Marketing Science

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

The Sense and Non-Sense of Holdout Sample Validation in the Presence of Endogeneity

Peter Ebbes, Dominik Papies, Harald J. van Heerde,

To cite this article:
Peter Ebbes, Dominik Papies, Harald J. van Heerde, (2011) The Sense and Non-Sense of Holdout Sample Validation in the Presence of Endogeneity. Marketing Science 30(6):1115-1122. https://doi.org/10.1287/mksc.1110.0666

MARKETING SCIENCE
CELEBRATING 30 YEARS
Vol. 30, No. 6, November–December 2011, pp. 1115–1122
ISSN 0732-2399 | EISSN 1526-548X | 11 | 3006 | 1115
http://dx.doi.org/10.1287/mksc.1110.0666

The Sense and Non-Sense of Holdout Sample Validation in the Presence of Endogeneity

Peter Ebbes
Fisher College of Business, Ohio State University, Columbus, Ohio 43210, ebbes.1@osu.edu

Dominik Papies
Institute for Marketing and Media, University of Hamburg, 20354 Hamburg, Germany, dominik.papies@uni-hamburg.de

Harald J. van Heerde
University of Waikato, Hamilton 3240, New Zealand; and Extramural Fellow at CentER, Tilburg University, 5000 LE Tilburg, The Netherlands, heerde@waikato.ac.nz

Market response models based on field-generated data need to address potential endogeneity in the regressors to obtain consistent parameter estimates. Another requirement is that market response models predict well in a holdout sample. With both requirements combined, it may seem reasonable to subject an endogeneity-corrected model to a holdout prediction task, and this is quite common in the academic marketing literature. One may be inclined to expect that the consistent parameter estimates obtained via instrumental variables (IV) estimation predict better than the biased ordinary least squares (OLS) estimates. This paper shows that this expectation is incorrect. That is, if the holdout sample is similar to the estimation sample so that the regressors are endogenous in both samples, holdout sample validation favors regression estimates that are *not corrected* for endogeneity (i.e., OLS) over estimates that *are corrected* for endogeneity (i.e., IV estimation). We also discuss ways in which holdout samples may be used sensibly in the presence of endogeneity. A key takeaway is that if consistent parameter estimates are the primary model objective, the model should be validated with an exogenous (rather than endogenous) holdout sample.

Key words: model validation; instrumental variables; IV estimation; endogeneity; exogeneity; predictive model; descriptive model; holdout sample; prediction

History: Received: May 25, 2010; accepted: June 5, 2011; Eric Bradlow and then Preyas Desai served as the editor-in-chief and Jean-Pierre Dubé served as associate editor for this article. Published online in *Articles in Advance* October 13, 2011.

1. Introduction

Market response models are mathematical models that represent the effects of marketing mix variables on performance measures such as sales or market shares (Hanssens et al. 2001, p. 4). Market response models have been developed to advance marketing knowledge and to aid managerial decision making (Leeflang et al. 2000, p. 3). These models may be classified in three categories according to their objectives. When the model is constructed to provide statements about the effectiveness of marketing instruments, it is considered to be a *descriptive* model (e.g., Franses 2005). *Predictive* models are developed to obtain accurate forecasts, whereas *normative* models are developed to offer a recommended course of action (Leeflang et al. 2000, pp. 38–39).

The marketing science community has developed a long and thorough checklist of requirements that a market response model should satisfy (e.g., Shugan 2004, Franses 2005, van Heerde et al. 2005). When field-generated (i.e., nonexperimental) data are used, a standard requirement is that market response models address potentially endogenous regressors (e.g., Villas-Boas and Winer 1999, Chintagunta 2001, Bronnenberg and Mahajan 2001, Shugan 2004). Endogeneity arises if regressors such as price and advertising are set based on demand shocks observable to the manager, but these shocks are unobserved by the researcher and omitted from the model. The resulting correlation between the error term and the regressors, if not properly addressed, biases the parameter estimates. When the regressors are varied experimentally, there is usually no endogeneity bias.

Endogeneity is especially problematic for descriptive models seeking unbiased estimates of the effects of marketing mix variables on performance. Correcting for endogeneity is required to obtain consistent estimates, which is frequently done through instrumental variables (IV) estimation (e.g., Berry 1994, Besanko et al. 1998). The exogenous information in

the IV is used to remove the endogenous variation in the regressors before these enter the regression for the dependent variable. The resulting regression coefficient represents the effect of an exogenous shock in the independent variable on the dependent variable.

Another standard requirement is that market response models predict well in a holdout sample. Holdout sample validation looks at the predictive performance for a part of the data set that was not used for model estimation (e.g., Cooil et al. 1987, Steckel and Vanhonacker 1993). The set of observations is split into two subsamples. First, the estimation sample is used to estimate the model parameters. Next, the fitted model is used in the holdout sample to predict the values of the dependent variable, which are then compared to the observed values. Holdout sample predictions may be made for new cross-sectional units (e.g., persons, stores, markets), for new time periods, or for both new cross-sectional units and new time periods. Holdout sample validation is often used to select a model from a number of competing models (e.g., Allenby 1990, Picard and Cook 1984) and to evaluate whether an estimated model generalizes to a new sample of observations not used for estimation (Steckel and Vanhonacker 1993).

Thus, endogeneity correction and holdout sample validation are standard requirements for marketing models that use field-generated data. Hence, it may seem reasonable to subject an endogeneity-corrected model to a holdout prediction task because one may be inclined to expect that the consistent parameter estimates obtained via IV estimation predict better than the biased and inconsistent ordinary least squares (OLS) estimates.[1] For example, Franses (2005, p. 10) suggests that "if a model is not performing well as a descriptive device, for example because its parameters are not estimated consistently, it is unlikely that the model will deliver good out-of-sample forecasts." Although this may be true in many cases, we investigate whether this also holds in the presence of endogeneity.

Holdout sample validation of endogeneity-corrected models is quite common in the academic marketing literature. Our review of major marketing journals[2] for the period of 1990–2010 identified 26 articles that used endogeneity correction in combination with a holdout sample validation task (see Appendix A of the electronic companion, available as part of the online version that can be found at http://mktsci.pubs.informs.org/). Examples include Neslin (1990), Besanko et al. (1998), van Dijk et al. (2004), Leenheer et al. (2007), and Andrews and Currim (2009). In several instances, the endogeneity-corrected model outperformed the noncorrected model in a holdout sample task (e.g., Leenheer et al. 2007, van Dijk et al. 2004), whereas in other cases (e.g., Neslin 1990, Besanko et al. 1998), the reverse was true. Hence, the outcomes of the comparison are mixed, which raises the issue of the sense (or non-sense) of holdout sample validation in the presence of endogeneity.

In this paper we formally compare the holdout sample performance of IV and OLS. We find that correcting for endogeneity comes at a cost: the holdout sample prediction performance deteriorates. That is, if the holdout sample is similar to the estimation sample so that the regressors are endogenous in both samples, holdout sample validation favors regression estimates that are *not corrected* for endogeneity (i.e., OLS) over estimates that *are corrected* for endogeneity (i.e., IV). The results hold without making distributional assumptions on the error term. Section 2 explains why it does not make sense to compare OLS to IV in a regular holdout sample. Next, §3 shows how holdout samples may be used sensibly to validate models in the presence of endogeneity. Section 4 concludes this paper.

2. Why OLS Predicts Better Than IV in a Holdout Sample

2.1. Intuition

We first consider a stylized example for why OLS predicts better than IV in a holdout sample. Using the context of modeling movie DVD sales, we assume that price is the only marketing instrument. Prices for the movie DVDs are set by the seller based on movie quality; i.e., prices are set relatively high in case of a high movie quality to benefit from the expected positive demand shock, and vice versa, in case of low movie quality. Consequently, relatively high prices are associated with relatively high sales. When sales are regressed on price, and movie quality is unobserved, this results in an OLS estimated price coefficient that is biased toward zero. IV may be used to correct for endogeneity, and the IV estimate for the price coefficient is expected to be more negative than the OLS estimate (Bijmolt et al. 2005).

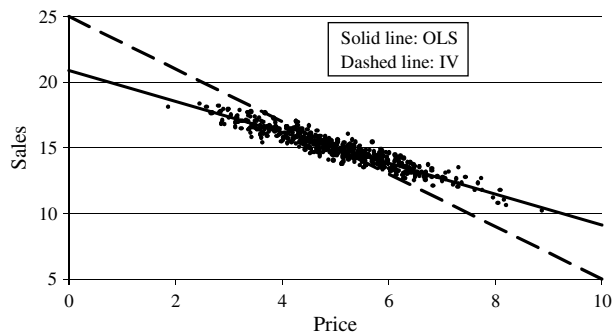
Figure 1 shows the scatterplot for a stylized estimation sample. The OLS fitted line represents the best-fit line through the scatter, whereas the IV line is tilted. We also see that the OLS line is flatter than the IV fitted line, indicating that the OLS price coefficient is biased toward zero.

Now consider validating the OLS and IV parameter estimates, both obtained from the estimation sample,

[1] In fact, we have seen several requests for holdout sample comparisons after endogeneity correction during the review processes at major academic marketing journals.

[2] *Marketing Science*, *Management Science*, the *Journal of Marketing Research*, the *Journal of Marketing*, the *International Journal of Research in Marketing*, and *Quantitative Marketing and Economics*.

Figure 1 **Estimation Sample**

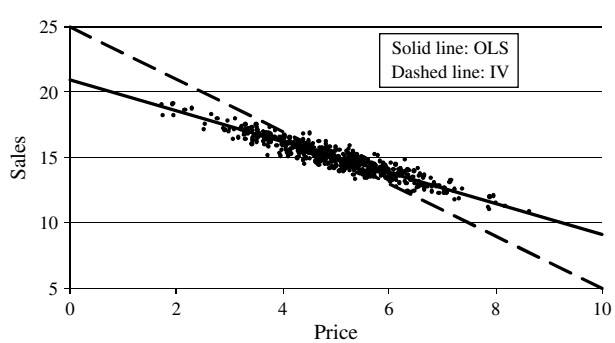


Note. In the estimation sample, the OLS regression line (solid) shows a weaker effect of price on sales than the IV regression line (dashed).

in a holdout prediction exercise based on the notion that a lower forecast error indicates a better model. Suppose we want to predict the sales for a relatively high-priced DVD movie. Because the price coefficient estimated by OLS is smaller than the price coefficient estimated by IV, the OLS fitted model predicts higher sales for this DVD than the IV fitted model, which turns out to be a better prediction. The intuition for this result is as follows. If the price level is relatively high, the OLS model "knows" that there is likely a positive demand shock (high movie quality) that has led to this price level, and hence it rightly predicts relatively high sales based on the biased estimated price coefficient. The IV model, instead, "does not know" there is such a positive demand shock for this high-priced DVD and predicts low sales.

We illustrate the superior predictive ability of OLS in Figure 2, which very much resembles Figure 1 because the sample was randomly split, and therefore it has the same underlying data generation mechanism. It shows that the OLS line fits through the middle of the scatter of holdout sample observations, whereas the IV fitted line is tilted away. Because the IV fitted model consistently underpredicts sales for high prices and overpredicts sales for low prices, the inconsistent OLS fitted model has smaller prediction errors than the consistent IV fitted model. This invalidates the notion that a lower forecasting error in a holdout sample is indicative of a better model, i.e., a model with consistent parameter estimates.

Figure 2 **Holdout Sample**



Note. In the holdout sample, the estimation-sample OLS regression line (solid) provides better predictions than the tilted estimation-sample IV regression line (dashed).

2.2. Formal Analysis

We now turn to a formal assessment of the predictive ability (in mean squared error) of OLS versus IV in holdout samples. All results are derived in Appendix B of the electronic companion. We assume that N_1 observations are available to estimate the model parameters and N_2 observations are used to validate the model, and the total sample size is $N = N_1 + N_2$. The sample might be either a cross section or a time series. There are K regressors, and the matrix of regressors X is of dimension $N \times K$, which is assumed to be of full rank. The dependent variable Y and the errors ϵ are both of dimension $N \times 1$. The regression model is given by

$$Y = X\beta + \epsilon. \tag{1}$$

We assume that the errors are independent and identically distributed (i.i.d.) with a mean of 0 and a variance of σ_ϵ^2. We indicate the data vectors and matrices for the estimation sample by subscript "1" and for the holdout sample by subscript "2"; i.e., X_1, Y_1, and ϵ_1 are the first N_1 rows from X, Y, and ϵ, respectively, and X_2, Y_2, and ϵ_2 are the last N_2 rows from X, Y, and ϵ, respectively. The OLS estimator $\hat{\beta}_{\text{OLS}} = (X_1'X_1)^{-1}X_1'Y_1$ is unbiased and consistent for β under the assumption that the regressors and errors are independent (e.g., Greene 2003, Chapter 5). When this is not the case, the regressors are endogenous, and the IV approach may be used to obtain a consistent estimator for β under the usual assumption that the instruments are valid (e.g., Davidson and MacKinnon 1993).[3] The IV regression model may be represented as a limited information simultaneous equation model (Kleibergen and Zivot 2003), where the single-equation model (1) is augmented with an equation for the endogenous regressors:

$$X = Z\Gamma + \Psi. \tag{2}$$

Here, Γ is a $L \times K$ matrix of regression coefficients denoting the effects of the instruments Z, and Ψ is a $N \times K$ matrix of error terms. The $(K+1) \times 1$ vector of errors ϵ_i and Ψ_i', where Ψ_i is the ith row from Ψ, are i.i.d. with a mean of 0 and a covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_\epsilon^2 & \Sigma_{\epsilon\Psi} \\ \Sigma_{\Psi\epsilon} & \Sigma_{\Psi\Psi} \end{bmatrix}. \tag{3}$$

[3] In this study, we assume that the available instruments Z are valid; i.e., the instruments have finite moments, have some correlation with X (at least asymptotically), and are exogenous (e.g., Davidson and MacKinnon 1993).

Let Z_1 be the $N_1 \times L$ matrix containing the IV for the estimation sample, where $L \geq K$ for identification. The IV estimator $\hat{\beta}_{\text{IV}} = (X_1' P_{Z_1} X_1)^{-1} X_1' P_{Z_1} Y_1$, where $P_{Z_1} = Z_1(Z_1'Z_1)^{-1}Z_1'$, is preferred to the OLS estimator when the regressors are endogenous, i.e., when $\Sigma_{\epsilon\Psi} \neq 0$, because the OLS estimator is biased and inconsistent.[4]

For making holdout predictions using the holdout regressors X_2, the predicted values of Y_2 using the OLS and IV fitted models are, respectively,

$$\hat{Y}_{\text{OLS},2} = X_2 \hat{\beta}_{\text{OLS}}, \tag{4}$$

$$\hat{Y}_{\text{IV},2} = X_2 \hat{\beta}_{\text{IV}}. \tag{5}$$

The asymptotic mean squared prediction errors (A.MSE) for the IV fitted model is

$$\text{A.MSE}_{\text{IV}} = \sigma_\epsilon^2, \tag{6}$$

whereas the A.MSE for the OLS fitted model is

$$\text{A.MSE}_{\text{OLS}} = \sigma_\epsilon^2 - \Sigma_{\epsilon\Psi} Q_{XX}^{-1} \Sigma_{\Psi\epsilon}, \tag{7}$$

where Q_{XX} is the (asymptotic) matrix of second-order moments of $X'X$. Since $\Sigma_{\epsilon\Psi} Q_{XX}^{-1} \Sigma_{\Psi\epsilon}$ is nonnegative, we have

$$\text{A.MSE}_{\text{OLS}} \leq \text{A.MSE}_{\text{IV}}. \tag{8}$$

Equation (8) shows that the IV fitted model is generally *inferior* to the OLS fitted model in a holdout prediction task.[5] Our derivations also imply that only the IV fitted model gives the correct size σ_ϵ^2 of the asymptotic mean squared prediction errors. The mean squared prediction errors from the OLS fitted model are asymptotically smaller than σ_ϵ^2, making the predictions "fit too well."

2.3. Numerical Illustration

Table 1 shows the results of a numerical example to illustrate expressions (6)–(8). We consider a simple regression model with one endogenous regressor ($K = 1$) and one observed instrument ($L = 1$) (the errors and the instrument have a mean of 0 and a variance of 1), and the instrument explains 50% of the variance in the regressor. Specifically, the model is $y_{i,j} = x_{i,j} + \epsilon_{i,j}$; $x_{i,j} = z_{i,j} + \Psi_{i,j}$, for $j = 1$ (estimation sample) and $j = 2$ (holdout sample), and for $i = 1, \ldots, N$. The covariance between the errors is $\sigma_{\epsilon\Psi} = 0$, 0.47, or 0.93, representing no endogeneity, a low extent of endogeneity, or a high extent of endogeneity, respectively.

Table 1 **Parameter Estimates and Holdout Sample Asymptotic MSE for OLS vs. IV**

	$\hat{\beta}$[a]		A.MSE	
Extent of endogeneity	OLS	IV	OLS	IV
No	1.00	1.00	1.00	1.00
Low	1.24	1.00	0.89	1.00
High	1.47	1.00	0.57	1.00

[a]The true value is $\beta = 1$.

When the regressors are exogenous (see the first of row Table 1), $\hat{\beta}_{\text{OLS}} = \hat{\beta}_{\text{IV}} = 1$ asymptotically, and all fitted models have the same asymptotic mean squared error equal to $\sigma_\epsilon^2 = 1$. In the low-endogeneity case (see the second row), the OLS estimates for the effect of X on Y are inconsistent; $\hat{\beta}_{\text{OLS}} = 1.24$, whereas $\hat{\beta}_{\text{IV}} = 1.00$. Despite the bias in the OLS estimate for β, the A.MSE from the OLS fitted model using X is 0.89, which is lower than the A.MSE for the IV fitted model, which is 1.00. For the high-endogeneity case (see the third row), the OLS coefficient is strongly biased ($\hat{\beta}_{\text{OLS}} = 1.47$), yet the A.MSE for OLS improves to 0.57. Hence, a seemingly straightforward holdout sample comparison between OLS and IV does not squarely identify the correct effect of X on Y. Instead, the biased and inconsistent OLS estimates predict better than the consistent IV estimates.

2.4. Using the IV in the Holdout Sample Prediction

So far, the prediction of Y is based on the *observed* endogenous regressor X_2 in the holdout sample. It may seem reasonable to also use the IV Z_2 in the holdout sample to improve the predictions. There are two conceivable approaches to do that.

One approach is to use the predicted $\hat{X}_2$, where $\hat{X}_2 = Z_2 \hat{\Gamma}_{\text{IV}}$, to predict the dependent variable; i.e., $\hat{Y}_{\text{IV},\hat{X},2} = \hat{X}_2 \hat{\beta}_{\text{IV}}$. However, this fitted model may or may not outperform the OLS or IV fitted models in (4) and (5) that use the *observed* X_2 instead. Whether the predictions using $\hat{X}_2$ outperform OLS or IV (in terms of A.MSE) depends on the value of β and the variance components in Σ (see Appendices C and D of the electronic companion). Hence using the predicted regressors $\hat{X}_2$ in a holdout validation task is not recommended because the predictive performance from this approach is unstable and may or may not be inferior to either IV or OLS.

Instead of replacing X_2 by $\hat{X}_2$ in (5), another approach for holdout sample predictions uses both

[4] The term $\hat{\beta}_{\text{IV}}$ can be interpreted as a local average treatment effect (LATE) when the monotonicity condition is fulfilled. This condition implies that the effect of Z on the endogenous regressor X may be heterogeneous across the response units, but either is ≥ 0 or ≤ 0 for all units. Then $\hat{\beta}_{\text{IV}}$ may be interpreted as the causal effect of X on Y for those observations whose value of X is affected by Z (Angrist and Imbens 1995, p. 434).

[5] Our results are large-sample results. We have also investigated whether our findings hold for finite samples in a simulation study and found that, indeed, OLS consistently outperforms IV in holdout samples. Only for very small sample sizes of 20 observations, and only in 25 out of 100 replications, did IV have a lower MSE in the holdout sample. For samples of size 50 or more, in nearly all cases, OLS had the lowest MSE.

the observed X_2 and Z_2 jointly to predict Y_2. This approach utilizes information in both (1) and (2) through either a control function specification (Petrin and Train 2010; Verbeek 2008, p. 144) or a likelihood approach under joint normality of the errors. Appendix D of the electronic companion shows that using information from X and Z jointly improves holdout sample performance for both IV and OLS. However, this approach still does not identify the correct estimated relationship between X and Y, because an inconsistent OLS approach that uses the same set of information (X_2 and Z_2) yields identical predictions compared to IV.

In sum, using information in the holdout sample IV to predict Y, either by first predicting $\hat{X}_2$ or by using X_2 and Z_2 jointly, does not help identify the correct estimated relationship between X and Y. In the next section, we discuss ways in which holdout sample validation may make sense in the presence of endogeneity.

3. How Holdout Sample Validation Can Be Used in the Presence of Endogeneity

3.1. Predicting with Exogenous Holdout Regressors X

The analyses so far assume that the extent of endogeneity is the same in the estimation as in the holdout sample. In the case when the holdout sample predictor is *exogenous*, the A.MSE_{IV} remains equal to σ_ϵ^2, but the $\text{A.MSE}_{\text{OLS}}$ becomes (see Appendix E of the electronic companion)

$$\text{A.MSE}_{\text{OLS}} = \sigma_\epsilon^2 + \Sigma_{\epsilon\Psi,1} Q_{\text{XX}}^{-1} \Sigma_{\Psi\epsilon,1}, \tag{9}$$

where $\Sigma_{\epsilon\Psi,1}$ is $\Sigma_{\epsilon\Psi}$ for the estimation sample. Hence,

$$\text{A.MSE}_{\text{IV}} \leq \text{A.MSE}_{\text{OLS}}, \tag{10}$$

meaning that if the holdout predictors are *exogenous*, the IV fitted model asymptotically yields *lower* holdout mean squared prediction errors than the OLS fitted model. The intuition is that OLS can no longer capitalize on the correlation between the error term and the regressor to improve predictions, whereas the IV estimator is designed to capture the effect of exogenous variation in the regressor on the dependent variable. Extending the numerical example from the previous section, the A.MSE for OLS is now 1.11 (1.43) for low (high) endogeneity in the estimation sample (see Table 2), whereas the A.MSE of the IV fitted model equals $\sigma_\epsilon^2 = 1$, i.e., the correct amount of unexplained variance in Y. Hence, an *exogenous* holdout sample identifies the correct estimated relationship between X and Y.

Table 2 **Holdout Sample Asymptotic MSE for OLS vs. IV for *Exogenous* Holdout Sample**

Extent of endogeneity in estimation sample	$\hat{\beta}$[a]		A.MSE	
	OLS	IV	OLS	IV
Low	1.24	1.00	1.11	1.00
High	1.47	1.00	1.43	1.00

[a]The true value is $\beta = 1$.

3.2. Using "Pseudo-Exogenous" Regressors for Holdout Validation

Experimentally varying the regressors (e.g., Hoch et al. 1994), which is required to obtain an exogenous holdout sample, may not always be practically feasible. However, there is an opportunity to create "pseudo-exogenous" regressors in the context of panel data. Suppose the observations follow a panel data structure with at least two indices—say, i and t—where dimensions i and t may refer to different types of cross sections (households, stores, countries) and time, respectively. Without loss of generality, suppose that the endogeneity in the regressors is primarily in dimension i but not in dimension t. We can now validate the endogeneity-corrected model by predicting deviations in the dependent variable based on deviations in the regressors relative to their means in dimension i. That is, instead of using the fitted model $X_{it,2}\hat{\beta}_{\text{IV}}$ for making holdout predictions, we can use $(X_{it,2} - \bar{X}_{i,2})\hat{\beta}_{\text{IV}}$ with $\bar{X}_{i,2} = (1/T)\sum_t X_{it,2}$. Hence the "pseudo-regressors" $(X_{it,2} - \bar{X}_{i,2})$ (deviations in X) are used to predict $(Y_{it,2} - \bar{Y}_{i,2})$ (deviations in Y). Appendix F in the electronic companion shows that IV now has superior predictive performance over OLS:

$$\text{A.MSE}_{\text{IV.pseudo}} \leq \text{A.MSE}_{\text{OLS.pseudo}}, \tag{11}$$

which holds for holdout samples consisting of new observations in either dimension i or t, or both.

This approach to holdout sample validation was used by van Dijk et al. (2004) to test the validity of shelf-space elasticity estimates. Their estimation sample has a cross-sectional dimension i (store) and a time-series dimension t (measurement periods), and the endogeneity was concentrated in the cross-sectional dimension. They found that the model that corrects for cross-sectional endogeneity predicts changes in holdout sample sales better than a model without such a correction.

A second example of this type of holdout sample validation is Leenheer et al. (2007). This study looks at the effect of loyalty program membership on behavioral loyalty (share of wallet). Because households that are already loyal are particularly more likely to sign up for a loyalty program, the regressor "loyalty program adoption" is mostly endogenous in the cross-sectional

(household) dimension. Leenheer et al. (2007) validate their endogeneity-corrected model in the time-series dimension by predicting the change in loyalty for households who (dis)adopted loyalty program membership in the subsequent year. The longitudinal change was better predicted by the endogeneity-corrected model than the model without such a correction. These two examples show that deliberately constructed pseudo-exogenous regressors may offer a way to assess the merits of an endogeneity correction.

3.3. The Merit of Holdout Sample Validation in Endogenous Holdout Samples

Although we argue against the OLS versus IV comparison in endogenous holdout samples (see §2), there are two cases in which holdout sample validation may have merit for endogenous holdout samples.

One case is as a guard against overfitting. This can be achieved by comparing the in-sample MSE to the holdout sample MSE for an IV fitted model. Equation (3) shows that the IV fitted model has an A.MSE in the estimation and holdout samples that is equal to the true variance of the error term. Therefore, if the holdout sample MSE (for IV) turns out to be substantially higher than the in-sample MSE (for IV), then this suggests that the estimated IV model captures idiosyncrasies of the estimation sample, but it cannot be readily generalized to other samples (Picard and Cook 1984). Drèze et al. (2004), for example, use holdout sample validation in the presence of endogeneity in this way.

Alternatively, we can compare the holdout sample MSEs for two IV fitted models that have the *same* set of instrumental variables. Suppose we compare two otherwise identical models, A and B, where model B has one additional *exogenous* regressor. Hagerty and Srinivasan (1991) show that the holdout sample MSE offers a trade-off between the squared bias in prediction and the variance in prediction error (see also Leeflang et al. 2000, p. 503; Shmueli 2010). If the additional regressor in model B causes a relatively large decrease in bias (e.g., because it is an important predictor) and only a small increase in variance, then the holdout sample MSE of model B will be lower than for model A. In that case we would conclude that the extra regressor in model B is a useful addition to the set of regressors. If the additional regressor adds relatively more noise (variance) than what it reduces in bias (e.g., because the predictor has a small effect on the dependent variable), model A will have a lower holdout sample MSE. In that case, the simpler model predicts better than the larger model, which signals that the larger model is overspecified.

The second case in which an endogenous holdout sample may have merit is to investigate the strength of IV. Suppose we have two candidate IV: Z_A and Z_B. We can use the reduced-form model to investigate the strength of the two IV as follows. The reduced-form model for Y regresses Y on all available exogenous regressors, including the instruments (e.g., Wooldridge 2002, p. 84). By comparing the reduced-form holdout sample fit for Y from the fitted model that uses Z_A to the fitted model that uses Z_B, we can gauge which instrumental variable is stronger in the holdout sample. Similarly, the reduced-form model for X can be used to assess which instrumental variable predicts X better in the holdout sample.

An important caveat here is that comparing the predictive power of two IV *only* makes sense when both instruments are *exogenous*. If an instrumental variable is not exogenous, it may predict X or Y better than the exogenous instrument, but that does not mean it should be preferred over the exogenous instrument. This comes back to the same argument that OLS should not be preferred over IV because it predicts better.

4. Conclusions

Endogeneity correction is important in understanding how marketing performance is affected by changes in the marketing mix. Holdout sample validation also has important merits because it can be used to select models and to assess whether the estimated relationships hold beyond the observations used for estimation.

Our literature review demonstrates that holdout sample validation in the presence of endogeneity is common in the marketing literature. Apparently, the literature is insufficiently aware that superior predictive performance and consistent estimates in regression models are incompatible objectives. Indeed, correcting for endogeneity implies that the model has a descriptive purpose: obtaining consistent estimates for marketing response parameters. Such a descriptive purpose, however, is incompatible with a predictive purpose of realizing the best holdout sample performance, as we show in this study. That is, the consistent IV fitted model in (5) does not outperform the inconsistent OLS fitted model in (4) in terms of asymptotic mean squared prediction errors. It is important to note that our results are derived without making distributional assumptions for the model error terms.

Although endogeneity correction and holdout sample validation are often combined in the literature, we identified only one citation that made a suggestion about the incompatibility of this combination: "In addition, other studies have found that simultaneous equations models in which parameter estimates have been obtained with OLS often predict just as well as when parameters are estimated through more sophisticated methods, even though the OLS

estimates are biased (Kennedy 1992)" (Besanko et al. 1998, p. 1542).

Our research has a number of practical implications. Most importantly, the researcher should first decide on the model objective and then develop and validate the model according to this objective (see also Shmueli 2010). This study provides two main recommendations for model development and validation in the presence of endogeneity:

1. Descriptive or Normative Models: If the model has a *descriptive* or *normative* purpose, consistent estimates are key, and an estimation approach that corrects for endogeneity, such as IV, is required. The researcher needs to argue that the IV are not correlated with the error term of the main equation and show that the IV have a sufficiently strong correlation with the endogenous regressor (e.g., Verbeek 2008, pp. 156–157). Next, the researcher can test for endogeneity using, e.g., the Hausman test (Verbeek 2008, p. 144). However, we argue against the use of regular random-split holdout sample validation to compare the OLS and IV fitted models. Such a split would lead to a similar extent of endogeneity in the estimation and holdout samples, which would favor the biased OLS estimates over the consistent IV estimates and does not identify the correct estimated relationship between Y and X.

Although we argue against the OLS versus IV comparison in endogenous holdout samples, we discussed two cases where validating with endogenous holdout samples may make sense. These cases include the use of holdout sample validation as a guard against overfitting and as a way to determine the strength of IV.

If the holdout sample contains *exogenous* variation in the regressors, holdout sample validation may be used to identify the correct estimated relationship between Y and X. As we show, an exogenous holdout sample results in a lower A.MSE for IV than for OLS. An exogenous holdout sample can be obtained through an experimental design, although this may not always be practically feasible. For panel data where the endogeneity is likely to be concentrated in one dimension, we show how a pseudo-exogenous holdout sample can be constructed.

2. Predictive Models: If *prediction* is the primary model objective, we recommend refraining from using IV estimation. If the data-generating process is the same in the holdout sample, the OLS fitted model will predict at least as well as the IV fitted model when both use the endogenous holdout sample regressor X only. If the OLS and IV fitted models each use both X and Z to form predictions, the models provide identical holdout sample performance (see §2.4). In neither case, however, does IV outperform OLS in an endogenous holdout sample.

We recommend future studies that use holdout sample validation in the presence of endogeneity to offer sufficient detail describing the holdout prediction exercise. First, which exact benchmarks are used, and do the benchmarks account for endogeneity or not? Second, how is the holdout sample selected, and is it similarly endogenous as the estimation sample, or exogenous?

We have made our case using a comparison between OLS and IV in a homogeneous regression model with a continuous dependent variable. Our intuition suggests that similar results apply to the case of limited dependent variables, because these models may be viewed as a realization of an underlying continuous (latent) variable regression model (Rossi et al. 2005). Furthermore, we conjecture that the direction of our results also generalizes to regression models with heterogeneous response parameters (e.g., Allenby and Rossi 1999). The intuition is that an endogeneity correction tilts the individual-specific regression lines away from the individual-specific least squares lines, similar to what is shown in Figure 1 for the homogeneous case. Two simulation studies in Web Appendices G and H in the electronic companion support our intuition for these generalizations.

This paper sheds some light on the sense and nonsense of holdout sample validation in the presence of endogeneity and extends the recent discussion about the usefulness of predictive validation (e.g., Shugan 2009, Tsang 2009). What we hope to achieve is a better awareness among researchers and reviewers alike that it is unrealistic to expect superior predictive performance after an endogeneity correction unless the holdout sample is exogenous. Because the luxury of an exogenous holdout sample is often unavailable, we also hope that this study inspires new research into ways of validating market response models that correct for endogeneity.

5. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mktsci.pubs.informs.org/.

Acknowledgments

The authors are listed alphabetically and contributed equally to this research. Part of this work was conducted while P. Ebbes was at the Smeal College of Business, the Pennsylvania State University. This paper is based on research that, in part, was conducted while D. Papies was a visiting scholar at the Waikato Management School in Hamilton, New Zealand. H. J. van Heerde acknowledges the New Zealand Royal Society (Marsden fund 10-UOW-068) for research support. The authors thank Greg Allenby, André Bonfrer, Marnik Dekimpe, John Geweke, and Gary Lilien for their insightful comments on previous versions

of this manuscript. The authors are also grateful for the feedback they obtained while presenting this paper on various occasions: the Marketing Science Conference 2010 (University of Cologne), University of Cologne, Marketing Dynamics Conference 2010 (Özyeğin University), Facultés Universitaires Catholiques de Mons, and the University of Technology, Sydney.

References

Allenby, G. M. 1990. Cross-validation, the Bayes theorem, and small-sample bias. *J. Bus. Econom. Statist.* **8**(2) 171–178.

Allenby, G. M., P. E. Rossi. 1999. Marketing models of consumer heterogeneity. *J. Econometrics* **89**(1/2) 57–78.

Andrews, R. L., I. S. Currim. 2009. Multi-stage purchase decision models: Accommodating response heterogeneity, common demand shocks, and endogeneity using disaggregate data. *Internat. J. Res. Marketing* **26**(3) 197–206.

Angrist, J. D., G. W. Imbens. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Amer. Statist. Assoc.* **90**(430) 431–442.

Berry, S. T. 1994. Estimating discrete-choice models of product differentiation. *RAND J. Econom.* **25**(2) 242–262.

Besanko, D., S. Gupta, D. Jain. 1998. Logit demand estimation under competitive pricing behavior: An equilibrium framework. *Management Sci.* **44**(11, Part 1) 1533–1547.

Bijmolt, T. H. A., H. J. van Heerde, R. G. M. Pieters. 2005. New empirical generalizations on the determinants of price elasticity. *J. Marketing Res.* **42**(2) 141–156.

Bronnenberg, B. J., V. Mahajan. 2001. Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables. *Marketing Sci.* **20**(3) 284–299.

Chintagunta, P. K. 2001. Endogeneity and heterogeneity in a probit demand model: Estimation using aggregate data. *Marketing Sci.* **20**(4) 442–456.

Cooil, B., R. S. Winer, D. L. Rados. 1987. Cross-validation for prediction. *J. Marketing Res.* **24**(3) 271–279.

Davidson, R., J. G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. Oxford University Press, Oxford, UK.

Drèze, X., P. Nisol, N. J. Vilcassim. 2004. Do promotions increase store expenditures? A descriptive study of household shopping behavior. *Quant. Marketing Econom.* **2**(1) 59–92.

Franses, P. H. 2005. On the use of econometric models for policy simulation in marketing. *J. Marketing Res.* **42**(1) 4–14.

Greene, W. H. 2003. *Econometric Analysis*, 5th ed. Prentice Hall, Upper Saddle River, NJ.

Hagerty, M. R., S. Srinivasan. 1991. Comparing the predictive powers of alternative multiple regression models. *Psychometrika* **56**(1) 77–85.

Hanssens, D. M., L. J. Parsons, R. L. Schultz. 2001. *Market Response Models, Econometric and Time Series Analysis*, 2nd ed. Kluwer Academic Publishers, Norwell, MA.

Hoch, S. J., X. Drèze, M. E. Purk. 1994. EDLP, hi-lo, and margin arithmetic. *J. Marketing* **58**(4) 16–27.

Kennedy, P. 1992. *A Guide to Econometrics*, 3rd ed. MIT Press, Cambridge, MA.

Kleibergen, F., E. Zivot. 2003. Bayesian and classical approaches to instrumental variables regression. *J. Econometrics* **114**(1) 29–72.

Leeflang, P. S. H, D. R. Wittink, M. Wedel, P. A. Naert. 2000. *Building Models for Marketing Decisions*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Leenheer, J., H. J. van Heerde, T. H. A. Bijmolt, A. Smidts. 2007. Do loyalty programs really enhance behavioral loyalty? An empirical analysis accounting for self-selecting members. *Internat. J. Res. Marketing* **24**(1) 31–47.

Neslin, S. A. 1990. A market response model for coupon promotions. *Marketing Sci.* **9**(2) 125–145.

Petrin, A., K. Train. 2010. A control function approach to endogeneity in consumer choice models. *J. Marketing Res.* **47**(1) 3–13.

Picard, R. R., R. D. Cook. 1984. Cross-validation of regression models. *J. Amer. Statist. Assoc.* **79**(387) 575–583.

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

Shmueli, G. 2010. To explain or to predict? *Statist. Sci.* **25**(3) 289–310.

Shugan, S. M. 2004. Endogeneity in marketing decision models. *Marketing Sci.* **23**(1) 1–3.

Shugan, S. M. 2009. Relevancy is robust prediction, not alleged realism. *Marketing Sci.* **28**(5) 991–998.

Steckel, J. H., W. R. Vanhonacker. 1993. Cross-validating regression models in marketing research. *Marketing Sci.* **12**(4) 415–427.

Tsang, E. W. K. 2009. Assumptions, explanation, and prediction in marketing science: "It's the findings, stupid, not the assumptions." *Marketing Sci.* **28**(5) 986–990.

van Dijk, A., H. J. van Heerde, P. S. H. Leeflang, D. R. Wittink. 2004. Similarity-based spatial methods to estimate shelf space elasticities. *Quant. Marketing Econom.* **2**(3) 257–277.

van Heerde, H. J., M. G. Dekimpe, W. P. Putsis Jr. 2005. Marketing models and the Lucas critique. *J. Marketing Res.* **42**(1) 15–21.

Verbeek, M. 2008. *A Guide to Modern Econometrics*, 3rd ed. John Wiley & Sons, Hoboken, NJ.

Villas-Boas, J. M., R. S. Winer. 1999. Endogeneity in brand choice models. *Management Sci.* **45**(10) 1324–1338.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.