Investigating Endogeneity Bias in Marketing

Qing Liu, Thomas Otter, Greg M. Allenby,

Please scroll down for article—it is on subsequent pages

Investigating Endogeneity Bias in Marketing

Qing Liu
Department of Marketing, University of Wisconsin–Madison, Madison, Wisconsin 53706, qliu@bus.wisc.edu

Thomas Otter
Faculty of Business and Economics, Johann Wolfgang Goethe University, Frankfurt, Germany,
otter@marketing.uni-frankfurt.de

Greg M. Allenby
Fisher College of Business, Ohio State University, 2100 Neil Avenue, Columbus, Ohio 43210, allenby.1@osu.edu

The use of adaptive designs in conjoint analysis has been shown to lead to an endogeneity bias in part-worth estimates using sampling experiments. In this paper, we re-examine the endogeneity issue in light of the likelihood principle. The likelihood principle asserts that all relevant information in the data about model parameters is contained in the likelihood function. We show that, once the data are collected, adhering to the likelihood principle leads to analysis where endogeneity becomes ignorable for estimation. The likelihood principle is implicit to Bayesian analysis, and discussion is offered for detecting and dealing with endogeneity bias in marketing.

Key words: likelihood principle; adaptive design; Bayes theorem; directed acyclic graphs
History: This paper was received January 31, 2006, and was with the authors 2 months for 2 revisions; processed by Eric Bradlow.

1. Introduction

The recent paper by Hauser and Toubia (2005), "The Impact of Utility Balance and Endogeneity in Conjoint Analysis," raises an interesting set of issues related to a wide class of marketing models. In their paper, they present evidence that adaptive designs, where answers to early questions in a conjoint interview are used to select later questions, induce biases in the estimated part-worths. While their paper focuses on analysis associated with Sawtooth Software's popular ACA (Adaptive Conjoint Analysis) software, the implications of their analysis reach beyond conjoint analysis, sequential analysis, and utility balance, touching on important philosophical issues at the core of statistical inference.

In this paper, we reexamine the endogeneity bias identified by Hauser and Toubia (HT) and explain its presence using traditional econometric methods. Bias is an aspect of statistical inference that relies on the notion of a sampling experiment, where hypothetical data sets are used to characterize the performance of an estimator. We argue that sampling experiments are useful to study properties of estimators and other procedures when real data are not available. However, when data are available, analysis should proceed according to the likelihood principle as originally proposed by Fisher (1922).

The likelihood principle asserts that the likelihood contains all the information about model parameters

(e.g., conjoint part-worths) in the data. We show that, according to the likelihood principle, the endogeneity created by adaptive questioning is not of concern for estimation, i.e., does not alter the likelihood function of the observed data. Our discussion of the likelihood principle raises a number of philosophical issues at the core of statistical inference that highlight the difference between classical (i.e., frequentist) and Bayesian philosophies.

Our analysis of endogeneity bias in conjoint models uses HT as a springboard for discussion and is not meant to criticize their findings. In fact, our analysis covers some of the same ground as their analysis, restating their findings in terms more familiar to the statistics literature. Our examples are sometimes similar to those examined by HT and sometimes depart from theirs to provide additional insight and analysis. We applaud their interest in analyzing the issue of endogeneity created by adaptive designs because of its importance to both practitioners and academic researchers.

The remainder of the paper is organized as follows. We begin with a review of endogeneity bias in regression models caused by adaptive designs, restating many of the points made by HT. We then introduce the likelihood principle and examine its implication for data analysis. Our analysis shows that, conditional on the dependent variable (y), the way the design is adaptively created is not informative

about model parameters. The mechanism employed to pick the design points, and the resulting endogeneity, are "ignorable" for the purpose of likelihood-based inference (Gelman et al. 2004, p. 203; Rubin 1976).[1] Thus, we believe the potential harm from endogeneity created by adaptive designs should be considered when selecting among experimental procedures (e.g., adaptive versus fixed designs), not in estimation. Discussion is provided about the importance of bias in the evaluation of procedures, and tools are offered for detecting when endogeneity created by adaptive questioning and other forms of sample selection will impact conditional inference—i.e., when it alters the likelihood function. Concluding comments are then offered, in which we advocate a Bayesian orientation to conducting analysis in marketing.

2. Endogeneity Bias

Endogeneity bias arises in regression analysis $y = X\beta + \varepsilon$ when the regressors, X, are not independent of the errors, ε. When a specific realization of the regressors, x'_t, is selected based on the outcome of previous choices, $y_{<t}$, as is the case for sequential design data, then x'_t is determined from within the system of study and is not independently determined. In this case, the traditional ordinary least-squares estimator is biased because the vector space of the regressors is dependent on the error realizations. Regression coefficient bias is defined as the difference between the true value of the parameter and the expected value of the estimate over hypothetical sets of data (D):

$$
\begin{aligned}
E_{D|\beta}[\hat{\beta}] &= E_{D|\beta}[(X'X)^{-1}X'y] \\
&= E_{D|\beta}[(X'X)^{-1}X'(X\beta + \varepsilon)] \\
&= \beta + E_{D|\beta}[(X'X)^{-1}X'\varepsilon].
\end{aligned} \quad (1)
$$

The presence of endogenous responses biases the regression estimate because the second term in Equation (1) is not equal to zero. To see this, consider a simple example where there is just one regressor without an intercept ($y = x\beta + \varepsilon$), and just two observations. The OLS estimator is given by

$$
\hat{\beta} = \frac{x_1 y_1 + x_2 y_2}{x_1^2 + x_2^2}. \quad (2)
$$

Then, if x_2 is determined by the value of y_1, i.e., $x_2 = f(y_1) = f(x'_1\beta + \varepsilon_1)$, we have

$$
E_{D|\beta}[\hat{\beta}] = \beta + E_{D|\beta}\left[\frac{x_1\varepsilon_1}{x_1^2 + x_2^2} + \frac{x_2\varepsilon_2}{x_1^2 + x_2^2}\right]. \quad (3)
$$

[1] We would like to thank the associate editor for pointing out that the result can be viewed as an extension/application of Rubin's (1976) framework for classifying missing data mechanisms.

The expectation of the first term in the brackets is not equal to zero because y_1, and hence ε_1, is used to determine x_2. Because ε_1 and x_2 are not independently determined, $E[f(x_2)\varepsilon_1] \neq E[f(x_2)]E[\varepsilon_1] = 0$, specifically

$$
E_{D|\beta}\left[\frac{x_1\varepsilon_1}{x_1^2 + x_2^2}\right] \neq E_{D|\beta}\left[\frac{x_1}{x_1^2 + x_2^2}\right]E_{D|\beta}[\varepsilon_1] = 0. \quad (4)
$$

Thus, for regression estimates to be unbiased it must be the case that all errors are independent of all the regressors (X).

The lack of independence between regressors and errors occurs in many models. Models with a lagged dependent variable (e.g., Koyck lag models, autoregressive models) violate the assumption that the regressors are independent of the error terms. In these models, the lagged regressors are themselves functions of the lagged error terms, and the assumption of independent regressors and errors is not valid. Endogeneity bias also occurs in models with sequential design, in which design points are selected based on previous responses. While the specific algorithm used to select design points (e.g., utility balance in ACA) may exacerbate the extent of endogeneity bias documented by frequentist analyses of small-sample situations with relatively large error variance, the bias is present whenever the vector of residuals, ε, is not independent of all the regressors.

Figure 1 illustrates the biasing effect of endogeneity, using a simple problem similar to that studied by HT. The analysis involves 1,000 replicates of samples, each consisting of 1,000 homogeneous respondents. Each respondent supplies three observations, where the regressor for the third observation is determined as a function of the first two observations. The

Figure 1   Expected Value of OLS Estimate Across 1,000 Replications
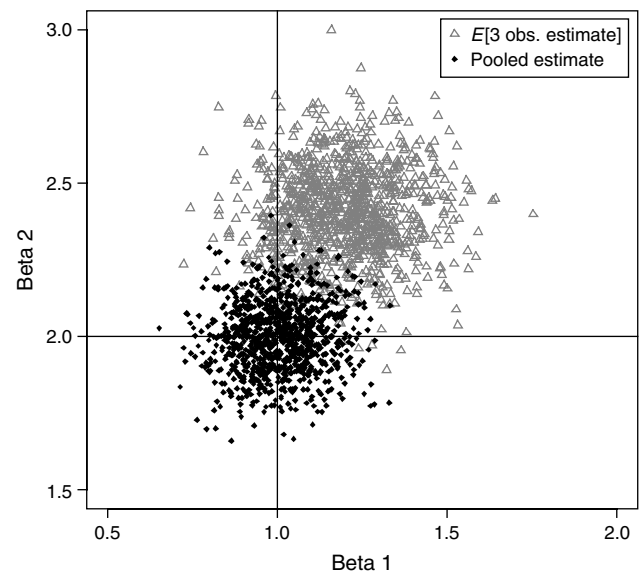
model is $y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$, $t = 1, 2, 3$ and $\varepsilon_t \sim$ Normal$(0, 25)$. The value of x for the first observation is $x_1' = (x_{11}, x_{21}) = (1, 0)$, the value of x for the second observation is $x_2' = (0, 1)$, and

$$x_3' = \begin{cases} (1, -1) & \text{if } y_1 y_2 > 0 \\ (1, 1) & \text{if } y_1 y_2 \leq 0 \end{cases}. \qquad (5)$$

The design rule in Equation (5) is meant to mimic the "utility balance" criterion used by the ACA software. The true values of the regression coefficients are $\beta_1 = 1$ and $\beta_2 = 2$. The figure shows that the OLS estimates exhibit positive bias, with $E[\hat{\beta}_1] = 1.198$ and $E[\hat{\beta}_2] = 2.406$. Each triangle character in the figure represents the mean of individual-level OLS estimates computed from one sample of $j = 1, \ldots, 1,000$ homogeneous respondents (i.e., $\sum_{j=1}^{1000} \hat{\beta}_j / 1,000$, where $\hat{\beta}_j$ is estimated with three observations $\{y_{jt}, x_{jt}\}$, $t = 1, 2, 3$).

Econometricians recognize the bias present in regression models when functions of lagged dependent variables are included in the model specification, and as a result, the finite sampling properties of these models are generally unknown (see Judge et al. 1988, p. 575). This has led to the use of asymptotics to characterize the sampling properties of estimators in such situations. In particular, the probability limit, or "plim" is used to describe the behavior of estimators as the sample size increases. The probability limit is a formal expression for the consistency of an estimator. An estimator is said to be consistent if the probability limit of obtaining an estimate arbitrarily close to the true parameter value equals one in infinite samples; i.e., the probability that an estimate from a sample of size T falls within the interval $[\beta - \varepsilon, \beta + \varepsilon]$ goes to one as the sample size increases, no matter how small ε:

$$\lim_{T \to \infty} P(|\hat{\beta}_T - \beta| < \varepsilon) = 1. \qquad (6)$$

A result known as Slutsky's theorem can be used to show that the OLS estimator is asymptotically consistent as defined above. Slutsky's theorem states that if $g(.)$ is a continuous function and z_T is some random variable that depends on T, then

$$\text{plim } g(z_T) = g(\text{plim } z_T). \qquad (7)$$

Thus, from Equation (1) we can derive the plim of the regression estimate as the sample size increases:

$$\begin{aligned} \text{plim } \hat{\beta} &= \beta + \text{plim}((X'X)^{-1} X'\varepsilon) \\ &= \beta + \text{plim}((X'X)/T)^{-1} \text{plim}(X'\varepsilon/T) \\ &= \beta, \qquad (8) \end{aligned}$$

where the last equality holds if $\text{plim}((X'X)/T)^{-1}$ converges to a finite-valued matrix (i.e., Σ_{XX}^{-1}), and $\text{plim}(X'\varepsilon/T)$ converges to zero. The latter condition

holds as long as $E_D[x_t \varepsilon_t] = 0$ for all t. Consistency only requires independence between regressor values x_t and their corresponding error term ε_t, and not the entire vector of errors. Thus, the requirements for asymptotic consistency are easier to obtain than the requirements of unbiasedness.

To illustrate the consistency of the OLS estimator in the simulation study, we obtained pooled OLS estimates for each of the replicated data sets, i.e., estimated regression coefficients using all 3,000 observations (1,000 respondents each supplying three observations) contained in each data set. The pooled estimates are plotted in Figure 1 as black diamonds. Clearly, the pooled estimates are not affected by the bias; their mean is $\hat{\beta}' = (1.005, 2.000)$ which is close to the true value of $\beta' = (1, 2)$. While the simulation study of HT does exhibit an endogeneity bias in small samples, the estimates are asymptotically consistent. Thus, endogenously determined covariates need not lead to inconsistent inferences, as is often assumed. As we show below, inconsistency arises from model misspecification, i.e., employing the wrong likelihood function for the data.

3. The Likelihood Principle

The study by HT touches on an important philosophical point about the role of sampling experiments in statistical analysis and managerial decision making. Bias is obviously a concern in conjoint analysis, where part-worth estimates are used to set prices and guide product formulation. Moreover, managers are often confronted with choosing the best way to collect and analyze data. Bayesians and frequentists disagree, however, on how to quantify knowledge given data and on how to study properties of statistical procedures and characterize their performance.

A principle of statistical inference first introduced by Fisher (1922), and adhered to in Bayesian analysis, is that the likelihood function contains all the information in the data about the model parameters. The likelihood principle implies that if two likelihoods for a parameter β are proportional, then we should make the same inference for β regardless of which likelihood we use. This principle has a number of significant and far-reaching implications. Berger and Wolpert (1984) provide an excellent review of this topic and provide many interesting examples. When considering implications of the likelihood principle, it is important to remember that the likelihood is always meant to represent the true data-generating mechanism, whose parameters have substantive meaning that guide decision making. Moreover, statistical inference involves the quantification of knowledge about these parameters.

An example of the application of the likelihood principle involves sequential sampling. Suppose we

observe the number of successes Z in n independent Bernoulli trials with success probability θ. The likelihood for the data is

$$\pi_1(Z = z \mid \theta) = \binom{n}{z} \theta^z (1 - \theta)^{n-z}. \quad (9)$$

Now, suppose that instead of holding fixed the number of trials, we were to decide to sample until we obtained z successes and then observed the realization of N, the number of trials to the zth success. In this scenario, N has a negative binomial distribution, so the model is

$$\pi_2(N = n \mid \theta) = \binom{n-1}{z-1} \theta^z (1 - \theta)^{n-z}, \quad (10)$$

and we have that, conditional on any pair $\{z, n\}$, $\pi_1(Z = z \mid \theta, n) \propto \pi_2(N = n \mid \theta, z)$. Despite the fact that the sampling scheme and the dependent variable are different, the likelihood principle states that we should ignore this and make the same inference about θ in both cases. A Bayesian analysis of these data would be equivalent because the likelihood principle implies that the stopping rule associated with a data collection is irrelevant. In contrast, a frequentist analysis of these data would incorporate the effects of the stopping rule in computing standard errors of point estimates and constructing confidence intervals. Frequentist confidence intervals are based on the idea of multiple realizations of the data, across which the number of trials (n) will vary in Equation (10) but not in Equation (9).

Analysis in marketing contains many instances where adherence to the likelihood principle will lead to different inference. Analysis in direct marketing, for example, involves initial testing of offer formats using many lists, and subsequent mailings of the best offers using lists with the highest response rates. Ignoring the uncertainty present in the initial test phase does not matter as long as the goal of the analysis is to simply identify the best list for a specific offer. However, if the goal is more general, involving learning about offer formats and list characteristics associated with high yields using both sets of data, then whether one conditions on the observed data or not will matter. By considering multiple hypothetical realizations of the initial test data, frequentist inference must admit the possibility of alternative best offers mailed out in the second round. Frequentist confidence intervals are misrepresented unless alternative hypothetical second-round offers are included in the analysis. In contrast, the Bayesian approach is more straightforward because it adheres to the likelihood principle and simply conditions on the observed data in both samples.

Another example involves the use of screening questions by Internet merchants, where answers are used to determine offers that are likely to be of value to the customer. Frequentist analysis of the observed purchases of the offers needs to account for possibly different answers to the screening questions, leading to offers different from those that were observed, to properly quantify the information in the data. In contrast, adherence to the likelihood principle implies that one should simply condition on the observed responses to the screening questions in the analysis. However, as discussed in greater detail later, simply conditioning on the particular offerings presented, and ignoring the responses to the screening questions, results in a misspecified likelihood.

Consider again the stylized example above, similar to the problem studied by HT. The likelihood function should acknowledge that x_3, the third design point, is determined from within the system of study and is therefore endogenous:

$$\pi(y_1, y_2, x_3, y_3 \mid \beta, \sigma^2) = \pi_1(y_1, y_2 \mid \beta, \sigma^2)$$
$$\times \pi_2(x_3 \mid y_1, y_2, \beta, \sigma^2)$$
$$\times \pi_3(y_3 \mid x_3, \beta, \sigma^2), \quad (11)$$

where the conditioning variables x_1 and x_2 are not written to improve readability. The first factor on the right side of Equation (11) corresponds to the first two observations where the design points x_1 and x_2 are determined beforehand. The likelihood for these observations corresponds to that found in a standard regression analysis:

$$\pi_1(y_1, y_2 \mid \beta, \sigma^2)$$
$$= \frac{1}{2\pi\sigma^2} \exp\left[\frac{-1}{2\sigma^2}((y_1 - x_1'\beta)^2 + (y_2 - x_2'\beta)^2)\right]. \quad (12)$$

Similarly, the third factor, which conditions on the realized design point x_3, is the same as found in standard analysis:

$$\pi_3(y_3 \mid x_3, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}((y_3 - x_3'\beta)^2)\right]. \quad (13)$$

The effect of endogeneity is therefore isolated in the second factor, π_2.

The likelihood for x_3 is determined by Equation (5), where a positive product of y_1 and y_2 leads to the selection of $x_3' = (1, -1)$, and a negative product yields $x_3' = (1, 1)$. Given y_1 and y_2, the selection of x_3 is deterministic, or

$$\pi_2(x_3 \mid y_1, y_2, \beta, \sigma^2) = \pi_2(x_3 \mid y_1, y_2) = 1. \quad (14)$$

Thus, while the likelihood for design point x_3 is unconditionally dependent on the model parameters β and σ^2, the likelihood is independent of these parameters given y_1 and y_2.[2] This makes intuitive sense

---

[2] Adding a random component to the selection mechanism implying a nondegenerate π_2 does not change the basic argument. Conditional on previous data, x_3 is again independent of β and σ^2.

because we learn about β and σ^2 from the respondent, not the mechanism used to select the design point. Following the likelihood principle implies that, given the data, selection mechanisms such as ACA's utility balance are ignorable; i.e., the likelihood of model parameters is unchanged if π_2 is included.

Heterogeneity and Selection Bias
In §6 of their paper, HT suggest that the presence of heterogeneity might compound the biasing effects of utility balance by also biasing the hyperparameters of the model that describe the mean part-worths (i.e., the mean of the random-effects distribution).[3] The likelihood principle again implies that, *conditional on all the data*, adaptive utility balance, or any other criterion to adaptively choose design points, is irrelevant to the analysis. In the presence of consumer heterogeneity, a respondent's contribution to the likelihood is

$$l(\beta_i) = \pi(D_i \mid \beta_i), \qquad (15)$$

and, by conditional independence, the sample likelihood is

$$l(\{\beta_i\}_{i=1}^N) = \prod_i \pi(D_i \mid \beta_i). \qquad (16)$$

Equation (14) demonstrates that adaptive questioning does not change the likelihood for any of the "i" respondents. The total sample likelihood is therefore not affected. In the setting of a Bayesian hierarchical model with heterogeneity, a random-effects distribution for β_i is introduced along with a prior distribution for the hyperparameters, τ:

$$\pi(\{\beta_i\}_{i=1}^N, \tau \mid \{D_i\}_{i=1}^N)$$
$$\propto \prod_i \pi(D_i \mid \beta_i) \times \pi(\beta_i \mid \tau) \times \pi(\tau). \qquad (17)$$

This corresponds to a data-generating process where the β_i vectors are modeled as draws from the random-effects distribution $\pi(\beta_i \mid \tau)$, and the data is generated from the likelihood $\pi(D_i \mid \beta_i)$. Because adaptive questioning does not change the likelihood, it is not relevant to the analysis, whether heterogeneity is present or absent. Moreover, posterior estimates of the hyperparameters τ will converge to the true data-generating values as the sample size increases, even if each consumer provides a limited amount of information and the individual-level parameters, β_i, experience shrinkage toward the hyperparameter τ. This result is due to exchangeability of consumers implied by the random-effects distribution—adaptive questioning does not affect the exchangeability of the response vectors obtained from each consumer.

As a result, estimates of hyperparameters are based on relatively large samples where data are pooled across respondents and will be close to their true values. Estimates of individual-level parameters, β_i, are biased because their posterior distributions are based on an individual's own data (D_i), all other data $\{D_j\}$ as shared through the mixing distribution $\pi(\beta_i \mid \tau)$, and the prior distribution $\pi(\tau)$. The mixing distribution and prior distribution have a nonnegligible effect on the individual-level estimates. We comment further on sampling properties of Bayesian estimators below.

We investigate the impact of endogenous covariates in a hierarchical Bayes model by adding heterogeneity to our earlier example. That is, instead of 1,000 identical respondents, we assume that the part-worths for the respondents follow a random-effects distribution:

$$\beta_i \sim \text{Normal}\left(\bar{\beta} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, V_\beta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$
$$i = 1, \ldots, 1{,}000, \quad (18)$$

where i indexes the respondents. Diffuse prior distributions, centered on the true values, are assumed for the hyperparameters:

$$\pi(\sigma^2) = \frac{v_0 s_0^2}{\chi_{v_0}^2} \quad \text{with } v_0 = 5 \text{ and } s_0^2 = 25; \qquad (19)$$

$$\pi(V_\beta) = IW(v_0, V_0) \quad \text{with } v_0 = 10 \text{ and } V_0 = 10I. \qquad (20)$$

Figure 2 compares Bayesian analysis of $\bar{\beta}$ for exogenous versus endogenous covariates. Time-series plots of the draws of $\bar{\beta}$ for these two sets of data show little difference. The posterior mean for the plot with exogenous x_3 is $\bar{\beta}' = (2.00, 0.96)$, with a posterior standard deviation of $(0.12, 0.11)$. For x_3 endogenous, the posterior mean is $\bar{\beta}' = (2.01, 0.96)$, with a posterior standard deviation of $(0.12, 0.12)$. These results agree with the implications of the likelihood principle—analysis that conditions on all of the observed data is not affected by endogenously determined covariates. Analysis with fewer respondents (e.g., 100) yields similar results, although the precision of the parameter estimates is decreased. Managerial decisions based on the analysis of the data are also unaffected.

4. Discussion
Our analysis shows that the presence of endogenously determined covariates leads to small-sample bias in conjoint part-worths, which diminishes as the sample size increases. The likelihood principle, on the other hand, asserts that one should ignore the fact that the covariates are endogenously determined in estimation. We note that, in general, likelihood-based estimates are biased whenever model parameters are nonlinearly related to the observed data

---

[3] The selection bias results of HT (2005) may be based on analysis that excluded the self-explicated data. We comment further on that procedure, which ACA refers to as "pairs only," below.

Figure 2   MCMC Draws for Exogenous and Endogenous Covariates 1,000 Respondents
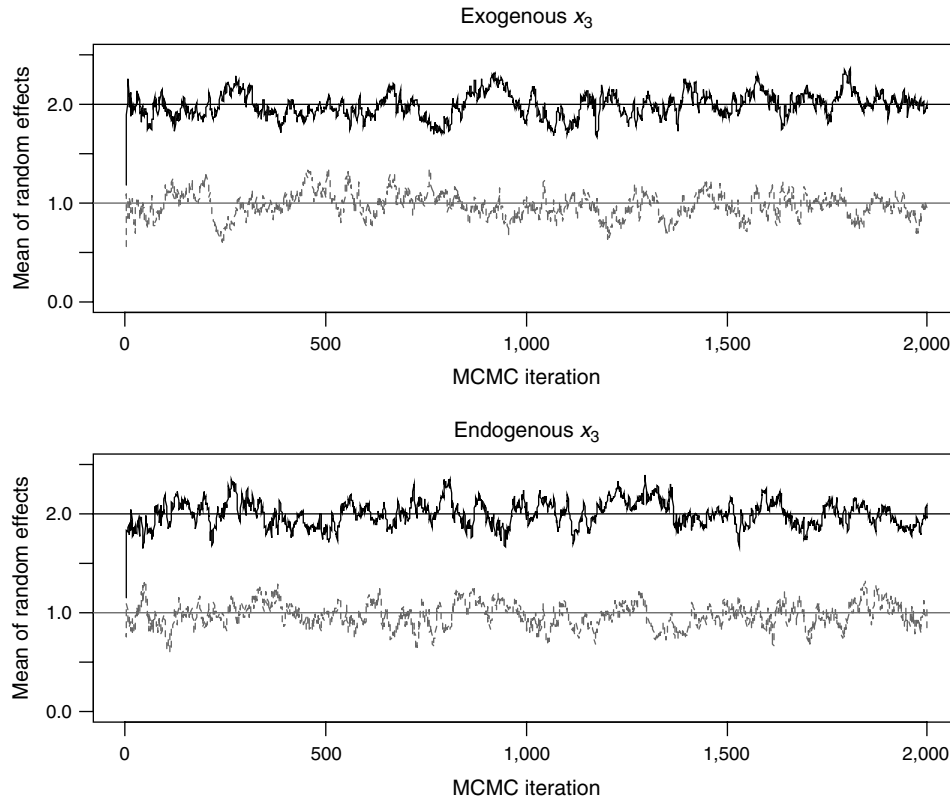
(e.g., variance estimates, logit coefficient estimates), including endogenously determined covariates. In this section we discuss the use of bias for evaluating the performance of estimators, and offer a method for determining when endogenously generated variables will create problems in analysis and decision making—i.e., are not ignorable.

Prior to data collection, managers might be concerned about the manner in which the data will be collected and the anticipated method used to analyze these data. While data collection (i.e., question selection, experimental design) and data analysis (i.e., parameter estimation) are related tasks, and decisions regarding them are often made jointly in light of commercially available software and user knowledge, they are not necessarily related. As discussed by HT and others, utility balance can be pursued for reasons that are difficult to quantify within a statistical model. Forcing respondents to select from among alternatives with nearly equal value might avoid scaling problems that occur when one alternative dominates the rest and might encourage respondents to more carefully evaluate the alternatives. These aspects are not reflected in the standard linear models used in traditional conjoint analysis and therefore do not enter into traditional analysis. Moreover, once the data are collected, one of a number of methods of estimation can be used in analysis (e.g.,

Bayesian, method of moments, maximum likelihood). It is therefore useful to think of data collection and data analysis separately.

Sampling Properties

Bias is one measure of the performance of an estimator, and while managers should be concerned about bias prior to data collection, bias should not be used as a litmus test for selecting an estimator or, as such, a scheme to generate design points adaptively. Bias is an aspect of statistical risk, defined as the expected loss from incorrectly estimating a respondent's true part-worths (β). For squared-error loss we have

$$
\begin{aligned}
\text{Risk}_{\hat{\beta}}(\beta) &= E_{D|\beta}[(\hat{\beta} - \beta)^2] \\
&= E_{D|\beta}[((\hat{\beta} - E_{D|\beta}[\hat{\beta}]) + (E_{D|\beta}[\hat{\beta}] - \beta))^2] \\
&= E_{D|\beta}[(\hat{\beta} - E_{D|\beta}[\hat{\beta}])^2 + (E_{D|\beta}[\hat{\beta}] - \beta)^2 \\
&\quad + 2(\hat{\beta} - E_{D|\beta}[\hat{\beta}])(E_{D|\beta}[\hat{\beta}] - \beta)] \\
&= \text{Var}(\hat{\beta}) + \text{Bias}(\hat{\beta})^2, \quad\quad (21)
\end{aligned}
$$

where the cross-product (third) term is zero because the expectation of its first factor in parentheses is zero. Bias can be traded off against variance to obtain lower risk, as it is in ridge regression, and is not usually pursued as a goal in and of itself. Many biased estimators have excellent sampling-theory properties.

Bayesian estimators ($\tilde{\beta}$), for example, minimize expected loss with respect to the posterior distribution (see Rossi et al. 2005, pp. 17–18):

$$\tilde{\beta} = \arg\min_{\hat{\beta}} \left\{ E_{\beta|D}[L(\hat{\beta}, \beta)] = \int L(\hat{\beta}, \beta)\pi(\beta \mid D)\, d\beta \right\}, \tag{22}$$

where $L(.)$ is the loss (e.g., squared-error loss) associated with using $\hat{\beta}$ to estimate β. Sampling properties of $\tilde{\beta}$ can be studied across multiple realizations of the data:

$$
\begin{aligned}
E_D[E_{\beta|D}[L(\tilde{\beta}, \beta)]] &= \iint L(\tilde{\beta}, \beta)\pi(\beta \mid D)\pi(D)\, d\beta\, dD \\
&= \iint L(\tilde{\beta}, \beta)\pi(D \mid \beta)\pi(\beta)\, dD\, d\beta \\
&= E_\beta[E_{D|\beta}[L(\tilde{\beta}, \beta)]] \\
&= E_\beta[\text{Risk}_{\tilde{\beta}}(\beta)]. \tag{23}
\end{aligned}
$$

The last equality shows that Bayes estimators have the property of minimizing expected risk, where the expectation is taken with respect to the prior distribution. Moreover, while Bayesian estimators are biased by the presence of the prior distribution, $\pi(\beta)$, they often outperform other estimators by successfully trading off increased bias for lower variance. Theoretically, bias (i.e., $E_{D|\beta}[\tilde{\beta} - \beta]$) is of diminished relevance to Bayesians because its computation requires knowledge of the unobserved true value of β.

An interesting issue is why one would want to engage in an adaptive design in the first place. After all, orthogonal designs are known to be optimal for linear models, and the use of procedures like utility balance will likely lead to a nonorthogonal design. As discussed by Chaloner and Verdinelli (1995), orthogonal designs are optimal only within the context of fixed designs in the absence of prior knowledge—i.e., designs where the questions are all selected prior to the collection of any data. Orthogonality is not necessarily optimal in the space of designs, D, that include design points that are selected sequentially. As pointed out by HT, utility balance as a criterion for adaptive design is unlikely to increase efficiency within the confines of the linear model. To the extent that the linear model is at best a local approximation, however, utility balance could help the design stay in the range of values that can be sensibly fitted by the local approximation. This is an empirical issue, the investigation of which is beyond the scope of this paper.

## Determining When Endogeneity Will Matter in Estimation

Directed acyclic graphs (DAGs) are useful tools for determining when conditioning on endogenously determined covariates results in a misspecified likelihood function. An introduction to DAGs for

Bayesian analysis can be found in Rossi et al. (2005, p. 67). The DAG for the utility-balance example discussed earlier is
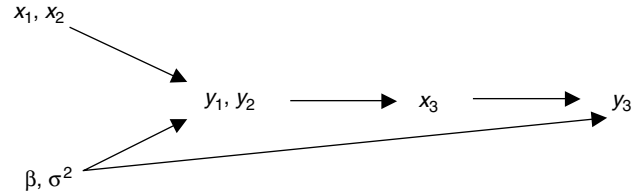

The graph is read from left to right, with the model parameters (β, σ^2) and two design points (x_1, x_2) used to generate responses y_1 and y_2. The third design point, x_3, is determined from the responses as per Equation (5), and used to generate the third response, y_3. The likelihood comprises all arrows connected to the model parameters (β, σ^2). Here we see that the third design point, x_3, is determined entirely from y_1 and y_2, and is not directly connected to the parameters. This implies that the mechanism for selecting x_3 can be ignored—i.e., is not relevant for inferences about (β, σ^2)—so long as y_1 and y_2 are included.

The direct-marketing example discussed earlier involved a pilot study (x_1) from which initial results (y_1) are obtained and used to design a subsequent study (x_2) whose results (y_2) are used to draw inferences about model parameters (θ). The DAG for this example is similar to the utility-balance example:
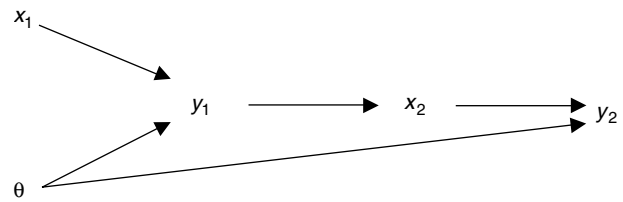
The arrow from y_1 to x_2 in the DAG above can represent a simple calculation as in Equation (5) above, or more complicated calculations involving the data, including the derivation and use of the posterior distribution $\pi(\theta \mid y_1, x_1) \propto \pi(y_1 \mid \theta, x_1)\pi(\theta)$ for choosing x_2. The posterior conditions on (x_1, y_1) and thus insulates x_2 from θ—i.e., if x_2 is chosen based on $\pi(\theta \mid y_1, x_1)$, x_2 cannot convey information about θ that is not already contained in $\pi(\theta \mid y_1, x_1)$. If, however, y_1 (e.g., self-explicated data, results of the pilot study, answers to screening questions) are not included in the analysis, or prior knowledge in the form of $\pi(\theta \mid y_1, x_1)$ used to select x_2 is discarded, the DAG becomes
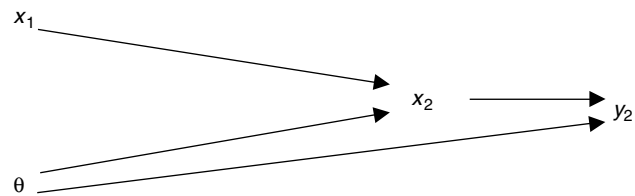
and the likelihood for the endogenously determined covariate (x_2) becomes a function of θ:

$$\pi(x_2, y_2 \mid \theta) = \int \pi(y_1, y_2, x_2 \mid \theta, x_1)\, d(y_1)$$
$$= \pi_1(x_2 \mid \theta, x_1) \times \pi_2(y_2 \mid \theta, x_2). \quad (24)$$

The likelihood contribution of x_2 is no longer ignorable. Nonignorable likelihoods occur whenever the data selection/creation mechanism for x is directly related to a parameter (θ) that is also related to the dependent variable y. In addition to the examples described above, marketing researchers often struggle with the issue of sample-selection bias, and various corrections have been proposed (e.g., Heckman 1976) to obtain consistent estimates. All the corrections are based on the presence of a common parameter (e.g., a correlation coefficient) related to variables x and y. Equation (24) indicates that, in principle, one should deal with an endogenously selected sample by including the likelihood for x as part of the model specification—i.e., to use the correct likelihood for all the data (x, y) (for marketing examples, see Manchanda et al. 2004, Zanutto and Bradlow 2006).

Effects of Omitting Self-Explicated Data in Conjoint Analysis

Consider what would happen if enough data were initially collected to make precise inferences about a respondent's part-worths. The utility balance algorithm would then generate candidate design points with responses close to zero for all choices. The likelihood π_3 in Equation (11) would then have multiple modes—one at zero, one at the true parameter value, and at all parameters values that are proportional to the true values. The mode at zero corresponds to one solution to fitting responses that are all equal to zero. The mode at the true parameter value is present as in any analysis, and the other values proportional to the true value will also yield zero responses for utility balance. Without access to the earlier data, or the original designer's knowledge to be used as a prior, inferences about the part-worths would not be identified by the data anywhere along the true parameter vector. The likelihood for stimulus selection (e.g., π_1 in Equation (24)) needs to be included in these analyses to avoid model misspecification. If past observations (in the sense that they are no longer available) were used to learn about model parameters, and design points were picked based on the acquired knowledge, then any analysis that proceeds without incorporating this fact is incomplete and the likelihood function misspecified.

Expanding on our earlier example, which mimics ACA, consider a model that has three regressors instead of two: $y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t$ and $\varepsilon_t \sim$ Normal $(0, 1)$. Assume we have three self-explicated

observations $y_{SE, i} = I\beta_i + \varepsilon_i$ and that we do not ask respondents questions about the attribute corresponding to the response (y_t) with the smallest (absolute) value. That is, the response data will be uninformative about one of the coefficients. ACA uses a similar procedure to avoiding asking questions about attributes that are unimportant to the respondent. Assume further that each respondent then provides answers to six pairwise comparisons corresponding to product profiles of the remaining attributes and that the distribution of heterogeneity is multivariate normal with mean $\bar{\beta} = (6, 3, 1)'$ and covariance matrix $V_\beta = I_4$. Finally, we assume that the self-explicated data are not used in the analysis.

The resulting posterior means of $\bar{\beta}$ based on 1,000 respondents are equal to $(6.23, 3.59, 1.67)$, with posterior standard deviations equal to $(0.06, 0.07, 0.14)$. Thus, the relative importances of the attributes are distorted. Moreover, the presence of nonzero heterogeneity covariance can be shown to increase the frequency of rank-order reversals of individual-level estimates by two to three times. It should be noted that the magnitude of these inconsistencies are functions of the unobserved parameters. Thus, there is no way to correct for the inconsistencies without knowing the desired parameter values beforehand. When the self-explicated data are included in the analysis (i.e., by including the likelihood for these data), posterior means are equal to $(6.05, 2.93, 0.88)$ with posterior standard deviation equal to $(0.06, 0.07, 0.07)$. These results indicate that violating the likelihood principle by ignoring the self-explicated data can easily result in economic harm to managers using conjoint analysis to inform product policy. Thus, ACA's "pairs only" option, which discards the self-explicated data when making inferences, should be avoided whenever the design in the pairs section is based on self-explicated data.

ACA offers several options for analysis. Empirically, "pairs only with constraints," where the constraints are derived from the self-explicated data, seems to perform best (Sawtooth Software 2003). This method, and the analysis option "pairs plus self-explicated" (ACA/Hierarchical Bayes v2.0), where the self-explicated data and the pairs data are analyzed jointly, can be shown to render ACA's adaptive design mechanism ignorable. For "pairs only with constraints," the self-explicated data are assumed to provide ordinal-level information without error that leads to ordinal constraints in the analysis of the pairs data (see Allenby et al. 1995). The "pairs plus self-explicated" analysis assumes that the data are generated from the same underlying model. Both analysis options condition on the self-explicated data, and the manner in which ACA adaptively generates questions is ignorable.

5.   Concluding Comments

The likelihood principle is implicit to the Bayesian approach to statistics, where the posterior distribution is derived from the prior distribution and the likelihood. Bayesian analysis conditions on the data to draw inferences about unobservable parameters in the analysis. In a conjoint analysis, it provides an answer to the question "Given the data at hand, what do I know about the part-worths?"

Sampling experiments are useful for understanding statistical properties such as bias when data have not yet been collected. They are not generally useful, however, once data are available for analysis. We demonstrate that the endogeneity identified by HT is irrelevant for analysis once data have been collected. A manager at a user firm who wants to make inferences about part-worths based on a specific data set should not worry about endogeneity created by adaptive questioning so long as the mechanism of choosing design points is ignorable. This will be the case if inference proceeds conditional on the data used to pick design points. Prior to data collection, however, we believe that adaptive design procedures should be evaluated based on risk.

Whether one should condition on the data and adhere to the likelihood principle, as in Bayesian analysis, or conduct sampling experiments, as in a frequentist analysis, is at the philosophical core of statistical inference. The Bayes-frequentist debate is sometimes dismissed as irrelevant because one can obtain about the same answer in some special cases (e.g., a normal likelihood in combination with diffuse priors). The issue of endogeneity bias raised by HT provides a counterexample to this view, where philosophical principles of inference play an important role in conducting analysis. We believe that analysis should condition on the data that is, strictly speaking, only possible if the mode of inference is Bayesian.

The advantage of taking on this orientation is that it greatly simplifies analysis while providing a coherent framework for inference.

References

Allenby, G. M., N. Arora, J. L. Ginter. 1995. Incorporating prior knowledge into the analysis of conjoint studies. *J. Marketing Res.* **32** 152–162.

Berger, J. O., R. L. Wolpert. 1984. *The Likelihood Principle.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, Instititute of Mathematical Statistics, Hayward, CA.

Chaloner, K., I. Verdinelli. 1995. Bayesian experimental design: A review. *Statist. Sci.* **10**(3) 273–304.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London, Ser. A* **222** 309–368.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Rubin. 2004. *Bayesian Data Analysis.* Chapman & Hall/CRC, Boca Raton, FL.

Hauser, J. R., O. Toubia. 2005. The impact of utility balance and endogeneity in conjoint analysis. *Marketing Sci.* **24** 498–507.

Heckman, J. J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Ann. Econom. Soc. Measurement* **5** 475–492.

Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lutkepohl, T.-C. Lee. 1988. *Introduction to the Theory and Practice of Econometrics*, 2nd ed. John Wiley & Sons, New York.

Manchanda, P., P. E. Rossi, P. K. Chintagunta. 2004. Response modeling with nonrandom marketing-mix variables. *J. Marketing Res.* **41** 467–478.

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

Rubin, D. B. 1976. Inference and missing data. *Biometrika* **63** 581–592.

Sawtooth Software. 2003. ACA/Hierarchical Bayes v2.0, Technical paper, http://www.sawtoothsoftware.com/technicaldownloads. shtml#acahbtech.

Zanutto, E. L., E. T. Bradlow. 2006. Data pruning in consumer choice. *Quant. Marketing Econom.* **4**(3) 267–287.