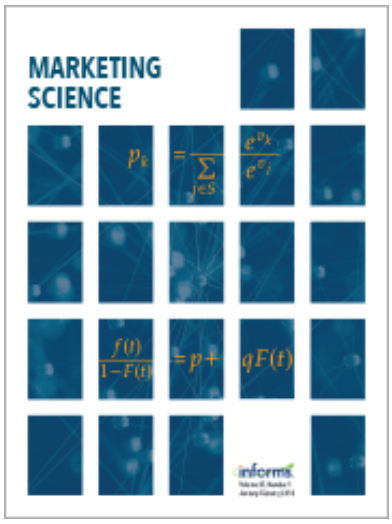


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Invited Paper—Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications

Peter E. Rossi

To cite this article:
Peter E. Rossi (2014) Invited Paper—Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications. Marketing Science 33(5):655-672. https://doi.org/10.1287/mksc.2014.0860

MARKETING SCIENCE
Vol. 33, No. 5, September–October 2014, pp. 655–672
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2014.0860

Invited Paper

Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications

Peter E. Rossi
Anderson School of Management, UCLA, Los Angeles, California 90095,
perossichi@gmail.com

Marketing is a field that is rich in data. Our data is of high quality, often at a highly disaggregate level, and there is considerable variation in the key variables for which estimates of effects on outcomes such as sales and profits are desired. The recognition that, in some general sense, marketing variables are set by firms on the basis of information not always observable by the researcher has led to concerns regarding endogeneity and widespread pressure to implement instrumental variables methods in marketing problems. The instruments used in our empirical literature are rarely valid and the IV methods used can have poor sampling properties, including substantial finite sample bias and large sampling errors. Given the problems with IV methods, a convincing argument must be made that there is a first order endogeneity problem and that we have strong and valid instruments before these methods should be used. If strong and valid instruments are not available, then researchers need to look toward supplementing the information available to them. For example, if there are concerns about unobservable advertising or promotional variables, then the researcher is much better off measuring these variables rather than using instruments (such as lagged marketing variables) that are clearly invalid. Ultimately, only randomized variation in marketing variables (with proper implementation and large samples) can be argued to be a valid instrument without further assumptions.

Keywords: endogeneity; instrumental variables; pricing; promotion; advertising
History: Received: October 29, 2013; accepted: April 17, 2014; Preyas Desai served as the editor-in-chief and Bart Bronnenberg served as associate editor for this article. Published online in *Articles in Advance* July 1, 2014.

1. Introduction

One of the principle attractions of marketing as a field for applied empirical work is the large quantity of high quality data. Scanner and Internet panels of a wide variety generate disaggregate data of a high quality. For example, the Nielsen Homescan panel data available from the Kilts Center, Booth School of Business, provides transaction level data on more than 50,000 households in the United States. These households record purchases using optical scanning equipment, assuring a high quality of data on purchases. A panel of more than 30,000 stores (the RMS data also available from the Kilts Center) provides weekly sales and prices of all Universal Product Codes for drug, grocery, mass merchant, and convenience stores. The social networking site, Twitter, provides an API that allows for direct access to a very rich source of text data. In fact, a persistent and resourceful researcher can generally gain access to virtually any industry-generated data set. Not only is this data well measured by the standards of applied work in economics but there is also considerable variation in key marketing variables such as price and promotion in this data.

In short, we are awash with abundant and potentially informative data in marketing applications. However, except in the cases of experimentally generated data,[1] the data we are so lucky to have is generated by passive observational methods. For this reason, there is a legitimate concern of how to generate causal inferences using standard conditional methods such as regression analyses. This concern is often not very precisely articulated, but rather is expressed as a concern for endogeneity biases. One traditional solution to the problem of endogeneity bias is to use Instrumental Variable (IV) methods. These methods, by definition, 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 that uses all of the variation in the data. IV methods are only asymptotically unbiased if

[1] While field experiments in marketing are relatively rare in the academic setting, all conjoint data is fundamentally experimental data which is devoid of so-called endogeneity problems by design.

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 do not 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 (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 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.[2] If there is measurement error in the independent variables, then the FE approach will generally magnify the errors-in-the-variables bias.

In §2, I will review omitted variable bias and recall the interpretation of endogeneity bias in terms of omitted variables. In §3, I will discuss the key assumptions used in IV methods as well as the difficulties with the finite sample distribution of IV estimators. Section 4 provides a brief summary of the last 10 years of empirical research that uses IV methods and will emphasize that IV methods are being applied with invalid instruments.

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

Throughout this discussion of IV and related methods, I identify not only generic issues but also consider issues unique or of highest relevance to marketing. For example, it is difficult to identify valid and strong instruments for advertising and promotional variables.

2. The Omitted Variables Interpretation of Endogeneity Bias

2.1. Omitted Variable Bias

In marketing applications, the omitted variable interpretation of endogeneity bias provides a very useful intuition. In this section, I will briefly review the standard omitted variables analysis and relate this to endogeneity bias. For those familiar with the omitted variables problem, this section will simply serve to set notation and as a very brief review (see, also treatments in Woolridge 2010, §4.3 or Angrist and Pischke 2009, §3.2.2). Consider a linear model with one independent variable (note: I have removed the intercept for notational simplicity)

$$y_i = \beta x_i + \varepsilon_i. \tag{1}$$

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 $\mathbb{E}[\varepsilon \mid x] = 0$ (or $\text{cov}(x, \varepsilon) = 0$)

$$\begin{aligned}\text{plim}\frac{x'y}{x'x} &= \beta + \text{plim}\frac{(x'\varepsilon)/N}{(x'x)/N} = \beta + \text{plim}\left(\frac{x'x}{N}\right)^{-1}\text{plim}\frac{x'\varepsilon}{N}\\ &= \beta + Q \times \text{cov}(x, \varepsilon).\end{aligned}$$

Here $Q^{-1} = \text{plim}(x'x)/N$. Thus, least squares will consistently estimate the structural parameter β only if (1) can be considered a valid regression equation (with an error term that has a conditional expectation of zero). If $\mathbb{E}[\varepsilon \mid x] \neq 0$, 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

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

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

$$\begin{aligned}\mathbb{E}[y \mid x] &= \beta x + \mathbb{E}[\gamma w + \varepsilon \mid x] = \beta x + \gamma\,\mathbb{E}[w \mid x]\\ &= \beta x + \gamma\pi x = \delta x.\end{aligned}$$

Here π is the coefficient of w in the conditional expectation of $w \mid x$. 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 nonzero 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 variations 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.

2.2. 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

$$y_i = \beta x_i + \alpha_y \nu_i + \varepsilon_{y,i}, \tag{2}$$

$$x_i = \alpha_x \nu_i + \varepsilon_{x,i}. \tag{3}$$

Here both ε_x, ε_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 (3) 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

$$\begin{aligned}\mathbb{E}[y \mid x] &= \beta x + \mathbb{E}[\alpha_y \nu + \varepsilon_y \mid x] \\ &= \beta x + \alpha_y \alpha_x \left(\frac{\sigma_\nu^2}{\alpha_x^2 \sigma_\nu^2 + \sigma_{\varepsilon_x}^2} \right) x. \end{aligned} \tag{4}$$

The ratio $\alpha_x(\sigma_\nu^2/(\alpha_x^2\sigma_\nu^2 + \sigma_{\varepsilon_x}^2))$ 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 (4). Whenever the unobservable has variation that comprises a large fraction of the total variation in x, and 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.

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 that has been omitted from the structural equation. This makes a concern for endogeneity a generic criticism that 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[3] 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 a direct way, only indirectly via x. Of course, there can be many instrumental variables. If we contend that a vector variables, z, are instrumental variables, then, in a very general sense, we are assuming that y and z are independent conditional on x, $y \perp z \mid x$.

3.1. The Linear Case

The case of a linear structural equation and linear IV model provides the intuition for the general case and includes many of the empirical applications of IV methods. However, it should be noted that due to the widespread use of choice models in marketing applications, there is a much higher incidence of the use of nonlinear models. I consider nonlinear choice models in §3.5. Equations (5) and (6) constitute the linear IV model

$$y = \beta x + \gamma' w + \varepsilon_y, \tag{5}$$

$$x = \delta' z + \varepsilon_x. \tag{6}$$

Equation (5) 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 that are exogenous in the sense that we assume that $\mathbb{E}[\varepsilon_y \mid w] = 0$. 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 $\mathbb{E}[\varepsilon_y \mid x, w] = f(x) \neq 0$, or that we still have an endogeneity problem.

The second Equation (6) is 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, $\mathbb{E}[\varepsilon_y \mid x, w] \neq 0$, is equivalent to $\mathrm{corr}(\varepsilon_x, \varepsilon_y) \neq 0$. 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 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 enter both equations is not identified.

3.2. Method of Moments and 2SLS

There are a number of ways to motivate inference for the linear IV model in (5) and (6). The most popular is the method of moments (MM) 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 MM 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 involves an assumption about the population or data generating model that $\mathbb{E}[\varepsilon_y z] = 0$. The MM principle defines an estimator by minimizing the discrepancy between the population and sample moments

$$\hat{\beta}_{\mathrm{MM}} = \arg\min_{\beta} \| \mathbb{E}[\varepsilon_y z] - (y - \beta x)' z \| = \frac{z'y}{z'x}. \tag{7}$$

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 $\mathbb{E}[\varepsilon_y z] = 0 = \mathrm{plim}((z'\varepsilon_y)/N)$) and asymptotically normal. If the structural equation errors are uncorrelated and homoskedastic, it can be shown (see, for example, Hayashi 2000, §3.8) that the particular MM estimator in (7) is the optimal Generalized Method of Moments (GMM) 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

[3] For simplicity, I will consider the case of only one right-hand side (rhs) 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.

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 MM estimator has better sampling properties than another 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 2SLS 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)

$$x = \hat{x} + e_x = \hat{\delta} z + e_x, \tag{8}$$

$$y = \hat{\beta}_{2\text{SLS}} \hat{x} + e_y. \tag{9}$$

This procedure yields the identical estimator as the MM estimator in (7).

If there is 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 is accounted for by the instruments and remains the same.

3.3. 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 estimating 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 that 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

$$y = \hat{\beta}_{\text{TSLS}} x + c e_x. \tag{10}$$

e_x is the residual from (8).

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. Because, in nonlinear 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, $\mathbb{E}[x \mid z] = f(z)$. 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 I 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; the standard errors produced by the standard nonlinear models software will be incorrect because they do not take into account that some of the variables are constructed. It is not clear 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.

3.4. Sampling Distributions

The ordinary least squares (OLS) estimator (conditional on the X matrix) is a linear estimator with a sampling distribution derived from the distribution of $\hat{\beta}_{\text{OLS}} - \beta = (X'X)^{-1}X'\varepsilon$. 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 nonlinear functions of the data. For example, the simple MM estimator (7) 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 nonlinearly, via the term $(z'x)^{-1}$, 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 (5) and (6). The sampling distribution of the IV estimator is approximated by using asymptotic methods. This is done by normalizing by $\sqrt{N}$ and applying a Central Limit Theorem (CLT)

$$\sqrt{N}(\hat{\beta}_{\mathrm{MM}} - \beta) = \left(\frac{z'x}{N}\right)^{-1} \sqrt{N}\frac{z'\varepsilon_y}{N}. \tag{11}$$

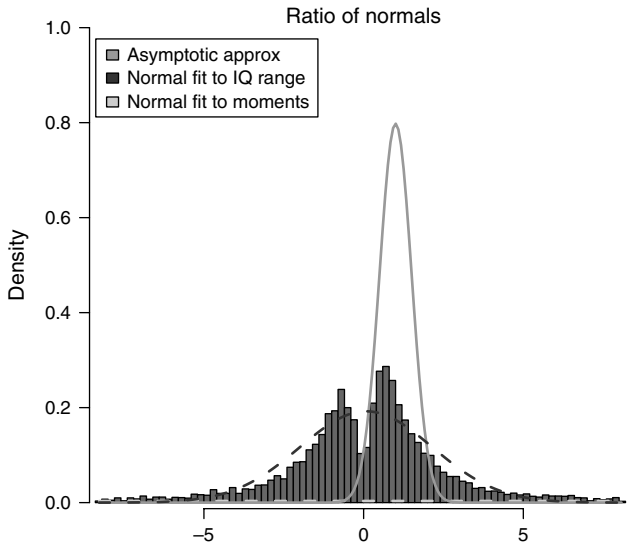
As N approaches infinity, the denominator of the MM estimator, $(z'x)/N$, converges to a constant by the Law of Large Numbers. The asymptotic distribution is entirely driven by the numerator, which has been expressed as $\sqrt{N}$ times a weighted sum of the error terms in the structural equation. The asymptotic distribution is then derived by applying a CLT 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 so 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 a 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 MM estimator is a ratio of a weighted average of y to the weighted average of x

$$\hat{\beta}_{\mathrm{MM}} = \frac{(z'y)/N}{(z'x)/N}.$$

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 the MM 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.

Figure 1 Distribution of a Ratio of Normals



To illustrate how ratios of normals can fail to have a normal distribution, consider the distribution of a ratio of a $N(1, 0.5)$ to a $N(0.1, 1)$ random variable.[4] The distribution is shown by the histogram in Figure 1 and is revealed to be bimodal with the positive mode having slightly more mass. This distribution exhibits outliers; the figure only shows the histogram of the data trimmed to remove the top and bottom 1% 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 light grey density curve in the figure. Clearly, this is a poor approximation that ignores the other mode and underestimates variability. The dashed light grey 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

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

by the dark dashed curve which is fit to the observed InterQuartile range. Even this approximation misses the bimodality 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 nonzero 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 the next section, we will explore what are the boundaries of the weak instrument problem.

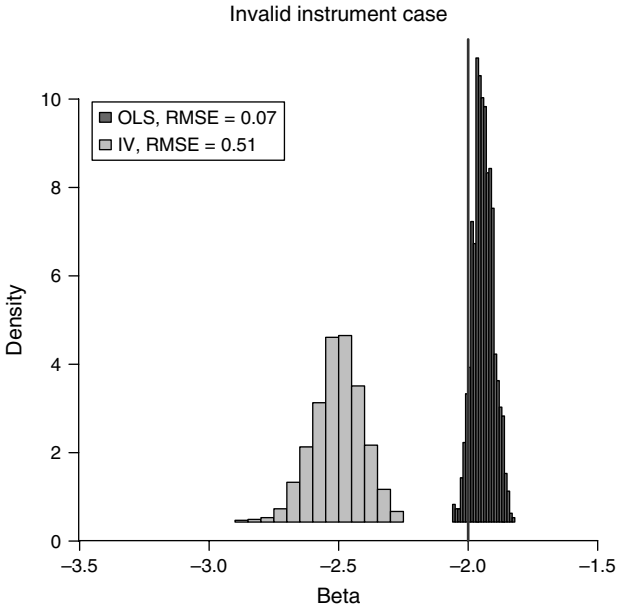
One point that is absent from the econometrics literature is that the sampling distribution of IV *estimators* is only considered *conditional* on the validity of the instruments. This is an untestable assumption which certainly is violated in many data sets. This form of misspecification is much more troubling than other forms of model misspecification such as non-normality of the error terms, conditional heteroskedasticity or nonlinearity. For each of these misspecification problems, we have tests for misspecification and alternative estimators. There are also methods to provide inference (i.e., standard errors and confidence intervals) that are robust to model misspecification for conditional heteroskedastic, auto-correlated, and non-normal errors. There are no methods that 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. I simulated data from the following model:

$$\begin{aligned} y &= -2x - z + \varepsilon_y, \\ x &= 2z + \varepsilon_x, \end{aligned}$$

$$\begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}\right); \quad z_i \sim \text{Unif}(0,1). \quad (12)$$

This is a situation with a relatively strong instrument (the population R-squared for the regression of x on z is about 0.25). I set $N = 500$, which is a large sample

Figure 2 Sampling Distributions of Estimators with Invalid Instruments



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. Figure 2 shows the sampling distribution of the MM estimator and the standard OLS estimator. Both estimators are biased and inconsistent. Moreover, the IV estimator has inferior sampling properties with a RMSE of more than seven times the OLS estimator. Because we cannot 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.

3.5. The Weak Instruments Problem

Linear Models. Not only are instruments potentially invalid, there is also 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, 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 subliterature 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 do not have more than a handful of potential instruments. One could also argue that truly weak instruments are rarely encountered in applications in marketing. This depends on what you mean

by weak and at what point standard asymptotics start to break down.

However, there is also a problem with strong instruments. If your instruments explain a great deal of the variation in x (a first-stage R-squared of more than 90%), then there cannot be an appreciable endogeneity problem in the first place. Most of the variation of x will not be driven by the unobserved quantity but by the instrument that is assumed to be exogenous. On the other hand, if there is a forceful argument that there truly is an endogeneity problem in your data, then it is hard to imagine how it would be possible to find both a strong and valid instrument.

Returning to the problem of weak instruments, there is a view by applied econometricians that failure of standard asymptotics 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% 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 are not finite sample biases of an even greater magnitude than these asymptotic biases.

The poor sampling properties of the IV estimator[5] can easily be shown even in cases where the instruments have a modest but not small first-stage R-squared. We simulate from the following system:

$$y = -2x + \varepsilon_y,$$
$$x = Z\delta + \varepsilon_x,$$
$$\begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} \sim N\left(0, \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}\right).$$

$N = 100$. Z is a $N \times p$ 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 0.10 using the formula, $\sqrt{(12\rho^2)/(p(1-\rho^2))}$, where ρ^2 is the desired R-squared value. Figure 3 shows the sampling distribution of the IV and OLS estimators in this situation with $p = 1$. The MM 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. Figure 4 provides the distribution of the first-stage F statistics for 2,000 replications. The vertical line in the figure is the median. This means that more than 50% 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, the IV estimator would require a sample size approximately four[6] times larger than the OLS estimator to deliver the same RMSE level.

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

Figure 3 **"Weak" Instruments Sampling Distribution: $p = 1$**

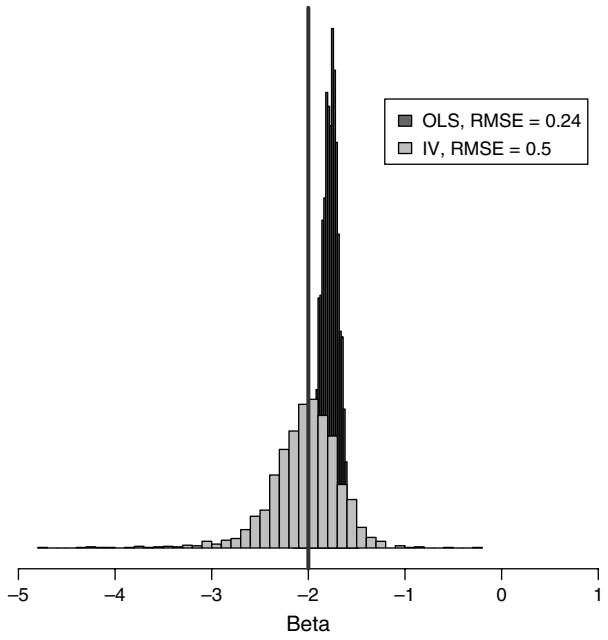
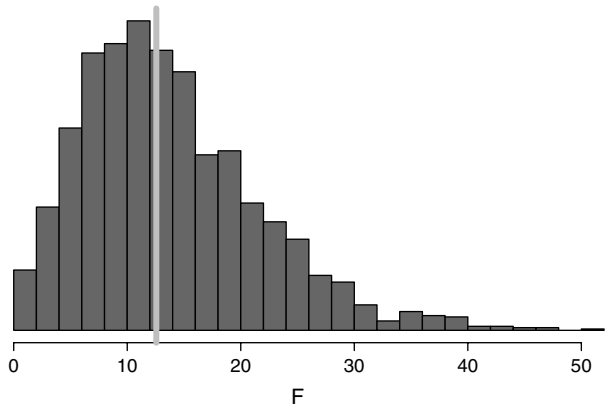


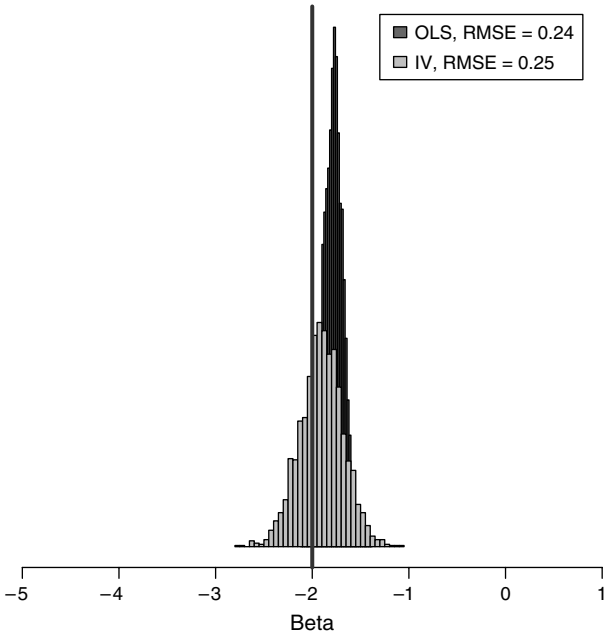
Figure 4 **Distribution of First-Stage F-Statistics**



[5] 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.

[6] $(0.5/0.24)^2$.

Figure 5 "Weak" Instruments Sampling Distribution: $p = 10$



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 that 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 in 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 (CLR) statistic and builds a testing theory that 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 that 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 (see, for example, 7) 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.

Choice Models. Much of the applied econometrics done in marketing uses 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 applies to a nonlinear 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 also 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 in 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

$$\Pr(j \mid t) = \frac{\exp(\alpha' c_{j,t} + \beta' m_{j,t} + \xi_{j,t})}{\sum_j \exp(\alpha' c_{j,t} + \beta' m_{j,t} + \xi_{j,t})}. \qquad (13)$$

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 $\xi_{j,t}$. Because 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 conditionally independent of $\xi_{j,t}$,

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

z_t is the vector of the p instrumental variables. As a practical matter, this means that I 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, I will consider the choice model in (13) along with a standard linear IV first-stage equation

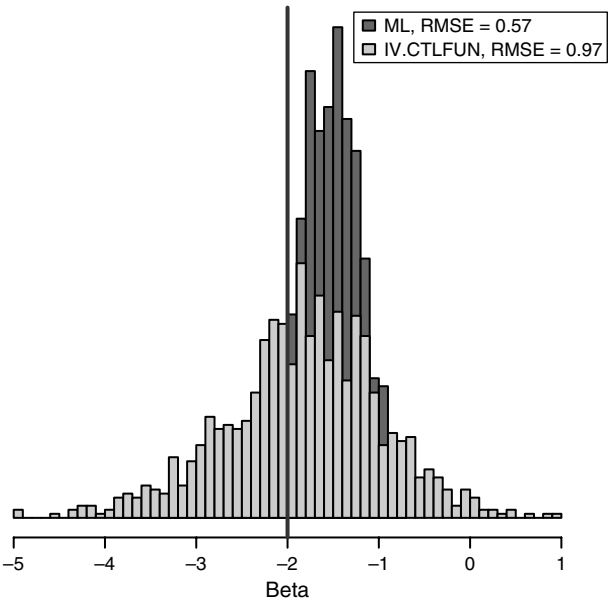
$$m_t = z_t\Delta + \nu_t,$$

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, I 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, I will assume that the unobserved demand shocks joint normal with the errors in the IV equation

$$\Pr(1) = \frac{\exp(\alpha+\beta m+\xi)}{1+\exp(\alpha+\beta m+\xi)},$$
$$m = \delta' z + \nu,$$
$$\begin{pmatrix}\xi\\ \nu\end{pmatrix} \sim N\left(0, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right).$$

Here I have arbitrarily written the probability of choice alternative 1. I used the same parameters as in the simulation of the weak instruments problem for the linear IV with one instrument, $N = 100$, $\rho = 0.25$, and δ is set for that first-stage R-squared is 0.10. Figure 6 shows the sampling distribution of the standard maximum likelihood estimator (MLE) which ignores the endogeneity (shown by the darker histogram). I used a control-function approach to compute the IV estimator for this problem under the assumption of a linear first stage: I regressed the endogenous rhs variable m, on the instruments and used the residual from this regression to create additional explanatory variables that were included in the logit model. In particular, I 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 light grey 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 an 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.

Figure 6 Sampling Distributions for Logit Models



3.6. 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 an 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 that 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.

4. What We Are Doing to Ourselves

In this section, I review some of the classic examples of instrumental variables of particular relevance to marketing applications and review the last 10 years of published research in marketing.

4.1. A Short Literature Review

I surveyed 10 years of two leading marketing journals. *Marketing Science* (volume 23 (2003), issue 1 through volume 32 (2013), issue 5) and *Quantitative Marketing and Economics (QME)* (volume 1 (2003), issue 1 through volume 10 (2012), issue 3) contain a total of 533 (*Marketing Science*) + 140 (*QME*) = 674 articles. Of these 674 articles, 463 (362 in *Marketing Science* and 101 in *QME*) were empirical in nature. Of these, 46 (29 in *Marketing Science* and 17 in *QME*) used IV methods of some kind. Here I took a very broad view of IV methods, which would also include methods specifying a full system of both demand and supply equations. The central idea is that these articles are screened to use instruments via some sort of exclusion restriction.

I summarize those variables considered endogenous in the 46 total articles (note that some articles include more than one endogenous variable; hence, the total is more than 46) in order of frequency:

- price (24),
- advertising (8),
- promotions (3),
- entry order (3),
- distribution (2),
- market structure (2),
- market share (2),
- revenue (2),
- networks (2), and
- other (11).

The instruments used are:

- lagged variables (18),
- costs—including input prices and wholesale prices (13),
- fixed effects—including interactions, brand dummies, time dummies (9),
- Hausman style variables from other markets (6),
- demographics (5),
- product characteristics a la Berry, Levinsohn, and Pakes (3),
- price indices (3),
- display and feature (2), and
- other (3).

Thus, the modal IV analysis is designed to address concerns regarding price endogeneity and tends to use lagged variables as instruments. Less than one half of the papers reported some sort of measure of the strength of the instruments such as the *R*-squared of a regression of the endogenous variable on the set of instruments. For those papers, where the first-stage *R*-squared is reported, the authors fail to report the incremental *R*-squared of the instruments. That is, the first stage-regressions reported included all exogenous variables in the analysis, including fixed effects and interactions. In these cases, the *R*-squared of the first-stage regression can be quite high even though the incremental contribution of the instruments to *R*-squared is very low. In addition, as outlined in §4.4, if brand dummies are interacted with instruments for use in a choice model, a different approach has to be taken to measure incremental *R*-squared. The combination of the failure to report any measure of fit in a first-stage analysis along with the failure to report the true incremental fit for those papers that include the first-stage *R*-squared means that we do not really know the strength of the instruments used in empirical work in marketing.

Even more disturbing is the virtual lack of discussion of why the instruments chosen by the authors are valid instruments. Discussion of the validity of the instruments is found in less than 10 of the IV papers. This occurs in spite of the fact that there is no way to test the validity of instruments and, thus, the only recourse is an economic argument regarding why a specific instrument is exogenous to the determinants of demand. Several authors appear to believe that the Hausman test[7] can be used to test the validity of instruments. The Hausman test can only be used to assess the validity of one set of instruments conditional on the validity of another set.

Finally, there are a number of papers that use demographics and price indices as instruments. The fundamental notion is that the validity of an instrument rests on an exclusion restriction. For example, in a demand model, we must make the argument that a valid instrument has no direct effect on demand, but rather works indirectly by influencing the marketing mix variables. It is hard to imagine how one could ever make the argument that demographic variables can be excluded from the demand equation. There is no discussion of the validity assumption in these papers and it appears that the authors believe any co-variate will suffice as an instrument.

4.2. 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). In particular, there has been a great deal of emphasis on discrete choice models of demand for differentiated products (for an overview, see Ackerberg et al. 2007, pp. 4178–4204). Typically, these are simple logit models

[7] See, for example, Hayashi (2000, pp. 232–233).

that allow marketing mix variables to influence demand and account for heterogeneity. The 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

$$MS(j\,|\,t) = \int \frac{\exp(\alpha' c_j + \beta' m_{jt} + \xi_{jt})}{\sum_{j=1}^{J} \exp(\alpha' c_{jt} + \beta' m_{jt} + \xi_{jt})} p(\alpha, \beta)\, d\alpha\, d\beta. \quad (14)$$

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 that represent 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 catalogue the sorts of instruments used in the literature.

Price Endogeneity. Equation (14) 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 that 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 arguments 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), we would expect that firms would set price as a function of ξ_{jt} as well as of the observed characteristics. This is 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 appear to be suboptimal 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 for 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.[8] 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 do not 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 that do not vary over time, then simply including market FE would eliminate the need for instruments. One has to argue that the unobserved demand shocks vary 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

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

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; 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. However, the FE approach is only available for linear models.

Price Endogeneity in Models of Discrete Choice with Panel Data. Villas-Boas and Winer (1999) make the point that even individual level choice models could be subject to price endogeneity. They advocate the use of lagged prices as instruments. This work has been revealed to be very influential on the practice of IV methods in marketing because lagged prices are often used in empirical work. Villas-Boas and Winer (1999) use panel data on the purchases of yogurt and ketchup. The endogeneity problem is caused by a possible contemporaneous correlation between prices and an unobservable demand shock which is modeled as uncorrelated across purchase occasions. The time scale in these data sets is weekly. Villas-Boas and Winer (1999) do not justify the validity of lagged prices as instruments other than to say that "a main advantage of using lagged prices as instruments is that they are readily available to the researcher" (p. 1327). Convenience is not a compelling argument that lagged prices are valid instruments.

Researchers have struggled to explain exactly what are these brand-specific demand shocks that are uncorrelated and observed at a weekly or even daily frequency. Some have suggested that these shocks represent the effects of unobserved advertising or promotional campaigns. In panel data sets in marketing, we typically have local store advertising variables such as feature and display in the model. This means that the unobservables in a choice model would have to be effects from unobserved mass advertising or manufacturer-based consumer promotions such as coupon drops. The problem here is that these activities are not planned at the weekly or daily frequency and there are very few instances of these activities in any given planning horizon such as a quarter or a year. If the unobservable aggregate demand shock is really driven by market-wide promotion or advertising, then we would expect that the variance of these shocks is very small and that these shocks would be highly correlated over time at the daily or weekly frequency.

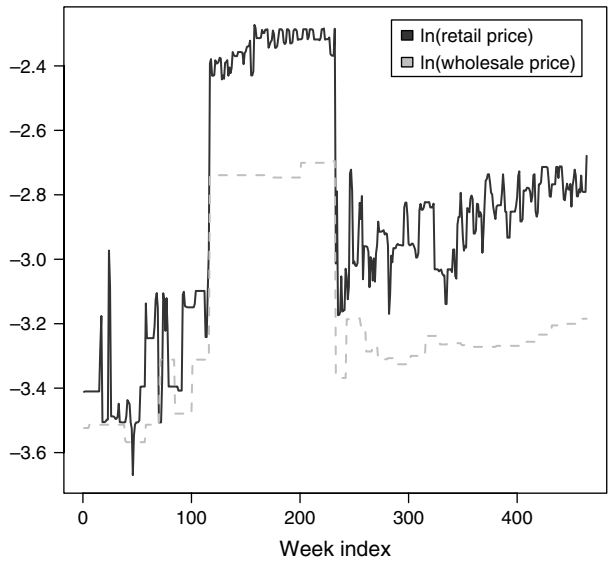
One important unobservable in scanner panel data on frequently purchased nondurable products is household inventories. If there are frequent sales, then household inventories today are almost surely related to prices yesterday (that is, if prices were low yesterday, then household inventories are probably higher than normal today). In this sort of consumer world, lagged prices are certainly invalid and would create substantial asymptotic biases in our estimates of the causal price effect even compared to standard least squares or ML, which do not consider the possibility of endogeneity.

Setting concerns about the validity of lagged prices as instruments aside, a legitimate point is that Villas-Boas and Winer (1999) find very large endogeneity effects in the sense that model parameters and price elasticities change substantially with and without the use of lagged price instruments. However, their model lacks heterogeneity and state-dependence, a well documented feature of packaged goods panels data (Allenby and Rossi 1999, Dubé et al. 2010). The correlation between the pricing equation error and the demand shocks found by Villas-Boas and Winer (1999) could simply be due to state dependence.

Wholesale Prices. While it is difficult to make a convincing argument that lagged prices are valid instruments, the arguments in favor of the use of wholesale prices as instruments are stronger. Standard models of the supply side for a retailer would suggest that wholesale prices are a valid instrument (however, we are still left with the question of what these demand shocks actually represent over time in the same market). Wholesale prices also vary by brand, eliminating one of the major problems with input or factor price series. Arguments against the use of wholesale prices are that manufacturers anticipate advertising and promotional events in setting wholesale prices.

For consumer packaged goods (CPG) products, there is an argument that adjustments in wholesale prices are made by the manufacturer via frequent trade promotions. These trade promotions often include temporary reductions in wholesale prices. That is, the retailer and manufacturer work together to monitor the market and the manufacturer adjusts to market conditions by varying the acquisition cost via trade discounts of various kinds. For some products, wholesale list prices are never reduced but the effective price is almost certainly lower due to frequent trade promotions. In most wholesale price databases, trade promotions are not netted from wholesale prices but are recorded separately. If the researcher attempts to compute the

Figure 7 Retail and Wholesale Margarine Prices



true acquisition cost of the products sold by the retailer, these constructed transactional wholesale prices are almost certainly not exogenous to demand shocks and would be invalid instruments.

Yet the problem with using wholesale prices is not only potential invalidity as an instrument, it is also the interpretation of response to changes in wholesale prices versus response to changes in retailer prices. Figure 7 displays weekly retail and wholesale price series for a margarine product obtained from Chintagunta et al. (2005). Wholesale prices are much less variable than retail prices as many retailers have frequent sales that involve large reductions in price for short periods of time. Wholesale price changes typically are associated with the much less frequent jumps in the regular or base price. Thus, if we were to use wholesale prices as instruments, we are projecting the highly variable retail price series on a much smoother series whose variation is at a much lower frequency. We can think of the part of the variation associated with the wholesale price instrument as the long-run variation in prices. Thus, if we were to contrast price effects from using retail price without instruments to the price effects estimated using wholesale price as an instrument, we would be essentially identifying the difference between long- and short-run price changes. Long-run price changes generally have lower price response due to the small set of substitution possibilities. Short-run price changes have larger price responses because of the possibility of forward buying and storage (a point emphasized by Erdem et al. 2003). The fact that the use of wholesale prices as instruments changes the estimated price elasticity may be due to the difference between long-run and short-run effects rather than endogeneity.

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 data sets 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. I am not aware of any economic arguments that can justify the use of lagged prices as instruments. In summary, I believe that, with panel or time-series data, the emphasis on price endogeneity has been misplaced.

4.3. 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 include what amounts to local forms of advertising from display signage, direct mail, and free standing inserts. In the pharmaceutical and healthcare products industry, large and highly compensated sales forces call on doctors and other healthcare 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.

4.4. Special Considerations for Choice Models

In any instrumental variables application, we require that the number of excluded (instrumental) variables be at least equal to the number of rhs endogenous variables. For example, consider a simple linear demand model

$$\ln S_t = \alpha + \beta_1 \ln P_{1,t} + \beta_2 \ln P_{2,t} + \varepsilon_t. \tag{15}$$

Here we are modeling the sales of a product (denoted by the subscript 1) as a function of own price and the price of a substitute competing product, and (2) as a standard log-linear model. If there is a concern that both price variables are endogenous (i.e., correlated with the error term due to omitted market-wide unobservables), then we require at least two instrumental variables. That is, we need two separate sources of variation, one that moves the price of product 1 and another with some independent variation in the price of product 2 (that is, the sources of variation can be correlated but not perfectly correlated). If we only have one candidate instrumental variable, then the model is unidentified. If we consider, for example, a standard supply side argument that the input or factor costs are valid instruments, we would have to acquire cost shifter data which is product specific. That is, we need shifts in the cost of product 1 that are not perfectly correlated with shifts in the costs of product 2. Standard input factor prices are not typically product specific. For example, suppose product 1 is regular Coke and product 2 is regular Pepsi. An input factor price like the price of sugar would not be sufficient to identify the separate effects of each price on the sales of Coke; we would expect that as sugar prices rise, the costs of manufacturing Coke and Pepsi increase proportionately. What we really need is some sort of cost shifter that would make it relatively more expensive to produce Coke than Pepsi. For example, if Coke contained high fructose corn syrup (HFCS), while Pepsi contained sugar as the sweetener, then we could use as instruments the two time series of HFCS and sugar prices.

Another way of seeing this is to consider a restricted version of the model in Equation (15) in which demand is driven only by the relative price of Coke to Pepsi, i.e., restrict $\beta_1 = -\beta_2$

$$\ln S_t = \alpha + \beta \ln\left(\frac{P_{1,t}}{P_{2,t}}\right) + \varepsilon.$$

For this model, we need only one instrument as we have reduced the number of parameters. However, the instrument must change the relative prices of the two products. In the Coke/Pepsi example with both using HFCS as sweeteners, the HFCS time series would not help identify the restricted model. That is, if both use HFCS, then as HFCS prices rise, both the price of Coke and Pepsi would rise proportionately and we would expect that the first-stage regression of the log of relative prices on the HFCS price series would have a very low R-squared indicating an exceptionally weak instrument. However, if it were the case that Pepsi was dramatically sweeter than Coke and used much more HFCS, then we might expect the HFCS price series to be a more powerful instrument that could change the relative price of Coke to Pepsi.

Choice models, by their very definition, are about the relative utility of various choice options and not the absolute value of utility. This means that it is the relative values of marketing mix variables such as price or advertising that matter for these models. A cost-based instrument that is non-brand specific cannot even be used in a standard linear IV manner (i.e., project the vector of stacked prices on the cost instrument) as this would simply drop out of the choice probabilities (at least in a logit formulation). What has to be argued is that the cost-based instrument affects the mark-ups for each brand in a different manner

$$p_j = f_j(z). \tag{16}$$

Here z is the instrument that does not vary over brands; but the mark-up function does. Again, this has to be derived from the production process. An argument has to be made that the instrument will shift around the relative prices of the products in the choice model. Shifting the level does not help identify any of the choice model parameters. It can be argued that most cost-based instruments primarily influence the level of prices, not relative prices and, therefore, are likely to be very poor instruments for a choice model.

In most applications, authors put in the interaction between the brand intercepts and the instrument in very much the same manner as one would enter a nonchoice alternative specific variable such as income. That is, we include a separate utility coefficient for every choice. This means that the income of the consumer would influence the attractiveness of different choice alternatives—a reasonable assumption for a group of products of varying quality. While this strategy is certainly feasible in a strict computational sense, we should recognize that the instrument is only useful to the extent to which it is associated with changes in relative prices, not the level of prices. This means that the power of nonbrand-specific instruments has to be judged in a different manner than simply reporting the R-squared of a first-stage linear projection of prices on the instrument interacted with brand intercepts.

We have to measure the ability of the instrument to predict relative price changes, not levels. This can be done by an incremental R-squared approach. Consider two regressions:

$$P_t = X_t\delta + v_t, \tag{17}$$

$$P_t = X_t^*\delta^* + v_t^*. \tag{18}$$

Here P_t is at the vector of the prices of the J choice alternatives. The matrices X and X^* reflect the differences between a simple projection of prices onto the instrument with and without interactions. We construct the X matrices from z_t (a scalar) as follows:

$$X_t = [\iota \quad \iota z_t],$$

$$X_t^* = \begin{bmatrix} \iota & 0' \\ & I_{J-1} \otimes z_t \end{bmatrix}.$$

To measure the strength or power of any given instrument, we must compute the incremental R-squared from the regression in (18) over the R-squared from the regression in (17). Most authors (if they report R-squared at all) only report the R-squared from (18) which can be misleading. For example, if the instrument explains a great deal of the fluctuation in the overall level of prices among the J brands, but nothing of the variation in relative prices, the reported first-stage R-squared from (18) will be very high even though the incremental R-squared will be zero. This means that some investigators are fooling themselves into believing they have strong instruments when, in fact, they have a weak instrument problem.

5. Conclusions

Endogeneity concerns are prominent in empirical research in both marketing and I/O. The notion that there exists some unobservable that is correlated with both sales/profit outcomes as well as marketing mix variables provides a powerful motivation for these concerns. In my view, these concerns are of utmost importance with cross-sectional data where the unobservable could be an unobserved product or market characteristic. With relatively high frequency time series data (such as weekly data), the notion that there exists some demand shock that varies from week to week (and possibly also from brand to brand) and that this unobservable also drives a non-negligible portion of the variation in marketing mix variables is strained. The evidence for the existence of endogeneity biases in time series or panel data consists entirely of model-based evidence via comparison of the results with and without IVs. There is no direct evidence from the firm side (for example, from pricing experiments) that endogeneity biases are large in panel or times series data.

While it is always possible to argue that unobservables correlated with marketing mix variables exist, there are other econometric problems, such as functional form and distributional assumptions, which can be viewed as equally important. Thus, I believe that a strong argument must be made that endogeneity problems are of the first order. I do not see convincing arguments along these lines in our empirical literature.

The solution to the endogeneity problem is often viewed as provided by IV methods. IVs are the extension of experimental methods to passively observed data. That is, a valid instrument is supposed to shift the marketing mix variable without directly affecting the relevant outcome measure such as demand or profits. The IV acts like a true randomized experiment in which the randomization of treatment allows for pure causal effects to be determined. The validity of the instruments depends critically on this exclusion restriction—the instrument only has an indirect effect, not a direct effect, on the outcome of interest. Contrary to the beliefs of some, it is not possible to test the assumption that an instrument is valid. This means that logical arguments must be offered as to why the advocate of an instrument believes it is valid. The strongest such arguments come from assumptions regarding firm behavior or the so-called supply side. A deep understanding of the supply side may provide guidance as to what sort of variables can be plausibly assumed to only have indirect effects on the demand side.

The empirical marketing literature has taken a cue from the empirical I/O literature and emphasized price endogeneity. In a cross-section of markets or with a cross-section of products, there are strong arguments that price endogeneity could be a first order problem. However, once a time series dimension is added, the arguments for IV methods become less strong. It is easy to add brand/market fixed effects that should remove the major price endogeneity problems without the need for instruments. Consideration of the supply side would suggest that input or factor prices might be logically valid instruments. However, these are not useful in marketing applications as we typically cannot find brand- or product-specific cost data and the time series variation of many factor prices is very limited. Wholesale prices are possibly more useful as they vary more over time and are brand specific. For many CPG products, wholesale prices can be obtained with some effort. However, the problem with projecting retail prices on wholesale prices is that we have now changed the price effect that we are estimating. If we use retail prices without instruments, we are estimating primarily the effects of short-run changes in prices. If we employ wholesale prices as instruments, then we are estimating long-run price effects. So by using wholesale prices as instruments we are not just attempting to correct

for endogeneity bias but we have also changed the structural quantity that is being estimated.

Rather than focusing on endogeneity problems with respect to price, it seems that stronger arguments could be made that advertising and promotional decisions require attention. Here endogeneity can be driven by unobservable demand shocks but also by the optimizing behavior of the firm. If, for example, the firm is allocating promotional budgets (such as a direct sales force or a trade promotional budget) across accounts, we would expect that the firm would make those allocations with at least partial knowledge of the account-level response parameters. In this case the unobservable is actually a response parameter and a different sort of econometric approach is required (see, for example, the approach advocated by Manchanda et al. 2004). In the case of the standard demand shock unobservable, instruments for advertising and promotional marketing mix variables are required for an IV approach. There are few practical and valid candidates for instruments with respect to these variables, further limiting the usefulness of the IV approach.

A large number of papers in marketing have turned to the use of lagged variables as instruments, in particular lagged marketing mix variables such as price. I know of no economic arguments that can justify this practice. Papers that employ lagged marketing mix variables do not even discuss the question of whether these are valid instruments. I recommend that IV methods not be used if the only instruments available are lagged marketing mix variables.

The econometric analysis of IV methods rests entirely on the use of asymptotic approximations. All of the approximations assume that the instruments are valid and do not consider the implications of invalidity for the inference.[9] IV methods are inconsistent unless the instruments are valid. Even conditional on validity of the instruments, the finite sample distribution of IV estimators can exhibit large bias and very fat tails. The fundamental intuition for poor sampling properties of IV estimators is that they are constructed from a ratio of normal random variables. I demonstrated that methods such as ML and OLS can have dramatically smaller RMSE than IV estimators for both linear and nonlinear models. It is commonly thought that the failure of asymptotic methods for IV estimators occurs only in conditions of very weak instruments. If the incremental R-squared due to the instruments is less than 20%, IV estimators have poor sampling properties relative to non-IV alternatives. For truly weak instruments (incremental R-squared less than 10%), then standard asymptotics should not be used to compute confidence intervals and standard errors.

[9] Conley et al. (2012) consider inference where IVs are not strictly valid but are only approximately exogenous.

Many applied researchers believe that IV methods are useful to implement a form of sensitivity analysis. That is, we compute both IV and non-IV estimates of the response parameters of interest. If these estimates (or functions of these estimates such as elasticities) differ, then many applied researchers conclude that this is evidence of endogeneity bias. This is only true if the instrument is valid and the model is properly specified. Invalid instruments can cause the estimates to differ even when there is no endogeneity bias.

Our concern in marketing should always be how to identify the causal effects of marketing variables. This means that we must be careful in interpreting standard regression style or correlational estimates as valid causal effects. Under certain conditions and with a comprehensive set of controls, we can interpret these estimates as causal. IV methods offer a very tempting alternative that appears to solve certain problems of identifying causal effects. However, this is true only for large and informative data sets with instruments for which there is a strong argument for validity based on economic principles. If relatively strong and valid instruments are not available, IV methods should not be used.

My hope is that this paper will spur a re-examination of the use of IV methods and generate a greater interest in both experimental methods that produce valid instruments by definition and measurement of the unobservables which create the endogeneity concerns in the first place.

Acknowledgments

The author thanks Wayne Taylor for his excellent research assistance. The author also thanks Wes Hartmann, Guenter Hitsch, Sanjog Misra, and Harikesh Nair for beneficial discussions. Funding from the Collins Chair, Anderson School of Management, UCLA is gratefully acknowledged.

References

Ackerberg D, Benkard CL, Berry S, Pakes A (2007) Econometric tools for analyzing market outcomes. Heckman J, Leamer E, eds. *Handbook of Econometrics* (Elsevier, Amsterdam), 4172–4271.

Allenby GM, Rossi PE (1999) Marketing models of consumer heterogeneity. *J. Econometrics* 89(1):57–78.

Angrist JD, Pischke J-S (2009) *Mostly Harmless Econometrics* (Princeton University Press, Princeton, NJ).

Bass FM (1969) A simultaneous equation study of advertising and sales of cigarettes. *J. Marketing Res.* 6(3):291–300.

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

Chintagunta PK, Nair H (2011) Discrete-choice models of consumer demand in marketing. *Marketing Sci.* 30(6):977–996.

Chintagunta PK, Dubé J-P, Goh KY (2005) Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household level brand choice models. *Management Sci.* 51(5):832–849.

Conley TG, Hansen CB, Rossi PE (2012) Plausibly exogenous. *Rev. Econom. Statist.* 94(1):260–272.

Dubé J, Hitsch G, Rossi PE (2010) State dependence and alternative explanations for consumer inertia. *RAND J. Econom.* 41(3): 417–445.

Erdem T, Imai S, Keane MP (2003) Brand and quantity choice dynamics under price uncertainty. *Quant. Marketing Econom.* 1(1):5–64.

Hansen C, Hausman J, Newey W (2008) Estimation with many instrumental variables. *J. Bus. Econom. Statist.* 26(4):398–422.

Hausman J (1996) The valuation of new goods under perfect and imperfect competition. Bresnahan T, Gordon R, eds. *The Economics of New Goods*, Vol. 58 (University of Chicago Press, Chicago), 209–237.

Hayashi F (2000) *Econometrics* (Princeton University Press, Princeton, NJ).

Manchanda P, Rossi PE, Chintagunta PK (2004) Response modeling with nonrandom marketing-mix variables. *J. Marketing Res.* 41(November):467–478.

Moreira MJ (2003) A conditional likelihood ratio test for structural models. *Econometrica* 71(4):1027–1048.

Narayanan S, Nair H (2013) Estimating causal installed-base effects: A bias-correction approach. *J. Marketing Res.* 50(1):70–94.

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

Petrin A, Train K (2010) Control function corrections for unobserved factors in differentiated product models. *J. Marketing Res.* 47(1): 3–13.

Stock JH, Wright JH (2000) GMM with weak identification. *Econometrica* 68(5):1055–1096.

Stock JH, Wright JH, Yogo M (2002) A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econom. Statist.* 20(4):518–529.

Villas-Boas JM, Winer RS (1999) Endogeneity in brand choice models. *Management Sci.* 45(10):1324–1338.

Woolridge JM (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).