect that the population distribution puts mass only on negative values and that the distribution would be highly skewed and possibly with a fat left tail. The normal random coefficient distribution would not be appropriate. It is a simple matter to reparameterize the price coefficient as in $\beta_p = -\exp(\beta_p^*)$ where we assume β_p^* is normal. But the general point that the normal distribution is restrictive is important to note. That is why we have enlarged these models to consider a finite or even infinite mixture of normals which can flexibly approximate any continuous distribution.

If we accept the normal form of the prior as reasonable, a method for assessing the prior hyper-parameters is required (Allenby and Rossi, 1999). It may be desirable to adapt the shrinkage induced by use of an informative prior to the characteristics of both the data for any particular cross-sectional unit as well as the differences between units. Both the location and spread of the prior should be influenced by both the data and our prior beliefs. For example, consider a cross-sectional unit with little information available. For this unit, the posterior should shrink toward some kind of "average" or representative unit. The amount of shrinkage should be influenced both by the amount of information available for this unit as well as the amount of variation across units. A hierarchical model achieves this result by putting a prior on the common parameter, τ. The hierarchical approach is a model specified by a sequence of conditional distributions, starting with the likelihood and proceeding to a two-stage prior.

$$\begin{gathered} p\left(y_h | \theta_h\right) \\ p\left(\theta_h | \tau\right) \\ p\left(\tau | a\right) \end{gathered} \tag{24}$$

The prior distribution on $\theta_h|\tau$ is sometimes called the first stage prior. In non-Bayesian applications, this is often called a random effect or random coefficient model and is regarded as part of the likelihood. The prior on τ completes the specification of a joint prior distribution on all model parameters and is often called the "second-stage" prior. Here a is a vector of prior hyper-parameters which must be assessed or chosen by the investigator.

$$p\left(\theta_1, \ldots, \theta_H, \tau | h\right) = p\left(\theta_1, \ldots, \theta_H | \tau\right) p\left(\tau | a\right) = \prod_{h=1}^{H} p\left(\theta_h | \tau\right) p\left(\tau | a\right) \tag{25}$$

One way of regarding the hierarchical model is just as a device to induce a joint prior on the unit-level parameters, that is we can integrate out τ to inspect the implied prior.

$$p\left(\theta_1, \ldots, \theta_H | a\right) = \int \prod_{h=1}^{H} p\left(\theta_h | \tau\right) p\left(\tau | a\right) d\tau \tag{26}$$

It should be noted that, while $\{\theta_h\}$ are independent conditional on τ, the implied joint prior can be highly dependent, particularly if the prior on τ is diffuse (note: it is sufficient that the prior on τ should be proper in order for the hierarchical model to specify a valid joint distribution). To illustrate this, consider a linear model, $\theta_h = \tau + v_h$. τ acts as common variance component and the correlation between any two θs is

$$Corr\left(\theta_h, \theta_k\right) = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_v^2}$$

As the diffusion of the distribution of τ relative to v increases, this correlation tends toward one.

3.2.2 Adaptive shrinkage

The popularity of the hierarchical model stems from the improved parameter estimates that are made possible by the two stage prior distribution. To understand why this is the case, let's consider a simplified situation in which each unit level likelihood is approximately normal with mean $\hat{\theta}_{MLE}$ and covariance matrix, I_h^{-1} (here we are abstracting from issues involving existence of the MLE). If we have a normal prior, $\theta_h \sim N\left(\bar{\theta}, V_\theta\right)$, then conditional on the normal prior parameters the approximate posterior mean is given by

$$\tilde{\theta} = \left(I_h + V_\theta^{-1}\right)^{-1} \left(I_h \hat{\theta}_h + V_\theta^{-1} \bar{\theta}\right) \tag{27}$$

This equation demonstrates the principle of shrinkage. The Bayes estimator is a compromise between the MLE (where the unit h likelihood is centered) and the prior mean. The weights in the average depend on the information content of the unit level likelihood and the variance of the prior. As we have discussed above, shrinkage is what gives Bayes estimators such excellent sampling properties. The problem becomes where should we shrink toward and by how much? In other words, how do we assess the normal mean and variance-covariance matrix. The two-part prior in the hierarchical model means that the data will be used (in part) to assess these parameter values. The "mean" of the prior will be something like the mean of the θ_h parameters over units and the variance will summarize the dispersion or extent of heterogeneity. This means that if we think that all units are very similar (V_θ is small) then we will shrink a lot. In this sense the hierarchical Bayes procedures has "adaptive shrinkage." Note also that, for any fixed amount of heterogeneity, units with a great deal of information regarding θ_h will not be shrunk much.

3.2.3 MCMC schemes

Given the independence of the units conditional on θ_h, all MCMC algorithms for hierarchical models will contain two basic groups of conditional distributions.

$$\begin{gathered} p\left(\theta_h | y_h, \tau\right), \quad h = 1, \ldots, H \\ p\left(\tau | \left\{\theta_h\right\}, a\right) \end{gathered} \tag{28}$$

As is well-known, the second part of this scheme exploits the conditional independence of y_h and τ. The first part of (28) is dependent on the form of the unit-level likelihood, while the second part depends on the form of the first stage prior. Typically, the priors in the first and second stages are chosen to exploit some sort of conjugacy and the $\{\theta_h\}$ are treated as "data" with respect to the second stage.

3.2.4 Fixed vs. random effects

In classical approaches, there is a distinction made between a "fixed effects" specification in which there are different parameters for every cross-sectional unit and random effects models in which the cross-sectional unit parameters are assumed to be draws from a super-population. Advocates of the fixed effects approach explain that the approach does not make any assumption regarding the form of the distribution or the independence of random effects from included covariates in the unit-level likelihood. The Bayesian analogue of the fixed effects classical model is an independence prior with no second-stage prior on the random effects parameters as in (23). The Bayesian hierarchical model is the Bayesian analogue of a random effects model. The hierarchical model assumes that each cross-sectional unit is exchangeable (possibly conditional on some observable variables). This means that a key distinction between models (Bayesian or classical) is what sort of predictions could be made for a new cross-sectional unit. In either the classical or Bayesian "fixed effects" approach, no predictions can be made about a new member of the cross-section as there is no model linking units. Under the random effects view, all units are exchangeable and the predictive distribution for the parameters of a new unit is given by

$$p\left(\theta_{h*} | y_1, \ldots, y_H\right) = \int p\left(\theta_{h*} | \tau\right) p\left(\tau | y_1, \ldots, y_H\right) d\tau \tag{29}$$

3.2.5 First stage priors

Normal prior

A straightforward model to implement is a normal first stage prior with possible covariates.

$$\theta_h = \Delta' z_h + v_h, \quad v_h \sim N\left(0, V_\theta\right) \tag{30}$$

where z_h is a $d \times 1$ vector of observable characteristics of the cross-sectional unit. Δ is a $d \times k$ matrix of coefficients. The specification in (30) allows the mean of each of the elements of θ_h to depend on the z vector. For ease of interpretation, we find it useful to subtract the mean and use an intercept.

$$z_h = (1, x_h - \bar{x})$$

In this formulation, the first row of Δ can be interpreted as the mean of θ_h.

(30) specifies a multivariate regression model and it is convenient, therefore, to use the conjugate prior for the multivariate regression model.

$$\begin{gathered} V_\theta \sim IW\left(\underline{V}, \underline{\nu}\right) \\ \delta = vec\left(\Delta\right) | V_\theta \sim N\left(\underline{\bar{\delta}}, V_\theta \otimes \underline{A}^{-1}\right) \end{gathered} \tag{31}$$

$\underline{A}$ is a $d \times d$ precision matrix. $vec\,()$ is the stacks the columns of a matrix up and $\otimes$ denotes Kronecker product. This prior specification allows for direct one-for-one draws of the common parameters, δ and V_θ.

Mixture of normals prior

While the normal distribution is flexible, there is no particular reason to assume a normal first-stage prior. For example, if the observed outcomes are choices among products, some of the coefficients might be brand specific intercepts. Heterogeneity in tastes for a product might be more likely to assume the form of clustering by brand. That is, we might find "clusters" of consumers who prefer specific brands over other brands. The distribution of tastes across consumers might then be multi-modal. We might want to shrink different groups of consumers in different ways or shrink to different group means. A multi-modal distribution will achieve this goal. For other coefficients such as a price sensitivity coefficient, we might expect a skewed distribution centered over negative values. Mixtures of multivariate normals are one way of achieving a great deal of flexibility (see, for example, Griffin et al., 2010 and the references therein). Multi-modal, thick-tailed, and skewed distributions are easily achieved from mixtures of a small number of normal components. For larger numbers of components, virtually any joint continuous distribution can be approximated. The mixture of normals model for the first-stage prior is given by

$$\begin{gathered} \theta_h = \Delta' z_h + v_h \\ v_h \sim N\left(\mu_{ind}, \Sigma_{ind}\right) \\ ind \sim MN\left(\pi\right) \end{gathered} \tag{32}$$

π is a $K \times 1$ vector of multinomial probabilities. This is a latent version of a mixture of K normals model in which a multinomial mixture variable, denoted here by *ind*, is used. In the mixture of normal specification, we remove the intercept term from z_h and allow v_h to have a non-zero mean. This allows the normal mixture components to mix on the means as well as on scale, introducing more flexibility. As before, it is convenient to demean the variables in z. A standard set of conjugate priors can be used for the mixture probabilities and component parameters, coupled with a standard conjugate prior on the Δ matrix.

$$\begin{gathered} \delta = vec\left(\Delta\right) \sim N\left(\underline{\bar{\delta}}, \underline{A}_{\delta}^{-1}\right) \\ \pi \sim D\left(\underline{\alpha}\right) \\ \mu_k \sim N\left(\underline{\bar{\mu}}, \Sigma_k \otimes \underline{a}_{\mu}^{-1}\right) \\ \Sigma_k \sim IW\left(\underline{V}, \underline{\nu}\right) \end{gathered} \tag{33}$$

Assessment of these conjugate priors is relatively straightforward for diffuse settings. Given that the θ vector can be of moderately large dimension (>5) and the θ_h parameters are not directly observed, some care must be exercised in the assessment of prior parameters. In particular, it is customary to assess the Dirichlet portion of the prior by using the interpretation that the $K \times 1$ hyper-parameter vector, $\underline{\alpha}$, is an observed classification of a sample of size, $\sum \underline{\alpha}_k$, into the K components. Typically, all components in $\underline{\alpha}$ are assessed equal. When a large number of components are used, the elements of α should be scaled down in order to avoid inadvertently specifying an informative prior with equal prior probabilities on a large number of components. We suggest a setting of $\underline{\alpha}_k = .5/K$ (see Rossi, 2014a, Chapter 1 for further discussion).

As in the single component normal model, we can exploit the fact that, given the $H \times k$ matrix, Θ, whose columns consist of each θ_h values and standard conditionally conjugate priors in (33), the mixture of normals model in (32) is easily handled by a standard unconstrained Gibbs sampler which includes augmentation to include the latent vector of component indicators (see Rossi et al., 2005, Section 5.5.1). The latent draws can be used for clustering as discussed below. We should note that any label-invariant quantity such as a density estimate or clustering is not affected by the "label-switching" identification problem (see Fruhwirth-Schnatter, 2006 for a discussion). In fact, the unconstrained Gibbs sampler is superior to various constrained approaches in terms of mixing.

A tremendous advantage of Bayesian methods when applied to mixtures of normals is that, with proper priors, Bayesian procedures do not overfit the data and provide reasonable and smooth density estimates. In order for a component to obtain appreciable posterior mass, there must be enough structure in the "data" to favor the component in terms of a Bayes factor. As is standard in Bayesian procedures, the existence of a prior puts an implicit penalty on models with a larger number of components. It should also be noted that the prior for the mixture of normals puts positive probability on models with less than K components. In other words, this is really a prior on models of different dimensions. In practice, it is common for the posterior mass to be concentrated on a set of components of much smaller size than K.

The posterior distribution of any ordinate of the joint (or marginal densities) of the mixture of normals can be constructed from the posterior draws of component parameters and mixing probabilities. In particular, a Bayes estimate of a density ordinate can be constructed.

$$\hat{d}(\theta) = \frac{1}{R}\sum_{r=1}^{R}\sum_{k=1}^{K}\pi_k^r \phi\left(\theta | \mu_k^r, \Sigma_k^r\right) \tag{34}$$

Here the superscript r refers to an MCMC posterior draw and $\phi(\cdot)$ the k-variate multivariate normal density. If marginals of sub-vectors of θ are required, then we simply compute the required parameters from the draws of the joint parameters.

3.2.6 Dirichlet process priors

While it can be argued that a finite mixture of normals is a very flexible prior, it is true that the number of components must be pre-specified by the investigator. Given that Bayes methods are being used, a practical approach would be to assume a very large number of components and allow the proper priors and natural parsimony of Bayes inference to produce reasonable density estimates. For large samples, it might be reasonable to increase the number of components in order accommodate greater flexibility. The Dirichlet Process (DP) approach can, in principle, allow the number of mixture components to be as large as the sample size and potentially increase with the sample size. This allows for a claim that a DP prior can facilitate general non-parametric density estimation. Griffin et al. (2010) provide a discussion of the DP process approach to density estimation. We review only that portion of this method necessary to fix notation for use within a hierarchical setting.

Consider a general setting in which each θ_h is drawn from a possibly different multivariate normal distribution.

$$\theta_h \sim N(\mu_h, \Sigma_h)$$

The DP process prior is a hierarchical prior on the joint distribution of $\{(\mu_1, \Sigma_1), \ldots, (\mu_H, \Sigma_H)\}$. The DP prior has the effect of grouping together cross-section units with the same value of (μ, Σ) and specifying a prior distribution for these possible "atoms."

The DP process prior is denoted $G(\alpha, G_0(\lambda))$. $G(\cdot)$ specifies a distribution over distributions that is centered on the base distribution, G_0, with tightness parameter, α. Under the DP prior, G_0 is the marginal prior distribution for the parameters for any one cross-sectional unit. α specifies the prior distribution on the clustering of units to a smaller number of unique (μ, Σ) values. Given the normal base distribution for the cross-sectional parameters, it is convenient to use a natural conjugate base prior.

$$G_0(\lambda): \quad \mu_h | \Sigma_h \sim N\left(\underline{\bar{\mu}}, \frac{1}{\underline{a}} \times \Sigma_h\right), \quad \Sigma_h \sim IW\left(\underline{V}, \underline{\nu}\right) \tag{35}$$

λ is the set of prior parameters in (35): $\left\{\underline{\bar{\mu}}, \underline{a}, \underline{\nu}, \underline{V}\right\}$.

In our approach to a DP model, we also put priors on the DP process parameters, α and λ. The Polya Urn representation of the DP model can be used to motivate the choice of prior distributions on these process parameters. α influences the number of unique values of (μ, Σ) or the probability that a new set of parameter values will be "proposed" from the base distribution, G_0. λ governs the distribution of proposed values. For example, if we set λ to put high prior probability on small values of Σ, then the DP prior will attempt to approximate the density of parameters with normal components with small variance. It is also important that the prior on μ put support on a wide enough range of values to locate normal components at wide enough spacing to capture the structure of the distribution of parameters. On the other hand, if we set very diffuse values of λ then this will reduce the probability of the "birth" of a new component via the usual Bayes Factor argument.

α induces a distribution on the number of distinct values of (μ, Σ) as shown in Antoniak (1974).

$$Pr\left(I^* = k\right) = \left\| S_n^{(k)} \right\| \alpha^k \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \tag{36}$$

$S_n^{(k)}$ are Sterling numbers of the first kind. I^* is the number of clusters or unique values of the parameters in the joint distribution of $(\mu_h, \Sigma_h, h = 1, \ldots, H)$. It is common in the literature to set a Gamma prior on α. Our approach is to propose a simple and interpretable distribution for α.

$$p(\alpha) \propto \left(1 - \frac{\alpha - \underline{\alpha}_l}{\underline{\alpha}_u - \underline{\alpha}_l}\right)^{\phi} \tag{37}$$

$\alpha \in \left(\underline{\alpha}_l, \underline{\alpha}_u\right)$. We assess the support of α by setting the expected minimum and maximum number of components, I^*_{min} and I^*_{max}. We then invert to obtain the bounds of support for α. Rossi (2014b), Chapter 2, provides further details including the assessment of the ϕ hyper-parameter. It should be noted that this device does not restrict the support of the number of components but merely assesses an informative prior that puts most of the mass of the distribution of α on values which are consistent with the specified range in the number of unique components. A draw from the posterior distribution of α can easily be accomplished as I^* is sufficient and we can use a griddy Gibbs sampler as this is simply a univariate draw.

Priors on λ (35) can be also be implemented by setting $\bar{\mu} = 0$ and letting $V = \nu \upsilon I_k$. If $\Sigma \sim IW(\nu \upsilon I_k, \nu)$, then this implies $mode(\Sigma) = \frac{\nu}{\nu+2} \upsilon I_k$. This parameterization helps separate the choice of a location for the Σ matrix (governed by υ) from the choice of the tightness on the prior for Σ (ν). In this parameterization, there are three scalar parameters that govern the base distribution, (a, υ, ν). We take them to be a priori independent with the following distributions.

$$\begin{aligned} p(a, \upsilon, \nu) &= p(a)\, p(\upsilon)\, p(\nu) \\ a &\sim U\left(\underline{a}_l, \underline{a}_u\right) \\ \upsilon &\sim U\left(\underline{\upsilon}_l, \underline{\upsilon}_u\right) \\ \nu &= \dim(\theta_h) - 1 + \exp(z), \ z \sim U\left(\underline{z}_l, \underline{z}_u\right), \ z_l > 0 \end{aligned} \tag{38}$$

It is a simple matter to write down the conditional posterior given that the unique set of (μ, Σ) are sufficient. The set of I^* unique parameter values is denoted $\Delta^* \left\{ \left(\mu_i^*, \Sigma_j^* \right), j = 1, \ldots, I^* \right\}$. The conditional posterior is given by

$$p\left(a, v, \nu | \Delta^*\right) \propto \prod_{i=1}^{I^*} \left| a^{-1} \Sigma_i^* \right|^{-1/2} \exp\left(-\frac{a}{2} \left(\mu_i^*\right)' \left(\Sigma_i^*\right)^{-1} \mu_i^* \right)$$

$$\left|\nu v I_k\right|^{\nu/2} \left|\Sigma_i^*\right|^{-(\nu+k+1)/2} \operatorname{etr}\left(-\frac{1}{2} \nu v \Sigma_i^* \right) p\left(a, v, \nu\right) \qquad (39)$$

We note that the conditional posterior factors and that, conditional on Δ^*, a and (ν, v) are independent.

3.2.7 Discrete first stage priors

Both the economics (Heckman and Singer, 1984) and marketing literatures (see, for example, references in Allenby and Rossi, 1999) considered the use of discrete random coefficient or mixing distributions. In this approach, the first state prior is a discrete distribution which puts mass on only a set of M unknown mass points. Typically, some sort of model selection criterion is used to select the number of mass points (such as BIC or AIC). As discussed in Allenby and Rossi (1999), discrete mixtures are poor approximations to continuous mixtures. We do not believe that consumer preferences consist of only a small number of types but, rather, a continuum of preferences which can be represented by a flexible continuous distribution. One can see this is a degenerate special case of the mixture of normals approach. Given that it is now feasible to use not only mixture of normals but mixtures with a potentially infinite number of components, the usefulness of the discrete approximation has declined. In situations where the model likelihood is extremely costly to evaluate (such as some dynamic models of consumer behavior and some models of search), the discrete approach retains some appeal from pure computational convenience.

3.2.8 Conclusions

In summary, hierarchical models provide a very appealing approach to modeling heterogeneity across units. Today almost all conjoint modeling (for details see Chapter 3) is accomplished using hierarchical Bayesian procedures applied to a unit level multinomial logit model. With more aggregate data, Bayesian hierarchical models are frequently employed to insure high dimensional systems of sales response equations produce reasonable coefficient estimates.

Given that there are a very large number of cross-sectional units in marketing panel data, there is an opportunity to move considerably beyond the standard normal distribution of heterogeneity. The normal distribution might not be expected to approximate the distribution of preferences across consumers. For example, brand preference parameters might be expected to be multi-modal while marketing mix sensitivity parameters such as a price elasticity or advertising responsiveness may be highly skewed and sign-constrained distributions.[15] For this reason, mixture-of-normal priors can be very useful (see, for example, Dube et al., 2010).

3.3 Big data and hierarchical models

It is not uncommon for firms to assemble panel data on millions of customers. Extremely large panels pose problems for the estimation and use of hierarchical models. The basic Gibbs sampler strategy in (28) means alternating between drawing unit level parameters (θ_h) and the common second-stage prior parameters (τ). Clearly, the unit level parameters can be draw in parallel which exploits a modern distributed computing environment with many loosely coupled processors. Moreover, the amount data which have to be sent to each processor undertaking unit-level computations is small. However, the draws of the common parameters require assembling all unit level parameters, $\{\theta_h\}$. The communications overhead of assembling unit level parameters may be prohibitive.

One way out of this computational bottleneck, while retaining the current distributed computer architecture, is to perform common parameter inferences on a subset (but probably a very large subset) of the data. The processor farm could be used to draw unit level parameters conditional on draws of the common parameters which have already been accomplished and reserved for this purpose. Thus, an initial stage of computation would be to implement the MCMC strategy on a large subset of units. Draws of the common parameters from this analysis would be reserved. In a second stage, common parameter draws would be sent down along with unit data to a potentially huge group of processors that would undertake parallel unit level computations. This remains an important area for computational research.

3.4 ML and hierarchical models

The Machine Learning (ML) literature has emphasized flexible models which are evaluated primarily on predictive performance. While the theory of estimation does not discriminate between two equally flexible approaches to estimating non-linear and unknown regression functions, as a practical matter investigators in the ML have found that certain functional forms or sets of basis functions appear to do very well in approximating arbitrary regression functions.

One could regard the entire hierarchical model approach entirely from a predictive point of view. To predict some unit level outcome variable, we should be able to use any function of the other observations on this unit. That is to predict, y_{ht_0}, we can use any other data on unit h except for y_{ht_0}. The hierarchical approach also suggests that summaries of the unit data (such as means and variance of unit parameters) might also be helpful in predicting y_{ht_0}. This suggests that we might consider ways of

[15] For example, the price coefficient could be reparameterized as in $\beta_p = -e^{\delta}$ and the first stage prior could be placed on δ. This will require care in assessment of priors as the δ parameter is on a log-scale while other parameters will be on a normal scale.

training flexible ML methods to imitate or approximate the predictions that arise from a hierarchical approach and use these as approximate solutions to fitting hierarchical models when it is impractical to implement the full-blown MCMC apparatus. Again, this might be a fruitful avenue for future research.

4 Causal inference and experimentation

As we have indicated, one fundamental goal of marketing research is to inform decisions which firms make about the deployment of marketing resources. At the core, all firm decisions regarding marketing involve counterfactual reasoning. For example, we must estimate what a potential customer would do had they not been exposed to a paid search ad in order to “attribute” the correct sales response estimate to this action. Marketing mix models pose a much more difficult problem of valid counterfactual estimates of what would happen to sales and profits if marketing resources were re-allocated in a different manner than observed in the past.

The importance of counterfactual reasoning in any problem related to optimization of resources raises the ante for any model of customer behavior. Not only must this model match the co-variation of key variables in the historical data, but the model must provide accurate and valid forecasts of sales in a new regime with a different set of actions. This means that we must identify the causal relationship between marketing variables and firm sales/profits and this causal relationship must be valid over a wide range of possible actions, including actions outside of the support of historical data.

The problem of causal inference has received a great deal of attention in the bio-statistics and economic literatures, but relatively little attention in the marketing literature. Given that marketing is, by its very nature, a decision-theoretic field, this is somewhat surprising. The problems in the bio-statistics and economics applications are usually evaluating the causal effect of a “treatment” such as a new drug or a job-training program. Typically, the models used in these literatures are simple linear models. Often the goal is to estimate a “local” treatment effect. That is, a treatment effect for those induced by an experiment or other incentives to become treated.

A classic example from this literature is the Angrist and Krueger (1991) paper which starts with the goal of estimating the returns to an additional year of schooling but ends up only estimating (with a great deal of uncertainty) the effect of additional schooling for those induced to complete the 10th grade (instead of leaving school in mid-year). To make any policy decisions regarding investment in education, we would need to know the entire causal function (or at least more than one point) for the relationship between years of education and wages. The analogy in marketing analytics is to estimate the causal relationship between exposures to advertising and sales. In order to optimize the level of advertising, we require the whole function not just a derivative at a point.

Much of the highly influential work of Heckman and Vytlacil (2007) has focused on the problem of evaluating job training programs where the decision to enroll in the program is voluntary. This means that those people who are most likely to benefit from the job training program or who have the least opportunity cost of enrolling (such as the recently unemployed) are more likely to be treated. This raises a host of thorny inference problems. The analogy in marketing analytics is to evaluate the effect of highly targeted advertising.

Randomized experimentation offers at least a partial solution to the problems of causal inference. Randomization in assignment to treatment conditions can be exploited as the basis of estimators for causal effects. Both academic researchers and marketing practitioners have long advocated the use of randomized experiments. In the direct marketing and credit card contexts, randomized field experiments have been conducted for decades to optimize direct marketing offers and manage credit card accounts. In advertising, IRI International used randomized experiments implemented through split cable to evaluate TV ad creatives (see Lodish and Abraham, 1995). In the early 1990s, randomized store-level experiments were used to evaluate pricing policies by researchers at the University of Chicago (see Hoch et al., 1994). In economics, the Income-Maintenance experiments of the 1980s stimulated an interest in randomized social experiments. These income maintenance experiments were followed by a host of other social experiments in housing and health care.

4.1 The problem of observational data

In the generic problem of estimating the relationship between sales and marketing inputs, the goal is to make causal inferences so that optimization is possible on the basis of our estimated relationship. The problem is that we often have only observational data on which to base our inferences regarding the causal nexus between marketing variables and sales. There is a general concern that not all of the variation in marketing input variables can be considered exogenous or as if the variation is the result of random experimentation. Concerns that some of the variation in the right hand side variables is correlated with the error term or jointly determined with sales mean that observational data may lead to biased or inconsistent causal inferences. For example, suppose we have aggregate time series data[16] on the sales of a product and some measure of advertising exposure.

$$S_t = f(A_t|\theta) + \varepsilon_t$$

Our goal is to infer the function, f, which can be interpreted as a causal function, that is, we can use this function to make valid predictions of expected sales for a wide range of possible values of advertising. In order to consider optimizing advertising,

[16] In the general case, assembling even observational data to fit a market response model can be difficult. At least three or possible four different sources are required: (1) Sales data, (2) Pricing and promotional data, (3) Digital advertising, and (4) Traditional advertising such as TV, Print, and Outdoor. Typically, these various data sources feature data at various levels of temporal, geographic, and product aggregation. For example, advertising is typically not associated with a specific product but with a line of products and may only be available at the monthly or quarterly level.

we require a non-linear function which, at least at some point, exhibits diminishing returns. Given that we wish to identify a non-linear relationship, we will require more extensive variation in A than if we assume a linear approximation. The question from the point of view of causal inference is whether or not we can use the variation in the observed data to make causal inferences.

As discussed in Section 2.3, the statistical theory behind any likelihood-based inference procedure for such a model assumes the observed variation in A is as though obtained via random experimentation. In a likelihood-based approach, we make the assumption that the marginal distribution of A is unrelated to the parameters, θ, which drive the conditional mean function. An implication of this assumption is that the conditional mean function is identified only via the effect of changes in A; the levels of A have no role in inference regarding the parameters that govern the derivative of $f()$ with respect to A. In practice, this may not be true. In general, if the firm sets the values of A observed in the data on the basis of the function $f()$, then the assumption that the marginal distribution of A is not related to θ is violated. In this situation, we may not be able to obtain valid (consistent) estimates of the sales response function parameters.[17] Manchanda et al. (2004) explain how a model in which both inputs are chosen jointly can be used to extract causal information from the levels of an advertising input variable. However, this approach requires additional assumptions about how the firm chooses the levels of advertising input.

Another possibility is that there is some unobservable variable that influences both advertising and sales. For example, suppose there are advertising campaigns for a competing product that is a close substitute and we, as data scientists, are not aware of or cannot observe this activity. It is possible that, when there is intensive activity from competitive advertising, the firm increases the scale of its advertising to counter or blunt the effects of competitive advertising. This means that we no longer estimate the parameters of the sales response function consistently. In general, anytime the firm sets A with knowledge of some factor that also affects Sales and we do not observe this factor, we will have difficulty recovering the sales response function parameters. In some sense, this is a generic and non-falsifiable critique. How do we know that such an unobservable does not exist? We can't prove it.

Typically, the way we might deal with this problem is to include as large a possible set of covariates in the sales equation as control variables. The problem in sales response model building is that we often do not observe any actions of competing products or we only observe these imperfectly and possibly at a different time frequency. Thus, one very important set of potential control variates is often not available. Of course, this is not the only possible set of variables observable to the firm but not observable to the data scientist. There are three possible ways to deal with this problem of "endogeneity."

[17] An early example is Bass (1969), with a model of the simultaneous determination of sales and advertising is calibrated using cigarette data. Bass suggested that ad hoc rules which allocate advertising budgets as some percentage of sales create a feedback loop or simultaneity problem.

1. We might consider using data sampled at a much higher frequency than the decisions regarding A are made. For example, if advertising decisions are made only quarterly, we might use weekly data and argue that the lion's share of variation in our data holds the strategic decisions of the firm constant.[18]
2. We might attempt to partition the variation in A into that which is "clean" or unrelated to factors driving sales and that which is. This is the logical extension of the conditioning approach of adding more observables to the model. We would then use an estimation method with uses only the "clean" portion of the variation.
3. We could consider experimentation to break whatever dependence there is between the advertising and sales.

Each of these ideas will be discussed in detail below. Before we embark on a more detailed discussion of these methods, we will relate our discussion of simultaneity or endogeneity to the literature on causal inference for treatment effects.

4.2 The fundamental problem of causal inference

A growing literature (see, for example, Angrist and Pischke, 2009 and Imbens and Rubin, 2014) emphases a particular formulation of the problem of causal inference. Much of this literature re-interprets existing econometric methods in light of this paradigm. The basis for this paradigm of causal inference was originally suggested by Neyman (1990) who conceived of the notion of potential outcomes for a treatment. The notation favored by Imbens and Rubin is as follows. Y represents the outcome random variable. In our case, Y will be sales or some sort of event (like a conversion or click) which is on the way toward a final purchase. We seek to evaluate a treatment, denoted D. For now, consider an binary treatment such as exposure to an ad.[19] We conceive of there being two potential outcomes:

- $Y_i\,(1)$: potential outcome if unit i is exposed to the treatment.
- $Y_i\,(0)$: potential outcome if unit i is not exposed to the treatment.

We would like to estimate the causal effect of the treatment which is defined as

$$\Delta_i = Y_i\,(1) - Y_i\,(0)$$

The fundamental problem of causal inference is that we only see one of two potential outcomes for each unit being treated. That is, we only observe $Y_i\,(1)$ for $D_i = 1$ and $Y_i\,(0)$ for $D_i = 0$. Without further assumptions or information, this statistical

[18] Here we are assuming that within variation in A is exogenous. For example, if promotions or ad campaigns are designed at the quarterly level, then we are assuming that within quarter variation is execution-based and unrelated to within quarter demand shocks. The validity of this assumption would have to be assessed in the same way that any argument for exogeneity is made. However, this exploits institutional arrangements that may well be argued are indeed exogenous.

[19] It is a simple matter to extend potential outcomes framework a more continuous treatment variables such as in causal inference with respect to the effect of price on demand.

problem is un-identified. Note that we have already simplified the problem greatly by assuming a linear model or restricting our analysis to only one "level" of treatment. Even if we simplify the model by assuming a constant treatment effect, $\Delta_i = \Delta \, \forall i$, the causal effect is still not identified.

To see this problem, let's take the mean differences in Y between those who were treated and not treated and express this in terms of potential outcomes.

$$\begin{aligned} E\left[Y_i | D_i = 1\right] - E\left[Y_i | D_i = 0\right] &= E\left[Y_i\left(1\right) | D_i = 1\right] - E\left[Y_i\left(0\right) | D_i = 0\right] \\ &= E\left[Y_i\left(1\right) | D_i = 1\right] - E\left[Y_i\left(0\right) | D_i = 1\right] \\ &\quad + E\left[Y_i\left(0\right) | D_i = 1\right] - E\left[Y_i\left(0\right) | D_i = 0\right] \end{aligned}$$

This equation simply states that what the data identifies is the mean difference in the outcome variable between the treated and untreated and this can be expressed as the sum of two terms.

The first term is the effect on the treated, $[Y_i\left(1\right) | D_i = 1] - E\left[Y_i\left(0\right) | D_i = 1\right]$, and the second term is called the selection bias, $E\left[Y_i\left(0\right) | D_i = 1\right] - E\left[Y_i\left(0\right) | D_i = 0\right]$. Selection bias occurs when the potential outcome for those assigned to the treatment differs in a systematic way from those who are assigned to the "control" or assigned not to be treated. This selection bias is what inspired much of the work of Heckman, Angrist, and Imbens to obtain further information. The classic example of this is the so-called "ability" bias argument in the literature on education. We can't simply compare the wages of college graduates with those who did not graduate for college, because it is likely that college graduates have greater ability even "untreated" with a college education. Those who argue for the "certification" view of higher education are the extreme point of this selection bias – they argue that the only point of education is not those courses in Greek Philosophy but simply the selection bias of finding higher ability individuals.

It is useful to reflect on what sort of situations are likely to have large selection bias in the evaluation of marketing actions. Mass media like TV or print are typically only targeted at a very broad demographic group. For example, advertisers on the Super Bowl are paying a great deal of money to target men aged 25-45. There is year-to-year variation in Super Bowl viewership which in principle would allow us to estimate some sort of regression based model of the effect of exposure to Super Bowl ads. The question is what is the possible selection bias? It is true that the effectiveness of a beer ad on those who view the Super Bowl versus a random consumer may be very different, but, that may not be relevant to the Super Bowl advertiser. The SB advertiser cares more about the effect on the treated, that is the effect of exposure on those in the target audience who view the SB. Are those who choose not to view the SB in year X different from those who view the SB in year Y? Not necessarily, viewership is probably driven by differences in the popularity of the teams in the SB. Thus, if our interest is the effect on the treated Super Bowl fan, there probably is little selection bias (under the assumption that the demand for beer is similar across the national population of SB fans).

However, selection bias is a probably a very serious problem in other situations. Consider a firm like North Face that markets outdoor clothing. This is a highly seasonal industry with two peaks in demand each year: one in the spring as people anticipate summer outdoor activities and another in the late fall as consumers are purchasing holiday gifts. North Face is aware of these peaks in demand and typically schedules much of its promotional and advertising activity to coincide with these peaks in demand. This means we can't simply compare sales in periods of high advertising activity to sales in periods of low as we are confounding the seasonal demand shift with the effect of marketing.

In the example of highly seasonal demand and coordinated marketing, the marketing instruments are still mass or untargeted for the most part (other than demographic and, possible, geographic targeting rules). However, the problem of selection bias can also be created by various forms of behavioral targeting. The premier example of this is the paid search advertising products that generate much of Google Inc.'s profits. Here the ad is triggered by the consumer's search actions. Clearly, we can't compare the subsequent purchases of someone who uses search keywords related to cars with those consumers who were not exposed to paid search ads for cars. There is apt to be a huge selection bias as most of those not exposed to the car keyword search ad are not in the market to purchase a car. Correlational analyses of the impact of paid search ads are apt to show a huge impact that is largely selection bias (see Blake et al., 2015 for analysis of paid search ads for eBay in which they conclude that they have little effect). There is no question that targeting ads based on the preferences of customers as revealed in their behavior is apt to become even more prevalent in the future. This means that, for all the talk of "big data" we are creating more and more data that is not amenable to analysis with our standard bag of statistical tricks.

4.3 Randomized experimentation

The problem with observational data is the potential correlation between "treatment" assignment and the potential outcomes. We have seen that this is likely to be a huge problem for highly targeted forms of marketing activities where the targeting is based on customer preferences. More generally, any situation in which some of the variation in the right hand side variables is correlated with the error term in the sales response equation will make any "regression-style" method inconsistent in estimating the parameters of the causal function. For example, the classical errors-in-variables model results in a correlation between the measured values of the rhs variables and the error term.

In a randomized experiment, the key idea is that assignment to the treatment is random and therefore uncorrelated with any other observable or unobservable variable. In particular, assignment to the treatment is uncorrelated with the potential outcomes. This eliminates the selection bias term.

$$E\left[Y_i\left(0\right)|D_i=1\right]-E\left[Y_i\left(0\right)|D_i=0\right]=0$$

This means that the difference in means between the treated and untreated populations consistently estimates not only the effect on the treated but also the average effect or the effect on the person chosen at random from the population.

However, it is important to understand that when we say person chosen at random from the "population" we are restricting attention to the population of units eligible for assignment in the experiment. Deaton and Cartright (2016) call the set of units eligible for assignment to a treatment cell (including the control cell) the *trial sample*. In many randomized experiments, the trial sample is anything but a random sample of the appropriate population to which we wish to extrapolate the results of the experiment.

Most experiments have a very limited domain. For example, if we randomly assign DMAs in the Northeast portion of the US, our population is only that restricted domain. Most of the classic social experiments in economics have very restricted domains or population to which the results can be extrapolated. Generalizability is the most restrictive aspect of randomized experimentation. Experimentation in marketing applications such as "geo" or DMA based experiments conducted by Google and Facebook start to get at experiments which are generalizable to the relevant population (i.e. all US consumers).

Another key weakness of randomization is that this idea is really a large sample concept. It is of little comfort to the analyst that treatments were randomly assigned if it turns out that randomization "failed" and did not give rise to a random realized sample of treated and untreated units. With a finite N, this is a real possibility. In some sense, all we know is that statements based on randomization only work asymptotically. Deaton and Cartright (2016) make this point as well and point out that only when all other effects actually balance out between the controls and the treated does randomization achieve the desired aim. This only happens in expectation or in infinite size samples. If there are a large number of factors to be "balanced out," then this may require very large N.

A practical limitation to experimentation is that there can be situations in which randomization results in samples with low power to resolve causal effects. This can happen when the effects of the variables being tested are small, the sales response model has low explanatory power, and the sales dependent variable is highly variable. A simple case might be where you are doing an analysis of the effect of an ad using individual data and no other covariates in the sales response model. The standard errors of the causal effect (here just the coefficient on the binary treatment variables) of course are decreasing only at rate $\sqrt{N}$ and increasing in the standard deviation of the error term. If the effects are small, then the standard deviation of the error term is about the same as the standard deviation of sales. Simple power calculations in these situations can easily result in experimental designs with thousands or even tens of thousands of subjects, a point made recently by Lewis and Rao (2014).

Lewis and Rao neglect to say that if there are other explanatory variables (such as price and promotion) included in the model, then even though sales may be highly variable, we still may be able to design experiments with adequate power even with smallish N. If there are explanatory variables included in the response model (in addition to dummy variables corresponding to treatment assignment), then the variance of the error term can be much lower than the variance of dependent variable (sales). In these situations, the power calculations that lead to pessimistic views regarding the number of experimental subjects could change dramatically and the conclusions of Lewis and Rao may not apply. It should be emphasized that this is true even though these additional control variables will be (by construction) orthogonal to the treatment regressors.

While Lewis and Rao's point regarding the difficulties in estimating ad effects is well-taken due to the small size of ad effects, this is not true regarding randomized experimentation on pricing. Marketing researchers have long observed that price and promotional effects are often very large (price elasticities exceeding 3 in absolute value and promotional lifts of over 100 per cent). This means that randomized experiments may succeed in estimating price and promotional effects with far smaller numbers of subjects than for advertising experiments (Dubé and Misra, 2018 constitute an example of recent attempts to use experimentation to optimize pricing).

While randomization might seem the panacea for estimation of causal effects, it has severe limitations for situations in which a large number or a continuum of causal effects are required. For example, consider the situation of two marketing variables and a possibly non-linear causal function: In order to maximize profits for choice of the two variables, we must estimate not just the gradient of this function at some point but the entire function. Clearly, this would require an continuum of experimental conditions. Even if we discretized the values of the variables used in the experiments, the experimental paradigm clearly suffers from the curse of dimensionality as we add variables to the problem. For example, the typical marketing mix model might include at least five or six marketing variables resulting in experiments with hundreds of cells. conjoint as an experiment

4.4 Further limitations of randomized experiments

4.4.1 Compliance in marketing applications of RCTs

The history of randomized experiments dates from agricultural experiments in which the "treatment" consists of various agricultural practices such as fertilization etc. When assigned to an experimental cell, there were no "compliance" issues – whatever treatment was prescribed was administered. However, in both medical and social experimentation applications, compliance can be an important problem. Much of Heckman's work centers around the evaluation of various labor market interventions such as job training programs. Even with complete randomized selection treatment, the government cannot compel US citizens to enroll in job training programs. The best we can do is randomized eligibility for treatment or assignment to treatment. Clearly, there can be selection bias in the acceptance of treatment by those assigned to treatment. Heckman and others have modeled this decision as a rational choice in which people consider the benefits of the job training program as well as their opportunity costs of time. In any event, the "endogeneity" of the actual receipt of treatment means that selection bias would affect the "naive" differences in means (or regression generalizations) approach to effect estimation. There are two ways to tackle this problem: (1) explicit modeling of the decision to accept treatment or (2) use of the treatment assignment as an instrumental variable (see discussion in Angrist and Pischke, 2009, Section 4.4.3). The Instrumental Variables estimator, in this case, is to simply scale the difference in means by the compliance rate.

In some situations such as geographically based ad experiments or store-level pricing experiments, compliance is not an issue in marketing.[20] If we randomize exposure to an ad by DMAs, all consumers in the DMA will have the opportunity to be exposed to the ad. Non-compliance would require consumers in the DMA to deliberately avoid exposure to the ad. Note that this applies to any ad delivery mechanism as long as everyone in the geographic area has the opportunity to become exposed. The only threat to the experimental design is "leakage" in which consumers in nearby DMAs are exposed to the stimulus. Store-level pricing or promotional experiments is another example where compliance is assured.

However, consider what happens when ad experiments are conducted by assigning individuals to an ad exposure treatment. For example, Amazon randomly assigns customers to be exposed to an ad or a "recommendation" while others are assigned not be exposed. However, not all those assigned to the treatment cell will actually be exposed to the ad. The only way to be exposed to the ad is to visit the Amazon site (or mobile app). Those who are more frequent visitors to the Amazon website will have a higher probability of being exposed to the ad than those who are less frequent or intense users. If the response to the ad is correlated to visitor frequency, then the analysis of the RCT for this ad will only reveal the "intent -to-treat" effect and not the average treatment effect on the treated. One way to avoid this problem is to randomize assignment by session and not by user (see Sahni, 2015 for details).

The compliance issue in digital advertising becomes even more complicated due to the "ad campaign" optimization algorithms used by advertising platforms to enhance the effect of ads. For example Facebook has established a randomized experimentation platform (see Gordon and Zettelmeyer, 2017 for analysis of some Facebook ad experiments). The idea is to allow Facebook advertisers to easily implement and analyze randomized experiments for ad campaigns. Facebook users will be randomly assigned to a "control" status in which they will not be exposed to ads. Ads are served by a complicated auction mechanism. For controls, if the ad campaign in question wins an auction for space on a Facebook page that a control is viewing, then the "runner-up" will be served. This insures one-sided compliance – the controls will never have access to the treatment. However, the "experimental" unit will only have an opportunity to view the ad if they visit Facebook as we have already pointed out. However, the problem of compliance is made worse by the Facebook ad campaign "optimization" feature. Any ad campaign is invoked for some span of time (typically 3-5 weeks on Facebook). Data from exposures to the ad early in the campaign is used

[20] There are always problems with implementation of store experiments. That is, the researcher must verify or audit stores to insure treatments are implemented and during the time periods of the experiment. This is not a "compliance" issue as compliance addresses whether or not experimental subjects can decide or influence exposure to the treatment. Consumers are exposed to the properly executed store experiment stimuli. Similarly, ad experiments do not have an explicit compliance problem. There may be a leakage problem across geographies but this is not a compliance problem.

to model who should be exposed to the ad in later stages of the campaign, in addition to whatever targeting criteria are embedded in the campaign. Thus, the probability of exposure to the ad can vary for two Facebook users who visit Facebook with the same frequency and intensity. This means that the Facebook experimentation platform can only estimate an intent to treat effect of the ad campaign and not the effect the treated. Johnson et al. (2017) construct a proxy control ad which they term a "Ghost Ad" which they claim avoids some of the un-intended consequences of the in-campaign optimization methods and can be implemented at lower cost than a more traditional approach in which the control group is exposed to a "dummy" ad or public service announcement. While the "ghost ad" approach appears promising as a way to reduce costs and deal with ad optimization questions, this must be achieved at some power costs which are not yet clear.

4.4.2 The Behrens-Fisher problem

We have explained that the simplest randomized experiment would consist of only control and one treatment cell and the effect of treatment would be estimated by computing the difference in means. Without extensive control covariates, the difference in means is apt to be a very noisy, but consistent (in the number of consumers in the experiment) of the causal effect of treatment. However, Deaton and Cartright (2016) point out that inference with standard methods faces what statisticians have long called the "Behrens-Fisher" problem. If variance of the outcome variable is different between control and treatment groups, then the distribution of the difference in means will be a function of the variance ratio (there is no simple t-distribution anymore). Since the distribution of the test-statistic is dependent on unknown variance parameters, standard finite sample testing methods cannot be used.

Given that there is random assignment to control and treatment groups, any differences in variability must be due to the treatment effect. In a world with heterogeneous treatment effects, we interpret the difference in means between controls and treated as measure the average effect of the treatment. Thus, having the heterogeneous treatments in the error term will create a variance component not present for the controls. For this reason, we might expect that the variance of the treated cell will be higher than the control cell and we are faced with the Behrens-Fisher inference problem. One could argue that Behrens-Fisher problem is apt to be minimal in advertising experiments as the treatment effects are small so that the variance component introduced in the ad exposure treatment cell would be small. However, in experiments related to pricing actions, it is possible that the Behrens-Fisher problem could be very consequential.

Many trained in modern econometrics would claim that the Behrens-Fisher problem could be avoided or "solved" by the use of so-called heteroskedasticity-consistent (White) variance-covariance estimators. This is nothing more than saying that, in large samples, the Behrens-Fisher problem "goes away" in the sense that we can consistently recover the different variances and proceed as though we actually know the variances of the two groups. This can also be seen as a special case of the "cluster" variance problem with only two clusters. Again, heteroskedastic-consistent estimators have long been advocated as a "solution" to the cluster variance problem.

However, it is well known that heteroskedastic consistent variance estimators can have very substantial finite sample biases (see Imbens and Kolesar, 2016 for explicit simulation studies for the case considered here of two clusters). There appears to be no way out of the Behrens-Fisher problem without additional information regarding the relative size of the two variances.

4.5 Other control methods

We have seen that randomization can be used to consistently estimate causal effects (or eliminate selection bias). In Section 5, we will discuss Instrumental Variables approaches. One way of viewing an IV is as a source of "naturally" occurring randomization (IVs) can help solve the fundamental problem of causal inference. Another approach is to add additional covariates to the analysis in hopes of achieving independence of the treatment exposure conditional on these sets of covariates. If we can find covariates that are highly correlated with the unobservables and then add these to the sales response model, then the estimate on the treatment or marketing variables of interest can be "cleaner" or less confounded with selection bias.

4.5.1 Propensity scores

If we have individual level data and are considering a binary treatment such as ad exposure, then conditioning on covariates to achieve approximate independence, simplifies to the use of propensity scores as a covariate. The propensity score[21] is nothing more than the probability that the individual is exposed to the ad as a function of covariates (typically the fitted probability from a logit/probit model of exposure). For example, suppose we want to measure the effectiveness of a YouTube ad for an electronic device. The ad is shown on a YouTube channel whose theme is electronics. Here the selection bias problem can be severe – those exposed to the ad may be pre-disposed to purchase the product. The propensity score method attempts to adjust for these biases by modeling the probability of exposure to the ad based on covariates such as demographics and various "techno-graphics" such as browser type and previous viewing of electronics YouTube channels. The propensity score estimate of the treatment or ad exposure effect would be from a response model that includes the treatment variable as well as the propensity score. Typically, effect sizes are reduced by inclusion of the propensity score in the case of positive selection bias.

Of course, the propensity score method is only as good as the set of co-variates used to form the propensity score. There is no way to test that a propensity score fully adjusts for selection bias other than confirmation via true randomized experimentation. Goodness-of-fit or statistical significance of the propensity score model is

[21] See Imbens and Rubin (2014), Chapter 13, for more details on propensity scores.

re-assuring but not conclusive. There is a long tradition of empirical work in marketing that demonstrates that demographic variables are not predictive of brand choice or brand preference.[22] This implies that propensity score models built on standard demographics are apt to be of little use reducing selection bias and obtaining better causal effect estimates.

Another way of understanding the propensity score method is to think about a "synthetic" control population. That is, for each person who is exposed to the ad, we find a "twin" who is identical (in terms of product preferences and ability to buy) who was not exposed to the ad. The difference in means between the exposed (treatment) group and this synthetic control population should be a cleaner estimate of the causal effect. In terms of propensity scores, those with similar propensity scores are considered "twins." In this same spirit, there is a large literature on "matching" estimators that attempt to construct synthetic controls (cf. Imbens and Rubin, 2014, Chapters 15 and 18). Again, any matching estimator is only as good as the variables used in implementing "matching."

4.5.2 Panel data and selection on unobservables

The problem of selection bias and the barriers to causal inference with observation data can also be interpreted as the problem of "selection on unobservables." Suppose our goal is to learn about the income effects for the demand for a class of goods such as private label goods (see, for example, Dubé et al., 2018). We could take a cross-section of households and examine the correlation between income and demand for private label goods. If we are interested in how the business cycle affects demand for private labels, then we want the true causal income effect. It could be that there is some unobservable household trait (such as pursuit of status) that drives both attainment of higher income as well as lowers the demand for lower quality private label goods. This unobservable would create a spurious negative correlation between household income and private label demand. Thus, we might be suspicious of cross-sectional results unless we can properly control (by inclusions of the appropriate covariate) for the "unobservables" by using proxies for the unobservable or direct measurement of the unobservable.

If we have panel data and we think that there are unobservables that are time invariant, then we can adopt a "fixed effects" style approach which uses only variation within unit over time to estimate causal effects. The only assumption required here is that the unobservables are time invariant. Given that marketing data sets seldom span more than a few years, this time invariance assumption seems eminently reasonable. It should be noted that if the time span increases a host of non-stationarities arise such as the introduction of new products, entry of competitors, etc. In sum, it is not clear that we would want to use a long time series of data without modeling the evolution of the industry we are studying. Of course as pointed out in Section 3.1 above, the fixed effects approach only works with linear models.

[22] See, for example, Fennell et al. (2003).

Consider the example of estimating the effect of a Super Bowl ad. Aggregate time series data may have insufficient variation in exposure to estimate ad effects. Pure cross-sectional variation confounds regional preferences for products with true useful variation in ad exposure. Panel data, on the other hand, might be very useful to isolate Super Bowl ad effects. Klapper and Hartmann (2018) exploit a short panel of six years of data across about 50 different DMAs to estimate effects of CPG ads. They find that there is a great deal of year-to-year variation in the same DMA in SB viewership. It is hard to believe that preferences for these products vary from year to year in a way that is correlated with the popularity of the SB broadcast. Far more plausible, is that this variation depends on the extent to which the SB is judged to be interesting at the DMA level. This could be because a home team is in the SB or it could just be due to the national or regional reputation of the contestants. Klapper and Hartmann estimate linear models with Brand-DMA fixed effects (intercepts) and find a large and statistically significant effect of SB ads by beer and soft drink advertisers. This is quite an achievement given the cynicism in the empirical advertising literature about ability to have sufficient power to measure advertising effects without experimental variation. Many, if not most, of the marketing mix models estimated today are estimated on aggregate or regional time series data.

The success of Klapper and Hartmann in estimating effects using more disaggregate panel data is an important source of hope for the future of marketing analytics. It is well known that the idea of using fixed effects or unit-specific intercepts does not generalize to non-linear models. If we want to optimize the selection of marketing variables then we will have to use more computationally intensive hierarchical modeling approaches to allowing response parameters to vary over cross-sectional units.

Advocates of the fixed effects approach argue that the use of fixed effects does not require any distributional assumptions nor the assumption that unit parameters are independent of the rhs variables. Given that it is possible to construct hierarchical models with a general distributional form as well as to allow unit characteristics to affect these distributions, it seems the time is ripe to move to hierarchical approaches for marketing analytics with non-linear response models.

4.5.3 Geographically based controls

In the area of advertising research, some have exploited a control strategy that depends on some of the historical institutional artifacts in purchase of TV advertising. In the day in which local TV stations were limited by the reach of their signal strength, it made sense to purchase local TV advertising on the basis of a definition of media market that include the boundaries of the TV signal. There are 204 such "Designated Market Areas" in the US. Local TV advertising is purchased by DMA. This means that there are "pairs" of counties on opposite sides of a DMA boundary, one of which receives the ad exposure while the does not. Geographical proximity also serves as a kind of "control" for other factors influencing ad exposure or ad response. Shapiro (2018) uses this strategy to estimate the effect of direct-to-consumer ads for various anti-depressant drugs. Instead of using all variation in viewership across counties and across time, Shapiro limits variation to a relatively small number of “paired” DMAs. Differences in viewership between these two “bordering” DMA is used to identify ad effects. Shapiro finds only small differences between ad effects estimated with his “border strategy” vs not. However, this idea of exploiting institutional artifacts in the way advertising is purchased is a general idea which might be applied in other ways. However, the demise of broadcast or even subscription TV in favor of streaming will likely render this particular “border strategy” increasingly irrelevant. But the idea of exploiting the discreteness in the allocation or exposure rule used by firms in a case of what is called a regression discontinuity design discussed below.

4.6 Regression discontinuity designs

Many promotional activities in marketing are conducted via some sort of threshold rule or discretized into various “buckets.” For example, consider the loyalty program of a gambling casino. The coin of the realm in this industry is the expected win for each customer which is simply a function of the volume of gambling and type of game. The typical loyalty program encourages customers to gamble more and come back to the casino by establishing a set of thresholds. As customers increase their expected win, they “move” from one tier or “bucket” in this program to the next. In the higher tiers, the customer receives various benefits like complementary rooms or meals. The key is that there is a discrete jump in benefits by design of the loyalty program. On the other hand, it is hard to believe that the response function of the customer to the level of complementary benefits is non-smooth or discontinuous. Thus, it would seem that we can “select” on the observables to compare those customers whose volume of play is just on either side of each discontinuity in the reward program.

As Hartmann et al. (2011) point out, as long as the customer is not aware of the threshold or the benefits from “selecting in” or moving to the next tier are small relative to the cost of greater play, this constitutes a valid Regression Discontinuity (RD) design. Other examples in marketing include direct mail activity (those who receive offers and or contact are a discontinuous function of past order history) and geographic targeting (it is unlikely people will move to get the better offer). But, if consumers are aware that are large promotions or rebates for a product and they can change their behavior (such as purchase timing), then an RD approach is likely to be invalid.

Regression discontinuity analysis has received a great deal of attention in economics as well (see Imbens and Lemieux, 2008). The key assumption is that the response function is continuous in the neighborhood of the discontinuity in the assignment of the treatment. There are both parametric and non-parametric forms of analysis, reflecting the importance of estimating the response function without bias that would aversely affect the RD estimates. Parametric approaches require a great deal of flexibility which may compromise power, while non-parametric methods rest on the promise to narrow the window of responses used in the vicinity of the threshold (s) as the sample size increases. This is not much comfort to the analyst with one finite sample. Non-parametric RD methods are profligate with data as, ultimately, most of the data is not used in forming treatment effect estimates.

RD designs result in only local estimates of the derivative of the response function. For this reason, unless the ultimate treatment is really discrete, RD designs do not offer a solution to the marketing analytics problem of optimization. RD designs may be helpful to corroborate the estimates based on response models fit to the entire dataset (the RD estimate and the derivative the response function at the threshold should be comparable).

4.7 Randomized experimentation vs. control strategies

Recent work in marketing compares inferences based on various control strategies (including propensity scores and various "matching" or synthetic control approaches with the results of large scale randomized experiments performed at Facebook. Gordon and Zettelmeyer (2017) find that advertising effects estimated from observational methods do not agree very closely with those based on randomized experimentation in the context of ad campaigns evaluated on the Facebook ad platform. If various control strategies are inadequate, then we might expect that ad effects estimated by observational data would be larger than those estimates that are based on randomized experimentation (at least up to sampling variation). Gordon and Zettelmeyer do not find any consistent pattern of this sort. They find estimates based on observation data to be, in some cases, smaller than those based on experimentation with non-overlapping confidence intervals. This result is difficult to understand and implies that there are important unobservables which are positively related to ad exposure and negatively related to ad effects. However, it is pretty clear that the jury is out on the efficacy of observational methods as Eckles and Bakshy (2017) find that observational methods (propensity scores) produce ad effect estimates which are close to those obtained from randomized experimentation in a similar context involving estimation of peer effects with Facebook data. It is possible that Facebook ad campaign "optimization" may make that comparison between the observational data-based effect estimates and the randomized trial results less direct than Gordon and Zettelmeyer imply.

4.8 Moving beyond average effects

We live in a world of heterogeneous treatment effects in which each consumer, for example, has a different response to the same ad campaign. In the past, the emphasis in economics is on estimating some sort of average treatment effects which is thought to be adequate for policy evaluation. Clearly, the distributional effects of policies are also important and, while the randomized experiment does identify the average treatment effect with minimal assumptions, randomized experimentation does not identify the distribution of effects without imposing additional assumptions.

In marketing applications, heterogeneity assumes even greater importance than in economic policy evaluation. This is because the policies in marketing are not applied uniformly to a subset of consumers but, rather, include the possibility of targeting policies based on individual treatment effects. A classic example of this problem is the problem in direct marketing of to whom a catalog or offer should be sent to from the very large set of customers whose characteristics are summarized in the "house" file of past order and response behavior. Classically, direct marketers built models that are standard marketing response models in which order response to a catalog or offer is modeled as a function of the huge set of variables that might be constructed using the house data file. This raises two inference problems. First, the model-builder must have a way of selecting from a set of variables that is may be even larger than the number of observations. Second, the model-builder should recognize that there may be unobservables that create the classic selection bias problem. The selection bias problem can be particularly severe when the variables used as "controls" are simply summaries of past response behavior as must be, by construction, from house file data.

How then does randomization help the model-builder? If there is a data-set where exposure to the marketing action is purely random, then there are no selection bias problems and there is nothing wrong with using regression-like methods to estimate or predict response to the new offering (i.e. "optimal targeting"). The problem then becomes more of a standard non-parametric modeling problem of selecting the most efficient summaries of the past behavior to be included as controls in the response model. Hitsch and Misra (2018) compare a number of different methods to estimate heterogeneous treatment effects based on a randomized trial and evaluate various estimators with respect to their potential profitability.

5 Instruments and endogeneity[23]

The problem of causal inference and the potential outcomes framework has recently assumed greater importance in the economics literature but that is not to say that the problem of causal inference has only recently been addressed. The original concern of the Cowles commission was to obtain consistent estimates of "structural" parameters using only observation data as well as the recognition that methods that assume right-hand-side variables are exogenous may not be appropriate in many applications. In most applications, the "selection" bias or "selection on unobservables" interpretation is appropriate and econometricians have dubbed this the "endogeneity" problem.

One basic approach to dealing with this problem is to find some way of partitioning the variation in the right-hand-side variable so that some of the variation can be viewed as "though random." This involves selection of an instrument. In this section, we provide a detailed discussion of the instrumental variables approach.

[23] This section was adapted in large part from Rossi (2014b).

As we have indicated, Instrumental Variable (IV) methods do not use all of the variation in the data to identify causal effects, but instead partition the variation into that which can be regarded as “clean” or as though generated via experimental methods and that which is “contaminated” and could result in endogeneity bias. “Endogeneity bias” is almost always defined as the asymptotic bias for an estimator which uses all of the variation in the data. IV methods are only asymptotically unbiased if the instruments are valid instruments. Validity is an unverifiable assumption. Even if valid, IV estimators can have poor sampling properties including fat tails, high RMSE, and bias. While most empirical researchers may recall that the validity assumption is important from their econometrics training, the poor sampling properties of IV estimators are not well appreciated.

Careful empirical researchers are aware of some of these limitations of IV methods and, therefore, sometimes view the IV method as a form of sensitivity analysis. That is, estimates of causal effects using standard regression methods are compared with estimates based on IV procedures. If the estimates are not appreciably different, then some conclude that endogeneity bias is not a problem. While this procedure is certainly more sensible than abandoning regression methods altogether, it is based on the implicit assumption that the IV method uses valid instruments. If the instruments are not valid, then the differences between standard regression style estimates and IV estimates don’t have any bearing on the existence or extent of endogeneity bias.

Closely related to the problem of endogeneity bias is the problem of omitted variables in cross-sectional analyses or pooled analyses of panel data. Many contend that there may exist unobservable variables that a set of control variables, no matter how exhaustive, cannot control for. For this reason, researchers often use a Fixed Effects (hereafter FE) approach in which cross-sectional unit specific intercepts are included in the analysis. In a FE approach, the slope coefficients on variables of interest are only identified using only “within” variation in the data. Cross-sectional variation is thrown out. Advocates for the FE approach argue that, in contrast to IV methods, the FE approach does not require any further assumptions than those already used by the standard linear regression analysis. The validity of the FE approach depends critically on the assumption of a linear model and the lack of measurement error in the independent variables.[24] If there is measurement error in the independent variables, then the FE approach will generally magnify the errors-in-the-variables bias.

5.1 The omitted variables interpretation of “endogeneity” bias

In marketing applications, the omitted variable interpretation of endogeneity bias provides a very useful intuition. In this section, we will briefly review the standard omitted variables analysis and relate this to endogeneity bias. For those familiar with

[24] If lagged dependent variables are included in the model, then the standard fixed effects approach is invalid, see Narayanan and Nair (2013); Nickell (1981).

the omitted variables problem, this section will simply serve to set notation and a very brief review (see also treatments in Section 4.3 of Woolridge, 2010 or Section 3.2.2 of Angrist and Pischke, 2009). Consider a linear model with one independent variable (note: the intercept is removed for notational simplicity).

$$y_i = \beta x_i + \varepsilon_i \tag{40}$$

The least squares estimator from a regression of y on x will consistently estimate parameters of the conditional expectation of y given x under the restriction that the conditional expectation is linear in x. However, the least squares estimator will converge to β only if $\mathbb{E}[\varepsilon|x] = 0$ (or $\operatorname{cov}(x, \varepsilon) = 0$).

$$\operatorname{plim} \frac{x'y}{x'x} = \beta + \operatorname{plim} \frac{\frac{x'\varepsilon}{N}}{\frac{x'x}{N}} = \beta + \operatorname{plim} \left(\frac{x'x}{N}\right)^{-1} \operatorname{plim} \frac{x'\varepsilon}{N} = \beta + Q \times \operatorname{cov}(x, \varepsilon)$$

Here $Q^{-1} = \operatorname{plim} \frac{x'x}{N}$. Thus, least squares will consistently estimate the "structural" parameter β only if (40) can be considered a valid regression equation (with an error term that has a conditional expectation of zero). If $\mathbb{E}[\varepsilon|x] \neq 0$, then least squares will not be a consistent estimator of β. This situation can arise if there is an omitted variable in the equation. Suppose there exists another variable, w, which belongs in the equation in the sense that the multiple regression of y on x and w is a valid equation.

$$\begin{aligned} &y_i = \beta x_i + \gamma w_i + \varepsilon_i \\ &E[\varepsilon|x, w] = 0 \end{aligned}$$

The least squares regression of y on x alone will consistently recover the parameters of the conditional expectation of y given x which will not necessarily be β

$$\mathbb{E}\left[y|x\right] = \beta x + \mathbb{E}\left[\gamma w + \varepsilon|x\right] = \beta x + \gamma \mathbb{E}[w|x] = \beta x + \gamma \pi x = \delta x$$

Here π is the coefficient of w in the conditional expectation of $w|x$. If $\pi \neq 0$, then the least squares estimator will not consistently recover β (sometimes called the structural parameter) but instead will recover δ. The intuition is that, in the simple regression of y on x, least squares estimates the effect of x without controlling for w. This estimate confounds two effects: (1) the direct effect of x (β) and (2) the indirect effect of x ($\gamma\pi$). The indirect effect (which is non-zero whenever x and w are correlated) also has a very straightforward interpretation: for each unit change in x, w will change by π units and this will, in turn, change y (on average) by γ units.

In situations where $\delta \neq \beta$, there is an omitted variable bias. The solution, which is feasible only if w is observable, is to run the multiple regression of y on x and w. Of course, the multiple regression does not use all of the variation in x to estimate the multiple regression coefficient – only that part of the variation in x which is uncorrelated with w. Thus, we can see that a multiple regression method is more demanding of the data in the sense that only part of the variation of x is used. In a true randomized experiment, there is no omitted variable bias because the values of x are assigned randomly and, therefore, are uncorrelated by definition with any other variable (observable or not). In the case of the randomized experiment, the only motivation for bringing in other covariates is to reduce the size of the residual standard error which can improve the precision of estimation. However, if the simple regression model produces statistically significant results, there is no reason for adding covariates.

The standard recommendation for limiting omitted variable bias is to include as many "control" variables or covariates as possible. For example, suppose that we observe demand for a given product across a cross-section of markets. If we regress quantity demanded on price across these markets, a possible omitted variable bias is that there are some markets where there is a higher demand for the product than others and that price is set higher in those markets with higher demand. This is a form of omitted variable bias where the omitted variable is some sort of indicator of market demand conditions. To avoid omitted variable bias, the careful researcher would add covariates (such as average income or wealth measures) which seek to "control" or proxy for the omitted demand variable and use a multiple regression. There is a concern that these control or proxy variables are only imperfectly related to true underlying demand conditions which are never perfectly predicted or "observable."

5.2 Endogeneity and omitted variable bias

Most applied empirical researchers will identify "endogeneity bias" as arising from correlation between independent variables and error terms in a regression. This is to describe a cold by its symptoms. To develop a strong intuitive understanding, it is helpful to give an omitted variables interpretation. Assume that there is an unobservable variable, ν, which is related to both y and x.

$$y_i = \beta x_i + \alpha_y \nu_i + \varepsilon_{y,i} \tag{41}$$

$$x_i = \alpha_x \nu_i + \varepsilon_{x,i} \tag{42}$$

Here both $\varepsilon_x, \varepsilon_y$ have 0 conditional mean given x and ν and are assumed to be independent. In our example of demand in a cross-section of markets, ν represents some unknown demand shifter variable that allows some markets to have a higher level of demand for any given price than others. Thus, ν is an omitted variable and has the potential to cause omitted variable bias if ν is correlated with x. The model listed in (42) builds this correlation in by constructing x from ν and another exogenous error term. The idea here is that prices are set partially as a function of this underlying demand characteristic which is observable to the firm but not observable to the researcher. In the regression of y on x, the error term is now $\alpha \nu_i + \varepsilon_{y,i}$ which is correlated with x. This form of omitted variable bias is called endogeneity bias. The term "endogeneity" comes from the notion that x is no longer determined "exogenously" (as if via an experiment) but is jointly determined along with y.

We can easily calculate the endogeneity bias by taking conditional expectations (or linear projections) of y given x.

$$\begin{aligned}\mathbb{E}\left[y|x\right] &= \beta x + \mathbb{E}\left[\alpha_y v + \varepsilon_y|x\right] \\ &= \beta x + \alpha_y \alpha_x \left(\frac{\sigma_v^2}{\alpha_x^2 \sigma_v^2 + \sigma_{\varepsilon_x}^2}\right) x \end{aligned} \tag{43}$$

The ratio $\alpha_x \left(\frac{\sigma_v^2}{\alpha_x^2 \sigma_v^2 + \sigma_{\varepsilon_x}^2}\right)$ is simply the regression coefficient from a regression of the composite error term (including the unobservable) on x. The endogeneity bias is thus the coefficient on x in (43). Whenever the unobservable has variation which comprises a large fraction of the total variation in x, and has the unobservable has a large effect on y, the endogeneity bias will be large.

If we go back to our example of price endogeneity in a cross-section of markets, this would mean that the demand differences across markets would have to be large relative to other influences that shift price. In addition, the influence of the unobservable demand shifter on demand (y) must be large.

5.3 IV methods

As we have seen the "endogeneity" problem is best understood as arising from an unobservable that is correlated both with the error in the "structural" equation and one or more of the right side variables in this equation. Regression methods were originally designed for experimental data where the x variable was chosen by the investigator as part of the experimental design. For observational data, this is not true and there is always the danger that there exists some unobservable variable which has been omitted from the structural equation. This makes a concern for endogeneity a generic criticism which can always be applied.

The ideal solution to the endogeneity problem would be to conduct an experiment in which the x variable is, by construction, uncorrelated via randomization with any unobservable. Short of this ideal, researchers opt to partition the variation in x variable[25] into two parts: (1) variation that is "exogenous" or unrelated to the structural equation error term and (2) variation that might be correlated with the error term. Of course, this partition always exists; the only question is whether or not the partition can be accessed by the use of observable variables. If such an observable variable exists, then it must be correlated with x variable but it must not enter the structural equation. Such a variable is termed an "instrumental variable." The idea of an instrument is that this variable moves around x but does not affect y in direct way, only indirectly via x. Of course, there can be many instrumental variables.

[25] For simplicity, I will consider the case of only one right hand side endogenous variable. There is no additional insight gained from the multiple rhs variable case and the great majority of applied work only considers endogeneity in one variable.

5.3.1 The linear case

The case of a linear structural equation and linear instrumental variable model provides the intuition for the general case and also includes many of empirical applications of IV methods. However, it should be noted that due the widespread use of choice models in marketing applications, there is a much higher incidence of the use of nonlinear models. We consider nonlinear choice models in Section 5.7. (44) and (45) constitute the linear IV model.

$$y = \beta x + \gamma' w + \varepsilon_y \tag{44}$$

$$x = \delta' z + \varepsilon_x \tag{45}$$

(44) is the structural equation. The focus is on estimation of the "structural" parameter, β, avoiding endogeneity bias. There is the possibility there are other variables in the "structural" equation which are exogenous in the sense that we assume that $\mathbb{E}\left[\varepsilon_y|w\right] = 0$. If these variables are comprehensive enough, meaning that almost all of the variation in the unobservable that is at the heart of the endogeneity problem can be explained by w, then the "endogeneity" problem ceases to be an issue. The regression methods will only use the variation in x that is independent of w and, under the assumption that the w controls are complete, then there should be no endogeneity problem. For the purpose of this exposition, we will assume that $\mathbb{E}\left[\varepsilon_y|x, w\right] = f\left(x\right) \neq 0$, or that we still have an endogeneity problem.

The second equation (45) is a just a linear projection of x on the set of instrumental variables and is often called the instruments or "first-stage" equation. In a linear model, the statement, $\mathbb{E}\left[\varepsilon_y|x, w\right] \neq 0$, is equivalent to $\text{corr}\left(\varepsilon_x, \varepsilon_y\right) \neq 0$. In the omitted variable interpretation, this correlation in the equation errors is brought about by a common unobservable. As the correlation between the errors increases, the "endogeneity bias" becomes more severe.

The critical assumption in the linear IV model is that the instrumental variables, z, do not enter into the structural equation. This means that the instruments only have an indirect effect on y via movement in x but no direct effect. This is restriction is often called the *exclusion* restriction or sometimes the over-identification restriction. Unfortunately, there is no way to "test" the exclusion restriction because the model in which the z variables enters both equations is not identified.[26]

5.3.2 Method of moments and 2SLS

There are a number of ways to motivate inference for the linear IV model in (44)-(45). The most popular is the method of moments approach. For the sake of brevity and notational simplicity, consider the linear IV model with only one instrument and no other "exogenous" variables in the structural equation. The method of moments estimator exploits the assumption that z (now just a scalar r.v.) is uncorrelated or orthogonal to the structural equation error. This is called a moment condition and

[26] The so-called "Hausman" test requires at least one instrument for which the investigator must assume the exclusion restriction holds.

involves an assumption about population or data generating model that $\mathbb{E}\left[\varepsilon_y z\right] = 0$. The method of moments principle defines an estimator by minimizing the discrepancy between the population and sample moments.

$$\hat{\beta}_{MM} = \underset{\beta}{\operatorname{argmin}} \left\| \mathbb{E}\left[\varepsilon_y z\right] - (y - \beta x)' z \right\| = \frac{z' y}{z' x} \tag{46}$$

Here y, x, z are $N \times 1$ vectors of the observations. It is easy to see that this estimator is consistent (because we assume $\mathbb{E}\left[\varepsilon_y z\right] = 0 = \operatorname{plim}\left(\frac{z'\varepsilon_y}{N}\right)$) and asymptotically normal. If the structural equation errors are uncorrelated and homoskedastic, it can be shown (see, for example, Hayashi, 2000, Section 3.8) that the particular method of moments estimator in (46) is the optimal Generalized Method of Moments Estimator. If the structural equation errors are conditionally heteroskedastic and/or autocorrelated, then the estimator above is no longer optimal and can be improved upon. It should be emphasized that when econometricians say that an estimator is optimal, this only means that the estimator has an asymptotic distribution with variance not exceeding that of any other estimator. This does not mean that, in finite samples, the method of moments estimator has better sampling properties than another other estimator. In particular, even estimators with asymptotic bias such as least squares can have lower mean-squared error than IV estimators.

Another way of motivating the IV estimator for the simple linear IV model is the principle of Two Stage Least Squares (2SLS). The idea of Two Stage Least Squares is much the same as how it is possible to perform a multiple regression via a sequence of simple regressions. The "problem" with the least squares estimator is that some of the variation in x is not exogenous and correlated with the structural equation error. The instrumental variables can be used to purge x of any correlation with the error term. The fitted values from a regression of x on z will be uncorrelated with the structural equation errors. Thus, we can use the fitted values from a "first-stage" regression of x on z and regress y on the fitted values from this first-stage (this is the second-stage regression).

$$x = \hat{x} + e_x = \hat{\delta} z + e_x \tag{47}$$

$$y = \hat{\beta}_{2SLS}\hat{x} + e_y \tag{48}$$

This procedure yields the identical estimator as the MM estimator in (46).

If there are more than one instrument, more than one rhs endogenous variable, or if we include a matrix of exogenous variables in the structural equation, then both procedures generalize but the principle of utilizing the assumption that there exists a valid set of instruments and that one should only use that portion of the rhs endogenous variables can is accounted for by the instruments remains the same.

5.4 Control functions as a general approach

A very useful way of viewing the 2SLS estimator is as a special case of the "control function" approach to obtaining an IV estimator. The control function interpretation of 2SLS comes from the fact that the multiple regression coefficient on x is estimated using only that portion of the variation of x which is uncorrelated with the other variables in the equation. If we put a regressor in the structural equation which contains only that part of x which is potentially correlated with ε_y, then the multiple regression estimator would be a valid IV estimator. In fact, the 2SLS estimator can also be obtained by regressing y on x as well as the residual from the first-stage IV regression.

$$y = \hat{\beta}_{TSLS} x + c e_x \tag{49}$$

e_x is the residual from (47).

Petrin and Train (2010) observe that the same idea can be applied to "control" for or eliminate (at least, asymptotically) endogeneity bias in a demand model with a potentially endogenous variable. For example, the control function approach can work even if the demand model is a nonlinear model such as a choice model. If x is a choice characteristic that might be considered potentially endogenous, then one can construct "control" functions from valid instruments and achieve the effects of an IV estimator simply by adding these control functions to the nonlinear model. Since, in non-linear models, the key assumption is not a zero correlation but conditional independence, it is necessary to not just project x on a linear function of the instruments, but to estimate the conditional mean function, $\mathbb{E}[x|z] = f(z)$. The conditional mean function is of unspecified form and this means that we need to choose functions of the instruments that can approximate any smooth function. Typically, polynomials in the instruments of high order should be sufficient. The residual, $e = x - \hat{f}$, is created and can be interpreted as that portion of x which is independent of the instruments. The controls required to be included in the nonlinear model must also allow for arbitrary flexibility in the way in which the residual is entered. Again, polynomials in the residual (or any valid set of basis functions) should work, at least for large enough samples, if we allow the polynomial order to increase with the sample size.

The control function approach has a lot of appeal for applied workers as all we have to do is a first stage linear regression on polynomials in the instruments and simply add polynomials in the residual from this first stage to the nonlinear model. For linear index models like choice models, this simply means that we can do one auxiliary regression and I can use any standard method to fit the choice model, but with constructed independent variables. The ease of use of the control function approach makes it convenient for checking to see whether an instrumental variables analysis produces estimates that are much different. However, inference in the control function approach requires additional computations as the standard errors produced by the standard non-linear models software will be incorrect as they don't take into account that some of the variables are "constructed." It is not clear that from an inference point of view that the control function approach offers any advantages over using the general GMM method which computes valid asymptotic standard errors. The control function approach has a number of assumptions required to show consistency. However, there is some evidence that it will closely approximate the IV solution in the choice model situation.

5.5 Sampling distributions

In the OLS estimator (conditional on the X matrix) is a linear estimator with a sampling distribution derived from the distribution of $\hat{\beta}_{OLS} - \beta = \left(X'X\right)^{-1} X'\varepsilon$. If the errors terms are homoskedastic and normal, then the finite sample distribution of the OLS sampling error is also normal. However, all IV estimators are fundamentally non-linear functions of the data. For example, the simple Method of Moments estimator (46) is a nonlinear function of the random variables. The proper way of viewing the linear IV problems is that, given a matrix of instruments, Z, the model provides the joint distribution of both y and x. Since x is involved non-linearly, via the term $\left(z'x\right)^{-1}$, we cannot provide an analytical expression for the finite sample distribution of the IV estimator even if we make assumptions regarding the distribution of the error terms in the linear IV model (44)-(45). The sampling distribution of the IV estimator is approximated by asymptotic methods. This is done by normalizing by $\sqrt{N}$ and applying a Central Limit Theorem.

$$\sqrt{N}\left(\hat{\beta}_{MM} - \beta\right) = \left(\frac{z'x}{N}\right)^{-1} \sqrt{N}\frac{z'\varepsilon_y}{N} \tag{50}$$

As N approaches infinity, the denominator of the MM estimator, $\frac{z'x}{N}$, converges to a constant by the Law of Large Numbers. The asymptotic distribution is entirely driven by the numerator which has been expressed as $\sqrt{N}$ times a weighted sum of the error terms in the structural equation. The asymptotic distribution is then derived by applying a Central Limit Theorem to this average. Depending on whether or not the error terms are conditional heteroskedastic or autocorrelated (in the case of time series data) a different CLT is used. However, the basic asymptotic normality results are derived by assuming that the sample is large enough to that we can simply ignore the contribution of the denominator to the sampling distribution. While asymptotics greatly simplifies the derivation of a sampling distribution, there is very good reason to believe that this standard method of deriving the asymptotic distribution is apt to be highly inaccurate under the conditions in which the IV estimator is often applied.

The finite sampling distribution can deviate from the asymptotic approximation in two important respects: (1) there can be substantial bias in the sampling distribution of the IV estimator even if the model assumptions hold and (2) the asymptotic approximation can be very poor and can dramatically understate the true sampling variability in the estimator. The simple Method of Moments estimator is a ratio of a weighted average of y to the weighted average of x.

$$\hat{\beta}_{MM} = \frac{\frac{z'y}{N}}{\frac{z'x}{N}}$$

The distribution of a ratio of random variables is very different from the distribution of a linear combination of random variables (the distribution of OLS). Even if the error terms in the linear IV model are homoskedastic and normal, then distribution of

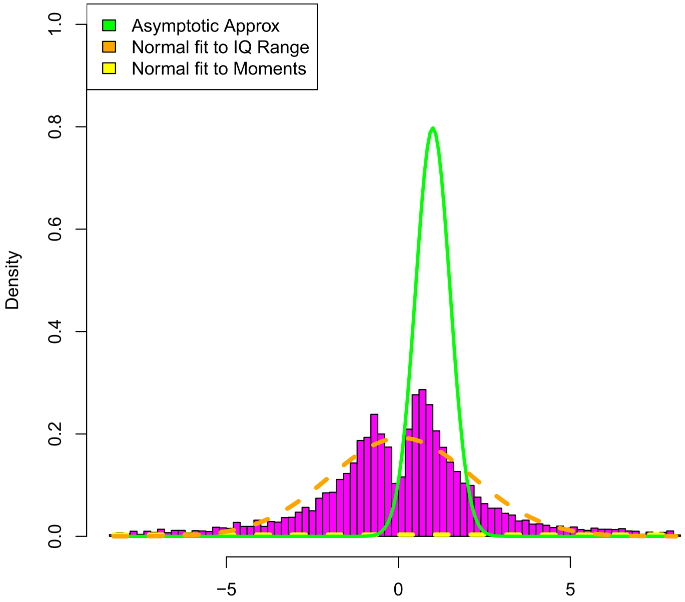

FIGURE 1

Distribution of a ratio of normals.

the Method of Moments IV estimator is non-normal. The denominator is the sample covariance between z and x. If this sample covariance is small, then the ratio can assume large positive and negative values. More precisely, if the distribution of the denominator puts appreciable mass near zero, then the distribution of the ratio will have extremely fat tails. The asymptotic distribution is using a normal distribution to approximate a distribution which has much fatter tails than the normal distribution. This means that the normal asymptotic approximation can dramatically understate the true sampling variability.

To illustrate how ratios of normals can fail to have a normal distribution. Consider the distribution of a ratio of an $N(1, .5)$ to an $N(.1, 1)$ random variable.[27] The distribution is shown by the magenta histogram in Fig. 1 and is revealed to be bimodal with the positive mode having slightly more mass. This distribution exhibits massive outliers and the figure only shows the histogram of the data trimmed to remove the top and bottom 1 per cent of the observations. The thick left and right tails are generated by draws from the denominator normal distribution which are close to the origin. The standard asymptotic approximation to the distribution of IV estimators simply ignores the denominator which is supposed to converge to a constant. The asymptotic approximation is shown by the green density curve in the figure. Clearly, this is a poor approximation that ignores the other mode and under-estimates variability. The

[27] Here the second argument in the normal distribution is the standard deviation.

dashed light yellow line in the figure represents a normal approximation based on the actual sample moments of the ratio of normals. The fact that this approximation is so spread-out is another way of emphasizing that the ratio of normals has very fat tails. The only reasonable normal approximation is shown by the medium dark orange curve which is fit to the observed InterQuartile range. Even this approximation misses the bi-modality of the actual distribution. Of course, the approximation based on the IQ range is not available via asymptotic calculations.

The degree to which the ratio of normals can be well-approximated by a normal distribution depends on both the location and spread of the distribution. Obviously, if the denominator is tightly distributed around a non-zero value, then the normal approximation can be highly accurate. The intuition that we have established is when the denominator has a spread-out distribution and/or places mass near zero, then the standard asymptotic approximation will fail for IV estimators. This can happen into two conditions: (1) in small samples and (2) where the instruments are "weak" in the sense they explain only a small portion of the variation in x. Both cases are really about lack of information. The sampling distributions of IV estimators become very spread out with fat tails when there is little information about the true causal effect in the data. "Information" should be properly measured by total covariance of the instruments with x. This total covariation can be "small" even in what appear to be "large" samples when instruments have only weak explanatory power. In next section, we will explore what are the boundaries of the "weak" instrument problem.

5.6 Instrument validity

One point that is absent from the econometrics literature is that the sampling distribution of IV *estimators* are only considered *conditional* on the validity of the instruments. This is an untestable assumption which certainly is violated in many datasets. This form of mis-specification is much more troubling than other forms of model mis-specification such as non-normality of the error terms, conditional heteroskedasticity, or non-linearity. For each of these mis-specification problems, we have tests for mis-specification and alternative estimators. There are also methods to provide inference (i.e. standard errors and confidence intervals) which are robust to model mis-specification for conditional heteroskedastic, auto-correlated, and non-normal errors. There are no methods which are robust to the use of invalid instruments. To illustrate this point, consider the sampling distribution of an IV estimator based on an invalid instrument. We simulate data from the following model.

$$\begin{aligned} y &= -2x - z + \varepsilon_y \\ x &= 2z + \varepsilon_x \\ \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} &\sim N\left(0, \begin{pmatrix} 1 & .25 \\ .25 & 1 \end{pmatrix}\right); \quad z_i \sim \text{Unif}(0,1) \end{aligned} \tag{51}$$

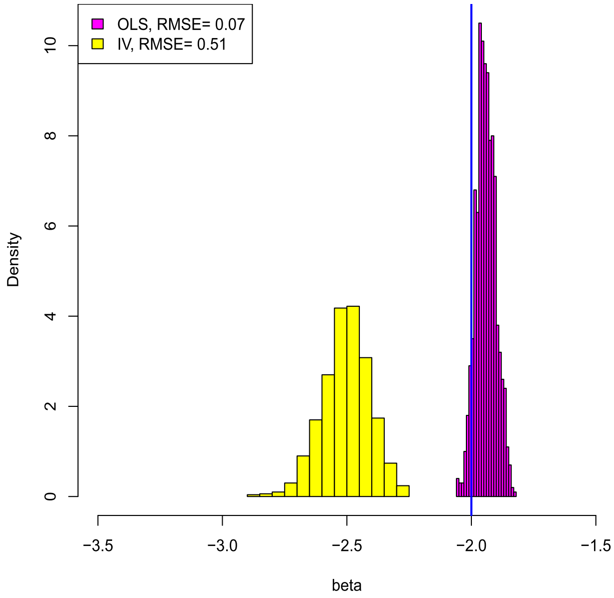

FIGURE 2

Sampling distributions of estimators with invalid instruments.

This is a situation with a relatively strong instrument (the population R-squared for the regression of x on z is about .25). Here $N = 500$ which is a large sample in many cross-sectional contexts. The instrument is invalid but with a smaller direct effect, -1, than an indirect effect, -4. Moreover, the structural parameter is also larger than the direct effect. Fig. 2 shows the sampling distribution of the method of moments estimator and the standard OLS estimator. Both estimators are biased and inconsistent. Moreover, the IV estimator has inferior sampling properties with a root mean-squared-error of more than seven times the OLS estimator. Since we can't know if the instruments are valid, the argument that the IV estimator should be preferred because it is consistent conditional on validity is not persuasive.

Conley et al. (2012) consider the problem of validity of instruments and use the term "plausibly exogenous." That is to say, except for true random variation, it is impossible to prove that an instrument is valid. In most situations, the best that can be said is that the instrument is approximately valid. Conley et al. (2012) define this as an instrument which does not exactly satisfy an exclusion restriction (i.e. the assumption of no direct effect on the response variable) but that the instrument has a small direct effect relative to the indirect effect. From both sampling and Bayesian points of view, Conley et al. (2012) argue that a sensitivity analysis with respect to the exclusion restriction can be useful. For example, if minor (i.e. small) violations of the exclusion restriction do not fundamentally change inferences regarding the key effects, then Conley et al. (2012) consider the analysis robust to violations of instrument validity within that range. For marketing applications, this seems to be a very useful framework. We do not expect any instrument to exactly satisfy the exclusion restriction but we might expect the instrument to be "plausibly exogeneous" in the sense of a small violation. The robustness or sensitivity checks developed by Conley et al. (2012) help assess if the findings are critically sensitive to violations of exogeneity in a plausible range. This provides a useful way of discussing and evaluating the instrument validity issue. This was absent in the econometrics literature and is of great relevance to researchers in marketing who rarely have instruments for which there are air-tight exogeneity arguments.

5.7 The weak instruments problem

5.7.1 Linear models

Not only are instruments potentially invalid, there is a serious problem when instruments are "weak" or only explain a small portion of the variation in the rhs endogenous variable. In situations of low information regarding causal effects (either because of small samples or weak instruments or both), then standard asymptotic distribution theory begins to break down. What happens is that asymptotic standard errors are no longer valid and are generally too small. Thus, confidence intervals constructed from asymptotic standard errors typically have much lower coverage rates than their nominal coverage probability. This phenomenon has spawned a large sub-literature in econometrics on the so-called weak or many instruments problem. In marketing applications, we typically do not encounter the "many" instrument problem in the sense that we don't have more than a handful of potential instruments.

There is a view among some applied econometricians that failure of standard asymptotics is only occurs for very small values of the first-stage R-squared or when the F-stat for the first stage is less than 10. This view comes from a misreading of the excellent survey of Stock et al. (2002). The condition of requiring the first stage F-stat be > 10 comes in the problem with only one instrument (in general, the "concentration parameter" or kF should be used). However, the results summarized in Stock et al. (2002) simply state that the "average asymptotic bias" will be less than 15 per cent in the case where $kF > 10$. This does not mean that confidence intervals constructed using the standard asymptotics will have good *actual* coverage properties (i.e. actual coverage close to nominal coverage). Nor does this result imply that there aren't finite sample biases of an even greater magnitude than these asymptotic biases.

The poor sampling properties of the IV estimator[28] can easily be shown even in cases where the instruments have a modest but not small first-stage R-squared. We

[28] Here I focus on the sampling properties of the estimator rather than the size of a test procedure. Simulations found in Hansen et al. (2008) show that coverage probabilities and bias can be very large even in situations where the concentration ratio is considerably more than 10.

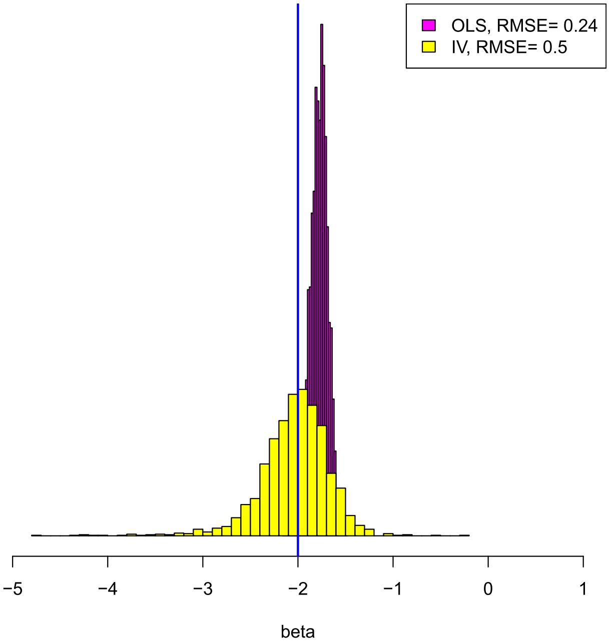

FIGURE 3

“Weak” instruments sampling distribution: $p = 1$.

simulate from the following system:

$$y = -2x + \varepsilon_y$$
$$x = Z\delta + \varepsilon_x$$
$$\begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} \sim N\left(0, \begin{bmatrix} 1 & .25 \\ .25 & 1 \end{bmatrix}\right)$$

$N = 100$. Z is an $N \times p$ matrix of iid Unif$(0, 1)$. The δ vector is made up of p identical elements, chosen to make the population first-stage R-squared equal to .10 using the formula, $\sqrt{\frac{12\rho^2}{p(1-\rho^2)}}$, where ρ^2 is the desired R-squared value. Fig. 3 shows the sampling distribution of the IV and OLS estimators in this situation with $p = 1$. The method of moments IV estimator has huge tails, causing it to have a much larger RMSE than OLS. OLS is slightly biased but without the fat tails of the IV estimator. Fig. 4 provides the distribution of the first-stage F statistics for 2000 replications. The vertical line in the figure is the median. This means that more than 50 per cent of these simulated samples had F-stats of greater than 10, showing the fallacy of this rule of thumb. Thus, for this case of only a moderately weak (but valid!) instrument,

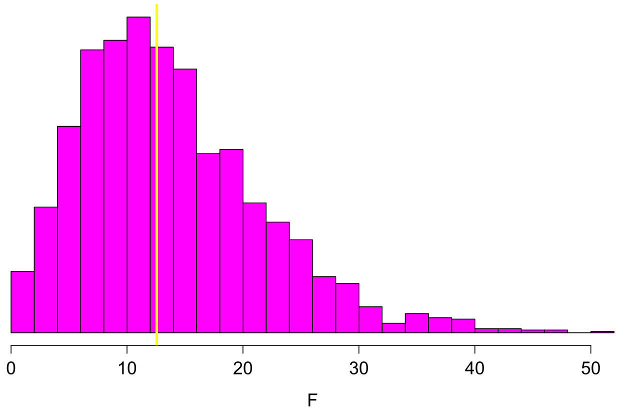

FIGURE 4

Distribution of first-stage F-statistics.

the IV estimator would require a sample size of approximately four[29] times larger than the OLS estimator to deliver the same RMSE level.

Lest the reader form the false impression that the IV estimator doesn't have appreciable bias, consider the case where there are 10 instruments instead of one but where all other parameters are held constant. Fig. 5 shows the sampling distributions in this case. The IV estimator now has both fat tails and finite sample bias.

The weak instruments literature seeks to improve on standard asymptotic approximations to the sampling distribution of the IV estimator. The literature focuses exclusively on improving inference which is defined as obtaining testing and confidence interval procedures which have correct size. That is, the weak instruments literature assumes that the researcher has decided to employ an IV method and just wants a test or confidence interval with the proper size. This literature does not propose new estimators with improved sampling properties but merely seeks to develop improved asymptotic approximation methods. This literature is very large and has made considerable progress on obtaining test procedures with actual size very close to nominal size under a variety of assumptions.

There are two major variants in this literature. One variant starts from what is called the Conditional Likelihood Ratio statistic and builds a testing theory which is exact under the homoskedastic, normal case (conditional on the error covariance matrix) (see Moreira, 2003 as an example). The other variant uses the GMM approach to define a test statistic which is consistent in the presence of heteroskedasticity and does not rely on normal errors (see Stock and Wright, 2000). The GMM variant will never deliver exact results but is potentially more robust. Both the CLR and GMM methods will work very well when the averages of y and x used in the IV estimator

[29] $\left(\frac{.5}{.24}\right)^2$.

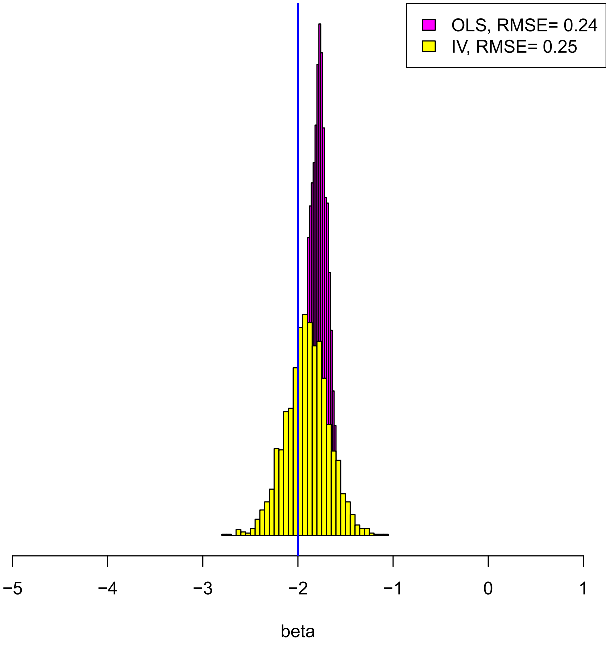

FIGURE 5

“Weak” instruments sampling distribution: $p = 10$.

(see, for example, (46)) are approximately normal. This happens, of course, when the CLT sets in quickly. The performance of these methods is truly impressive even in small samples in the sense that the nominal and actual coverage of confidence intervals is very close. However, the intervals produced by the improved methods simply expose the weakness of the IV estimators in the first place, that is the intervals can be very large (in fact, the intervals can be of infinite length). The fact that proper size intervals are very large simply says that if you properly measure sampling error, it can be very large for IV estimators. This reflects the fact that an IV approach uses only a portion of the total sample variability or information to estimate the structural parameter.

5.7.2 Choice models

Much of the applied econometrics done in marketing is done in the context of a logit choice model of demand rather than a linear structural model. Much of the intuition regarding the problems with IV methods in linear models carries over to the nonlinear case. For example, the basic exclusion restriction that underlies the validity of an instrument also applied to a non-linear model. The idea that the instruments partition the variability of the endogenous rhs variable still applies. The major difference is that the GMM estimator is now motivated not just by the assumption that the structural errors are uncorrelated with the instruments but on a more fundamental notion that the instruments and the structural errors are conditionally independent. Replacing zero conditional correlation with conditional independence means that the moment conditions used to develop the GMM approach can be generated by not just assuming that the error terms are orthogonal to the instruments but also to any function of the instruments. This allows a greater flexibility than in the linear IV model. In the linear IV model, we need as many (or more) instruments as there are included rhs endogenous variables to identify the model. However, in a nonlinear model such as the choice model, any function of the instruments is also a valid instrument and can be used to identify model parameters. To make this concrete, consider a very simply homogeneous logit model.

$$\Pr(j|t) = \frac{\exp\left(\alpha' c_{j,t} + \beta' m_{j,t} + \xi_{j,t}\right)}{\sum_j \exp\left(\alpha' c_{j,t} + \beta' m_{j,t} + \xi_{j,t}\right)} \tag{52}$$

Here $\xi_{j,t}$ is an unobservable, $m_{j,t}$ are the marketing mix variables for alternative j observed at time t, and $c_{j,t}$ are the characteristics of choice alternative j. The endogeneity problem comes from the possibility that firms set the marketing mix variables with partial knowledge of the unobservable demand shock and, therefore, the marketing mix variables are possibly a function of the $\xi_{j,t}$. Since the choice model is a linear index model, this is the same as suggesting that the unobservables are correlated with the marketing mix variables. The IV estimator would be based on the assumption that there exists a matrix, Z, of observations on valid instruments which are variables are conditionally independent of $\xi_{j,t}$.

$$\mathbb{E}\left[\xi_t g\left(z_t\right)'\right] = 0 \text{ for any measurable function, } g\left(\right)$$

z_t is the vector of the p instrumental variables. As a practical matter, this means that one can use as valid instruments any polynomial function of the z variables and interactions between the instruments, greatly expanding the number of instruments. However, the identification achieved by expanding the set of instruments in this manner is primarily from the model functional form.

To illustrate the problem with IV estimators for the standard choice model, we will consider the choice model in (52) along with a standard linear IV "first-stage" equation.

$$m_t = z_t \Delta + \nu_t$$

ν_t and ξ_t are correlated, giving rise to the classic omitted variable interpretation of the endogeneity problem. To examine the sampling properties of the IV estimator for this situation, we will consider the special case where there is only one endogenous marketing mix variable, there is only one instrument, and the choice model is a binary choice model. To generate the data, we will assume that the unobserved demand shocks joint normal with the errors in the IV equation.

$$\Pr(1) = \frac{\exp\left(\alpha + \beta m + \xi\right)}{1 + \exp\left(\alpha + \beta m + \xi\right)}$$

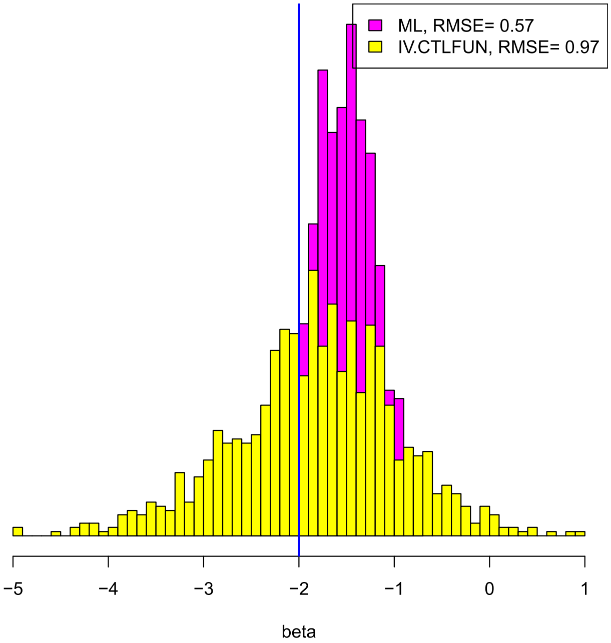

FIGURE 6

Sampling distributions for logit models.

$$m = \delta' z + \nu$$

$$\begin{pmatrix} \xi \\ \nu \end{pmatrix} \sim N\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

Here we have arbitrarily written the probability of choice alternative 1. We used the same parameters as in the simulation of the weak instruments problem for the linear IV with one instrument, $N = 100$, $\rho = .25$, and δ is set for that first-stage R-squared is 0.10. Fig. 6 shows the sampling distribution of the standard ML estimator which ignores the endogeneity (shown by the darker “magenta” histogram). We used a control-function approach to compute the IV estimator for this problem under the assumption of a linear first stage: We regressed the endogenous rhs variable, m, on the instruments and used the residual from this regression to create additional explanatory variables which were included in the logit model. In particular, we used the residual, the residual-squared, the residual-cubed, and exp of the residual as control functions. The sampling distribution of this estimator is shown by the yellow histogram. The sampling performance of the IV estimator is considerably inferior to that of the MLE which ignores the endogeneity problem. The fat tails of the IV estimator contribute to a RMSE of about twice that of the MLE. The IV estimator appears to be approximately unbiased, however, this goes away quickly if you increase the number of instruments, while the RMSE remains high.

5.8 Conclusions regarding the statistical properties of IV estimators

We have seen that an IV estimator can exhibit substantial finite sample bias and tremendous variability, particularly in the case of small samples, non-normal errors, and weak to moderate instruments. The failure of standard asymptotic theory applies not just to extreme case of very weak instruments, but also to cases of moderate strength instruments. All of these results assume that the instruments used are valid. If there are even "small" violations of the exclusion restriction (i.e. the instruments have a direct effect on y), then the statistical performance of the IV estimator degrades even further.

The emphasis in the recent econometrics literature on instruments is on improved testing and confidence interval construction. This emphasis is motivated by a "theory-testing" mentality. That is, researchers want to test hypotheses regarding whether or not a causal effect exists. The emphasis is not on predicting y conditional on a change in x. This exposes an important difference between economics and marketing applications. In many (but not all) marketing applications, we are more interested in conditional prediction rather than testing a hypothesis. If our goal is to help the firm make better decisions, then the first step is to help the firm make better predictions of the effects of changes in marketing mix variables. One may actually prefer estimators which do not attempt to adjust for endogeneity (such as OLS) for this purpose. OLS can have a much lower RMSE than an IV method.

In sum, IV methods are costly to apply and prone to specification errors. This serves to underscore the need for caution and the requirement that arguments in support of potential endogeneity bias and validity must be strong.

5.9 Endogeneity in models of consumer demand

Much empirical research in marketing is directed toward calibrating models of product demand (see, for example, the Chintagunta and Nair, 2011 survey and Chapter 1 of this volume). In particular, there has been a great deal of emphasis on discrete choice models of demand for differentiated products (for an overview, see pp. 4178–4204 of Ackerberg et al., 2007). Typically, these are simple logit models which allow marketing mix variables to influence demand and account for heterogeneity. An innovation of Berry et al. (1995) was to include a market wide error term in this logit structure so that the aggregate demand system is not a deterministic function of product characteristics and marketing mix variables.

$$MS(j|t) = \int \frac{\exp\left(\alpha' c_j + \beta' m_{jt} + \xi_{jt}\right)}{\sum_{j=1}^{J} \exp\left(\alpha' c_{jt} + \beta' m_{jt} + \xi_{jt}\right)} p(\alpha, \beta)\, d\alpha d\beta \tag{53}$$

There are J products observed either in T time periods or in a cross section of T markets. c_j is a vector characteristic of the jth product, m_{jt} is a vector of market mix variables such as price and promotion for the jth product, and ξ_{jt} represents an error term which is often described as a "demand shock." The fact that consumers are heterogeneous is reflected by integrating the logit choice probabilities over a distribution of parameters which represents the distribution of preferences in the market. This basic model represents a sort of intersection between marketing and I/O and provides a useful framework to catalog the sorts of instruments used in the literature.

5.9.1 Price endogeneity

(53) provides a natural motivation for concerns regarding endogeneity using an omitted variables interpretation. If we could observe the ξ_{jt} variable, then we would simply include this variable in the model and we would be able to estimate the β parameters which represent the sensitivity to marketing mix variables. However, researchers do not observe ξ_{jt} and it is entirely possible that firms have information regarding ξ_{jt} and set marketing mix variables accordingly. One of the strongest argument made for endogeneity is the argument of Berry et al. (1995) that if ξ_{jt} represents an unobserved product characteristic (such as some sort of product quality) that we would expect that firms would set price as a function of ξ_{jt} as well as of the observed characteristics. This a very strong argument, when applied in marketing applications, as the observed characteristics of many consumer products are often limited to packaging, package size, and a subset of ingredients. For consumer durable goods, the observed characteristics are also limited as it is difficult to quantify design, aesthetic, and performance characteristics. We might expect that price and unobserved quality are positively correlated, giving rise to a classic downward endogeneity bias in price sensitivity. This would result in what appears to be sub-optimal prices.

There are many possible interpretations of the ξ_{jt} terms other than the interpretation as an unobserved product characteristic. If the demand is observed in cross-section of markets, we might interpret the ξ_{jt} as unobserved market characteristics that make particular brands more or less attractive in this market. If the t index is time, then others have argued that the ξ_{jt} represent some sort of unobserved promotional or advertising shock.

These arguments for endogeneity bias in the price coefficient have led to the search for valid instruments for the price variable. The obvious place to look for instruments is the supply side which consists of cost and competitive information. The idea here is that costs do not affect demand and therefore serve to push around price (via some sort of mark-up equation) but are uncorrelated with the unobserved demand shock, ξ_{jt}. Similarly, the structure of competition should be a driver of price but not of demand. If a particular product lies in a crowded portion of the product characteristics space, then we might expect smaller mark-ups than a product that is more isolated.

The problem with cost-based instruments is lack of variability and observability. For some types of consumer products, input costs such as raw material costs may be observable and variable, but other parts of marginal cost may be very difficult to measure. For example, labor costs, measured by the Bureau of Labor Statistics, are based on surveys with a potentially high measurement error. Typically, those costs that are observable do not vary by product so that input costs are not usable as instruments for the highly differentiated product categories studied in marketing applications.

If the data represent a panel of markets observed over time, then the suggestion of Hausman (1996) can be useful. The idea here is that the demand shocks are not correlated across markets but that costs are.[30] If this is the case, then the prices of products in other markets would be valid instruments. Hausman introduced this idea to get around the problem that observable input costs don't vary by product. To evaluate the usefulness and validity of the Hausman approach, one must take a critical view of what the demand shocks represent. If these error terms represent unobservable market level demand characteristics which do not vary over time, then simply including market fixed effects would eliminate the need for instruments. One has to argue that the unobserved demand shocks are varying both by market and by time period in the panel. For this interpretation, authors often point to unobserved promotional efforts such as advertising and coupon drops. If these promotional efforts have a much lower frequency than the sampling frequency of the data (e.g. feature promotions are planned quarterly but we observe weekly demand data), then it is highly unlikely that these unobservables explain much of the variation in demand and that this source of endogeneity concerns is weak.

For products with few observable characteristics and for cross-sectional data, Berry et al. (1995) make a strong argument for price endogeneity. However, their arguments for the use of characteristics of other products as potential instruments are not persuasive for marketing applications. Their argument is that the characteristics of competing products will influence mark-up independent of demand shocks. This may be reasonable. However, their assumption that firms observed characteristics are exogenous and set independently of the unobservable characteristic is very likely to be incorrect. Firms set the bundle of both observed and unobserved characteristics jointly. Thus, the instruments proposed by Berry et al. (1995) are unlikely to be valid. With panel data, there is no need to use instruments as simple product specific fixed effects would be sufficient to remove the "endogeneity" bias problem as long as the unobserved product characteristics do not vary across time.

5.9.2 Conclusions regarding price endogeneity

Price endogeneity has received a great deal of attention in the recent marketing literature. There is no particular reason to single out price as the one variable in the marketing mix which has potentially the greatest problems of endogeneity bias. In fact, the source of variation in prices in most marketing datasets consists of cost variation (including wholesale price variation) and the ubiquitous practice of temporary price promotions or sales. Within the same market over time, it is hard to imagine what the unobservable demand shocks are that vary so much over time and by brand. Retailers set prices using mark-up rules and other heuristics that do not depend on market wide variables. Cost variables are natural price instruments but lack variation over time and by brand. Wholesale prices, if used as instruments, will confuse long and short run price effects. We are not aware of any economic arguments which can justify the use of lagged prices as instruments. In summary, we believe that, with panel or time-series data, the emphasis on price endogeneity has been misplaced.

[30] Given that, for many products, there are national advertising and promotional campaigns suggests that the Hausman idea will only work if there are advertising expenditure variables included in the model.

5.10 Advertising, promotion, and other non-price variables

While the I/O literature has focused heavily on the possibility of price endogeneity, there is no reason to believe, a priori, that the endogeneity problem is confined to price. In packaged goods, demand is stimulated by various "promotional" activities which an include what amount to local forms of advertising from display signage, direct mail, and newspaper inserts. In the pharmaceutical and health care products industry, large and highly compensated sales forces "call on" doctors and other health care professionals to promote products (this is often called "detailing"). In many product categories, there is very little price variation but a tremendous expenditure of effort on promotional activities such as detailing. This means that for many product categories, the advertising/promotional variables are more important than price. An equally compelling argument can be made that these non-price marketing mix variables are subject to the standard "omitted" variable endogeneity bias problem.

For example, advertising would seem to be a natural variable that is chosen as a function of demand unobservables. Others have suggested that advertising is determined simultaneously along with sales as firms set advertising expenditures as a function of the level of sales. In fact, the classical article (Bass, 1969) uses linear simultaneous equations models to capture this "feedback" mechanism for advertising.

The standard omitted variables arguments apply no less forcefully to non-price marketing mix variables. This motivates a search for valid instruments for advertising and promotion. Other than costs of advertising and promotion, there is no set of instruments that naturally emerge as valid and strong instruments. Even the cost variables are unlikely to be brand or product-specific and may vary only slowly over time, maximizing the "weak" instruments problem.

We have observed that researchers have argued persuasively that some kinds of panel data can be used to infer causal effects of advertising by using fixed effects to control for various concerns that changes advertising allocations over time or that specific markets receive allocations that depend on possible responsiveness to the ad or campaign in question (see, for example, Klapper and Hartmann, 2018). In the panel setting, researchers limit the source of variation so as to reduce concerns for endogeneity bias. The IV approach is to affirmatively seek out "clean" or exogenous variation. In the same context of measuring return to Super Bowl ads, Stephens-Davidowitz et al. (2015) use whether or not the home team is in the Super Bowl as an explicit instrument (exposure to the ad because viewership changes if the home team is in the Super Bowl which they argue is genuinely unpredictable or random at the point at which advertisers bid on Super Bowl slots. This is clearly a valid instrument in the same way as Angrist's Vietnam draft lottery is a valid instrument and no further proof is required. However, such truly random instruments are extremely rare.

5.11 Model evaluation

The purpose of causal inference in marketing applications is to inform firm decisions. As we have argued, in order to optimize actions of the firm, we must consider counterfactual scenarios. This means that the causal model must predict well in conditions that can be different from those observed in the data. The model evaluation exercise must validate the model's predictions across a wide range of different policy regimes. If we validate the model under a policy regime that is the same or similar to the observational data, then that validation exercise will be uninformative or even misleading.

To see this point clearly, consider the problem of making causal inferences regarding a price elasticity. The object of causal inference is the true price elasticity in a simple log-log approximation.

$$\ln Q_t = \alpha + \eta \ln P_t + \varepsilon_t$$

Imagine that there is an "endogeneity" problem in the observational data in which the firm has been setting price with partial knowledge of the demand shocks which are in the error term. Suppose further, that the firm raises price when it anticipates a positive demand shock. This means that a OLS estimate of the elasticity will be too small and we might conclude, erroneously, that the firm should raise its price even if the firm is setting prices optimally. Suppose we reserve a portion of our observational data for out-of-sample validation. That is, we will fit the log-log regression on observations, $1, 2, \ldots, T_0$, reserving observations $T_0 + 1, \ldots, T$ for validation. If we were to compare the performance of the inconsistent and biased OLS estimator of the price elasticity with any valid causal estimate using our "validation" data, we would conclude that OLS is superior using anything like the MSE metric. This is because OLS is a projection-based estimator that seeks to minimize mean squared error. The only reason OLS will fare poorly in prediction in this sort of exercise is if the OLS model is highly over-parameterized and the OLS procedure will over-fit the data. However, the OLS estimator will yield non-profit maximizing prices if used in a price optimization exercise because it is inconsistent for the true causal elasticity parameter. Thus, we must devise a different validation exercise in evaluating causal estimates. We must either find different policy regimes in our observational data or we must conduct a validation experiment.

6 Conclusions

Marketing is an applied field where the emphasis is on providing advice to firms on how to optimize their interaction with customers. The decision-oriented emphasis motivates an interest in both inference paradigms compatible with decision making as well as response functions which are nonlinear. Much of the recent emphasis in econometrics is focused on estimating effects of a policy variable via linear approximations or even via a "local" treatment effect or (LATE). Unfortunately, local treatment effects are only a beginning of an understanding of policy effects and policy optimization. This means that, in marketing, we will have to impose something like a parametric model (which can be achieved via strong priors and non-parametric approaches) in order to make progress on the problem of optimizing marketing policies.

On the positive side, marketing researchers have access to an unprecedented amount of detailed observational data regarding the reactions of customers to a ever increasing variety of firm actions (for example, the many types of advertising or marketing communications possible today). Large scale randomized experimentation holds out the possibility of valid causal inferences regarding the effects of marketing actions. Thus, the amount of both observational and experimental data at the disposal of a marketing researcher is truly impressive and growing as new sources of spatial (via phone or mobile device pinging) and other types of data are possible. Indeed, some data ecosystems such as Google/Facebook/Amazon in the U.S. and Alibaba/JD in China aim to amass data on purchases, search and product consideration, advertising exposure, and social interaction into one large dataset. At the same time, firms are becoming increasingly sophisticated in targeting customers based on information regarding preferences and responsiveness to marketing actions. This means that the evaluation of targeted marketing actions may be difficult or even impossible with even the richest observational data and may require experimentation.

We must use observational data to the greatest extent possible as it is impossible to optimize fully marketing actions on the basis solely of experimentation. The emphasis in the econometrics literature has been on using only a subset of the variation in observational data so as to avoid concerns that invalidate causal inference. We can ill afford a completely purist point of view that we should only use a tiny fraction of the variation in our data to estimate causal effects or optimize marketing policies. Instead, our view is that we can restrict variation in observational data only when there are strong prior reasons to expect that use of a given dimension of variation will produce inconsistent estimates of marketing effects.

Experimentation must be combined with observational data to achieve the goals of marketing. It is highly unlikely that randomized experimentation will ever completely replace inferences based on observational data. Many problems in marketing are not characterized by attempts to infer about an average effect size but, rather, to optimize firm actions over a wide range of possibilities. Optimization cannot be achieved in any realistic situation only by experimental means. It is likely, therefore, that experiments should play a role in estimating some of the critical effects but models calibrated on observational data will still be required to make firm policy recommendations. Experiments could also be used to test key assumptions in the model such as functional form or exogeneity assumptions without requiring that policy optimization be the direct result of experimentation.

References

Ackerberg, D., Benkard, C.L., Berry, S., Pakes, A., 2007. Econometric tools for analyzing market outcomes. In: Handbook of Econometrics. Elsevier Science, pp. 4172–4271 (Chap. 63).

Allenby, G.M., Rossi, P.E., 1999. Marketing models of consumer heterogeneity. Journal of Econometrics 89, 57–78.

Angrist, J.D., Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics 106, 979–1014.

Angrist, J.D., Pischke, J.-S., 2009. Mostly Harmless Econometrics. Princeton University Press, Princeton, NJ, USA.

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics 2 (6), 1152–1174.

Bass, F.M., 1969. A simultaneous equation study of advertising and sales of cigarettes. Journal of Marketing Research 6 (3), 291–300.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. John Wiley & Sons.

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

Blake, T., Nosko, C., Tadelis, S., 2015. Consumer heterogeneity and paid search effectiveness: a large scale field experiment. Econometrica 83, 155–174.

Chen, Y., Yang, S., 2007. Estimating disaggregate models using aggregate data through augmentation of individual choice. Journal of Marketing Research 44, 613–621.

Chintagunta, P.K., Nair, H., 2011. Discrete-choice models of consumer demand in marketing. Marketing Science 30 (6), 977–996.

Conley, T.G., Hansen, C.B., Rossi, P.E., 2012. Plausibly exogenous. Review of Economics and Statistics 94 (1), 260–272.

Deaton, A., Cartright, N., 2016. Understanding and Misunderstanding Randomized Controlled Trials. Discussion Paper 22595. NBER.

Dube, J., Hitsch, G., Rossi, P.E., 2010. State dependence and alternative explanations for consumer inertia. The Rand Journal of Economics 41 (3), 417–445.

Dubé, J.-P., Fox, J.T., Su, C.-L., 2012. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. Econometrica 80 (5), 2231–2267.

Dubé, J.-P., Hitsch, G., Rossi, P.E., 2018. Income and wealth effects in the demand for private label goods. Marketing Science 37, 22–53.

Dubé, J.-P., Misra, S., 2018. Scalable Price Targeting. Discussion Paper. Booth School of Business, University of Chicago.

Eckles, D., Bakshy, E., 2017. Bias and High-Dimensional Adjustment in Observational Studies of Peer Effects. Discussion Paper. MIT.

Fennell, G., Allenby, G.M., Yang, S., Edwards, Y., 2003. The effectiveness of demographic and psychographic variables for explaining brand and product use. Quantitative Marketing and Economics 1, 223–244.

Fruhwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. Chapman and Hall.

George, E.I., McCulloch, R.E., 1997. Approaches for Bayesian variable selection. Statistica Sinica 7, 339–373.

Gilbride, T.J., Allenby, G.M., 2004. A choice model with conjunctive, disjunctive, and compensatory screening rules. Marketing Science 23 (3), 391–406.

Gordon, B.R., Zettelmeyer, F., 2017. A Comparison of Approaches to Advertising Measurement. Discussion Paper. Northwestern University.

Griffin, J., Quintana, F., Steel, M.F.J., 2010. Flexible and nonparametric modelling. In: Geweke, J., Koop, G., Dijk, H.V. (Eds.), Handbook of Bayesian Econometrics. Oxford University Press.

Hansen, C., Hausman, J., Newey, W., 2008. Estimation with many instrumental variables. Journal of Business and Economic Statistics 26 (4), 398–422.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50 (4), 1029–1054.

Hartmann, W.R., Nair, H.S., Narayanan, S., 2011. Identifying causal marketing mix effects using a regression discontinuity design. Marketing Science 30, 1079–1097.

Hausman, J., 1996. The valuation of new goods under perfect and imperfect competition. In: Bresnahan, T., Gordon, R. (Eds.), The Economics of New Goods, vol. 58. University of Chicago, pp. 209–237.

Hayashi, F., 2000. Econometrics. Princeton University Press.

Heckman, J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models. Econometrica 52 (2), 271–320.

Heckman, J., Vytlacil, E.J., 2007. Econometric evaluation of social programs. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier, pp. 4779–4874.

Hitsch, G., Misra, S., 2018. Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. Discussion Paper. Booth School of Business, University of Chicago.

Hoch, S.J., Dreze, X., Purk, M.E., 1994. EDLP, Hi-Lo, and margin arithmetic. Journal of Marketing 58 (4), 16–27.

Imbens, G.W., Kolesar, M., 2016. Robust standard errors in small samples: some practical advice. Review of Economics and Statistics 98 (4), 701–712.

Imbens, G.W., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. Journal of Econometrics 142, 807–828.

Imbens, G.W., Rubin, D.B., 2014. Causal Inference. Cambridge University Press.

Jiang, R., Manchanda, P., Rossi, P.E., 2009. Bayesian analysis of random coefficient logit models using aggregate data. Journal of Econometrics 149, 136–148.

Johnson, G.A., Lewis, R.A., Nubbemeyer, E.I., 2017. Ghost ads: improving the economics of measuring online ad effectiveness. Journal of Marketing Research 54, 867–884.

Klapper, D., Hartmann, W.R., 2018. Super bowl ads. Marketing Science 37, 78–96.

Lewis, R.A., Rao, J.M., 2014. The Unfavorable Economics of Measuring the Returns to Advertising. Discussion Paper. NBER.

Lodish, L., Abraham, M., 1995. How T.V. advertising works: a meta-analysis of 389 real world split cable T.V. advertising experiments. Journal of Marketing Research 32 (2), 125–139.

Manchanda, P., Rossi, P.E., Chintagunta, P.K., 2004. Response modeling with nonrandom marketing-mix variables. Journal of Marketing Research 41, 467–478.

McFadden, D.L., Train, K.E., 2000. Mixed MNL models for discrete response. Journal of Applied Econometrics 15, 447–470.

Moreira, M.J., 2003. A conditional likelihood ratio test for structural models. Econometrica 71, 1027–1048.

Musalem, A., Bradlow, E.T., Raju, J.S., 2009. Bayesian estimation of random-coefficients choice models using aggregate data. Journal of Applied Econometrics 24, 490–516.

Narayanan, S., Nair, H., 2013. Estimating causal installed-base effects: a bias-correction approach. Journal of Marketing Research 50 (1), 70–94.

Neyman, J., 1990. On the application of probability theory to agricultural experiments: essay on principles. Statistical Science 5, 465–480.

Nickell, S., 1981. Biases in dynamic models with fixed effects. Econometrica 49 (6), 1417–1426.

Park, T., Casella, G., 2008. The Bayesian lasso. Journal of the American Statistical Association 103 (482), 681–686.

Petrin, A., Train, K., 2010. Control function corrections for unobserved factors in differentiated product models. Journal of Marketing Research 47 (1), 3–13.

Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods, second ed. Springer.

Rossi, P.E., 2014a. Bayesian Non- and Sem-Parametric Methods and Applications. The Econometric and Tinbergen Institutes Lectures. Princeton University Press, Princeton, NJ, USA.

Rossi, P.E., 2014b. Even the rich can make themselves poor: a critical examination of IV methods in marketing applications. Marketing Science 33 (5), 655–672.

Rossi, P.E., Allenby, G.M., McCulloch, R.E., 2005. Bayesian Statistics and Marketing. John Wiley & Sons.

Sahni, N., 2015. Effect of temporal spacing between advertising exposures: evidence from online field experiments. Quantitative Marketing and Economics 13 (3), 203–247.

Scott, S.L., 2014. Multi-Armed Bandit Experiments in the Online Service Economy. Discussion Paper. Google Inc.

Shapiro, B., 2018. Positive spillovers and free riding in advertising of pharmaceuticals: the case of antidepressants. Journal of Political Economy 126 (1).

Stephens-Davidowitz, S.H., Varianc, H., Smith, M.D., 2015. Super Returns to Super Bowl Ads? Discussion Paper. Google Inc.

Stock, J.H., Wright, J.H., 2000. GMM with weak identification. Econometrica 68 (5), 1055–1096.

Stock, J.H., Wright, J.H., Yogo, M., 2002. A survey of weak instruments and weak identification in generalized method of moments. Journal of Business and Economic Statistics 20 (4), 518–529.

Woolridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA, USA.