Industrial Marketing Management 77 (2019) 57–62

Contents lists available at ScienceDirect

Industrial Marketing Management

journal homepage: www.elsevier.com/locate/indmarman

Rejoinder to "Endogeneity bias in marketing research: Problem, causes and remedies"

Richard T. Gretz*, Ashwin Malshe

Department of Marketing, College of Business, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249-0631, United States of America

ABSTRACT

There are a few significant errors in Zaefarian, Kadile, Henneberg, and Leischnig's (2017) (henceforth ZKHL) discussion of remedies to address endogeneity issues in their recent survey article in *Industrial Marketing Management*. Most notably, they incorrectly describe 2-stage least squares (2SLS) estimation procedure. We provide the correct methodology here, along with sample data and code, and compare estimations using ZHKL's suggested methodology with proper 2SLS and ordinary least squares. We show that the method they suggest will actually result in greater bias in coefficient estimates while proper 2SLS addresses the endogeneity problem. Also, they incorrectly describe 3-stage least squares (3SLS), both in terms of implementation procedure and appropriate setting. We address both of these issues by describing when 3SLS would be used and the benefits of using 3SLS versus other methods to estimate simultaneous equations models. We discuss further issues with their paper and provide R code and simulated data.

1. Introduction

Zaefarian, Kadile, Henneberg, and Leischnig (2017) (henceforth ZKHL) recently published a survey article in this journal titled "Endogeneity bias in marketing research: Problem, causes and remedies." The goal of the piece was to inform applied researchers of the estimation problems caused by endogeneity issues, potential sources of endogeneity, and probably most importantly, an overview of empirical remedies to address endogeneity problems.

It is the latter part of the article with which we take issue. Several methodologies outlined in ZKHL are either incorrect, missing, or misleading. The goal of this comment is to address these errors, provide researchers with an overview of the appropriate methodology, compare the consequences of using the incorrect methodology outlined in ZKHL instead of the appropriate methodology, and give researchers examples to work with as they consider implementing these methodologies in practice.

First, we review endogeneity bias and provide some examples. Then we discuss two critical problems with ZKHL – their presentation of the methodology for 2-stage least squares (2SLS) and 3-stage least squares (3SLS). In addition to providing the appropriate methodology, we also provide estimation examples (along with code and data) to highlight the appropriate methodology and contrast with ZKHL's suggested methodology where appropriate. Finally, we discuss some additional issues with their paper to provide greater clarity for future researchers.

2. Endogeneity bias

One of the key assumptions of ordinary least squares (OLS) estimation is that the independent variables are exogenous in that they are not correlated with the econometric error term (Angrist & Pischke, 2008; Cameron & Trivedi, 2005; Greene, 2018; Wooldridge, 2010, 2013). Endogeneity bias occurs when this assumption is violated. An independent variable is said to be endogenous when it is correlated with the error term. ZKHL provide a detailed discussion of different sources of endogeneity and we do not repeat them here for brevity. Critically, if any of the independent variables are endogenous then OLS coefficient estimates are "biased" in that the expected value of the estimator is different from the true value (Angrist & Pischke, 2008; Cameron & Trivedi, 2005; Greene, 2018; Wooldridge, 2010). In other words, coefficients estimates from OLS will not reflect the true impact of independent variables on the dependent variable. Citing Semadeni, Withers, and Trevis Certo (2014), ZKHL correctly note that "endogeneity may affect the causal inferences that researchers make with regard to the hypothesized associations between variables, and failure to account for this may lead to spurious findings resulting in misleading theoretical as well as managerial implications (p. 39)".

For example, applied researchers often worry that price is endogenous when estimating product demand.[1] It is often assumed that price is set by managers, who consider information that is not observable to the econometrician. The impact of these unobservable

* Corresponding author.
E-mail addresses: richard.gretz@utsa.edu (R.T. Gretz), ashwin.malshe@utsa.edu (A. Malshe).

[1] Rossi (2014) notes that over 50% of empirical papers he examined in a recent survey of instrumental variable methods in marketing focused on price as an endogenous variable.

https://doi.org/10.1016/j.indmarman.2019.02.008

Available online 23 February 2019
0019-8501/

factors will be captured in the econometric error term, thereby leading to a correlation between the econometric error term and price. To make this more concrete, let's assume that we are interested in estimating the impact price has on quantity demanded for some product. Let's also assume that consumers value higher quality, mangers typically charge higher prices for higher quality products, and quality is unobservable in that we only have data on prices and quantities. OLS estimation of the impact of price on quantity demanded will likely be positively biased in this case. That is, the negative impact we expect price to have on quantity demanded will be confounded with the positive impact of higher quality. This demonstrates the endogeneity problem – the OLS estimate of the impact of price is biased because price is correlated with the impact of unobserved (to the econometrician) quality which is captured by the econometric error term.

Additionally, OLS estimates are "inconsistent" which means that the estimated coefficients do not converge to the population coefficients even as the estimation sample approaches infinity. This means that an applied researcher cannot simply address this problem by getting more data, which is a common strategy for dealing with issues such as non-normal distributed error terms or multicollinearity (Greene, 2018). What's worse, endogeneity bias and inconsistency are not limited to coefficient estimates of the offending variable or variables. It impacts *all* coefficient estimates, even for those variables in the model that are demonstrably exogenous (Greene, 2018; Rossi, 2014; Wooldridge, 2010). Given this backdrop, it is admirable that ZKHL take on the task of describing some methods to address endogeneity issues in applied work. However, their description of 2SLS—one of the most fundamental and oft-used methods to deal with endogeneity—is incorrect and will exacerbate any endogeneity problem. We discuss this next.

3. 2-stage least squares (2SLS)

2SLS is an instrumental variable (IV) method, requiring at least one additional exogenous variable to help identify the impact the offending endogenous variable has on the dependent variable. An IV has to be "relevant" in that it is correlated with the suspected endogenous variable and "valid" in that it is not correlated with the econometric error term (Cameron & Trivedi, 2005; Wooldridge, 2010). The second condition is also referred to as the "exclusion restriction" in that IVs do not directly impact the main dependent variable in the model, and hence should not be included in the main estimation (Cameron & Trivedi, 2005; Wooldridge, 2010).

2SLS earns its name from the two stages of estimations necessary to implement the procedure.[2] The first stage involves regressing each endogenous variable on all excluded instruments and exogenous variables in the main model. The second stage involves regressing the dependent variable on all the exogenous variables and the ***predicted values*** of endogenous variables from the first stage. Although ZKHL correctly describe the first stage, their description of the second stage is incorrect. They state that each first-stage regression "… residual is saved. In the second step, the dependent variable is regressed on the residual *in lieu* of the endogenous independent variable (p. 41, emphasis in original)" and go on to cite Bascle (2008) and Wooldridge (2010). However, this will not alleviate the endogeneity problem. Indeed, this method will result in biased and inconsistent estimates.

Rather than use the ***residuals*** from the first-stage regressions, 2SLS methodology uses the ***predictions*** of the endogenous covariates from the first-stage in the second-stage estimation (Angrist & Pischke, 2008; Bascle, 2008; Cameron & Trivedi, 2005; Greene, 2018; Wooldridge, 2010). The intuition is that the predictions of the endogenous covariate from excluded instruments and exogenous variables in the main model will capture the exogenous part of the offending variable. This results in consistent estimates since all regressors in the second stage are exogenous by construction. The method ZKHL describes would include *only* the endogenous part of the offending variable in the second stage and *leave out* the exogenous part, resulting in biased and inconsistent estimates. Indeed, we show in the online appendix[3] that ZKHL's method will *increase* bias above and beyond standard OLS (see the section titled "The Bias Due to ZKHL Methodology"). And with more exogenous instruments, the bias above and beyond OLS gets worse.

ZKHL motivate the discussion of 2SLS with an example where a researcher is interested in estimating the influence of "Trust" on "Supplier Performance." However, OLS regression of "Supplier Performance" on "Trust" may produce biased and inconsistent estimates because "Trust" is likely endogenous as it might be correlated with unobservables captured in the error term of an OLS estimation. We follow their lead and provide an example as well, along with data and estimation code, in order to demonstrate the difference between the incorrect methodology they describe and the correct methodology presented here, as well as Bascle (2008) and Wooldridge (2010), and the other graduate econometric textbooks cited above.

Say we want to estimate the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where y is the dependent variable; x_1 and x_2 are independent variables; β_0, β_1, and β_2 are coefficients to be estimated; ε is the econometric error term, and we believe x_1 is endogenous in that it is correlated with ε. In order to get consistent estimates of β_0, β_1, and β_2 using 2SLS we would need an additional variable, say z_1, that is correlated with x_1 (i.e. relevant), but is neither correlated with ε nor has a direct impact on y (i.e. valid).

The correct methodology is as follows: first, regress x_1 on x_2 and z_1 and obtain predicted values of x_1, say $\hat{x}_1$; second, regress y on $\hat{x}_1$ and x_2.

In the following example we demonstrate the correct methodology using a simulated dataset and compare with results using the incorrect methodology described in ZKHL. We assume a data generating process (DGP), which involves a dependent variable, y, two independent variables x_1 and x_2, and a latent variable *omVar*. The latent variable is not observed and not measured. Therefore, from all the estimations it will be omitted. The DGP for y is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 omVar + \eta \quad (1)$$

where $\beta_0 = 2$, $\beta_1 = 3$, $\beta_2 = 1$, $\beta_3 = 3$ and $\eta \sim N(\mu = 0, \sigma = 10)$. However, there is more to this DGP. We also put various restrictions on x_1, x_2, and *omVar*:

1. $x_2 \sim N(\mu = 0, \sigma = 1)$
2. $x_1 = z_1 + z_2 + omVar$; $z_1 \sim N(\mu = 0, \sigma = 1)$, $z_2 \sim N(\mu = 0, \sigma = 1)$, and $omVar \sim N(\mu = 0, \sigma = 1)$
3. $\text{cov}(z_1, z_2) = \text{cov}(z_1, omVar) = \text{cov}(z_2, omVar) = 0$
4. $\text{cov}(z_1, x_2) = \text{cov}(z_2, x_2) = \text{cov}(x_2, omVar) = \text{cov}(x_1, x_2) = 0$

Because we do not observe *omVar*, the model that we are estimating is:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\varepsilon} \quad (2)$$

Restrictions 2 states that x_1 is a function of z_1, z_2 and *omVar*, but *omVar* will be captured by the error term in the estimation, $\hat{\varepsilon}$, since it is never observed. This means that x_1 is correlated with $\hat{\varepsilon}$ (through *omVar*) and is therefore endogenous. Also, z_1 and z_2 are relevant instruments since they are related to x_1. Restriction 3 states that z_1 and z_2 are not correlated with each other or, more importantly, *omVar*. This means z_1 and z_2 are also valid instruments in that they are not correlated with the error term. Finally, restrictions 1 and 4 give the

[2] Though we note that modern econometric software (Stata, Eviews, etc.) typically performs 2SLS in one step.

[3] The online appendix is available at: https://www.ashwinmalshe.com/post/imm-2sls/

Table 1
Comparison estimations using simulated data: OLS vs. ZKHL proposed method vs. 2SLS vs. control function.

Estimated parameters	True parameter	Estimation method			
		OLS	ZKHL	2SLS	Control Function
Intercept	2	2.0868[a] (0.3351)	2.0538[a] (0.3438)	2.0774[a] (0.3400)	2.0774[a] (0.3324)
x_1	3	3.7352[a] (0.1945)	4.3416[a] (0.2505)	2.6805[a] (0.3267)	2.6805[a] (0.3194)
x_2	1	1.2846[a] (0.3521)	1.0807[a] (0.3611)	1.2270[a] (0.3575)	1.2270[a] (0.3495)
Control function correction					1.6611[a] (0.4008)
Bias in x_1	0	0.7352[a] (0.1945)	1.3416[a] (0.2505)	−0.3195 (0.3267)	−0.3195 (0.3194)
Unadjusted R^2		0.2750	0.2369	0.2536	0.2873
Adjusted R^2		0.2735	0.2353	0.2521	0.2851
N		1000	1000	1000	1000

Notes: Estimates without superscript are nonsignificant at $p = .1$. Standard errors in parentheses.
[a] Indicates $p \leq 0.01$

distributional characteristics of x_2 and state that it is not correlated with either the instruments ($z_1$and z_2) or *omVar* – the latter means that x_1 is exogenous in our setting.

We simulate a dataset with 1000 observations using the relationships and characteristics described above. We discuss the simulation in detail in the online appendix (see the section titled "Simulations"). The generated data is available from the authors on request. Table 1 shows several estimations using this data. We also display the True Parameters (i.e. the population parameters that we are trying to estimate) for reference. We note that we only use z_1 as an instrument for x_1 in the ZKHL, 2SLS, and Control Function estimations presented in Table 1 – we present additional results using both z_1 and z_2 as instruments in the online appendix (see the section titled "What if we have more than One Instrument for x_1?")

First note that the OLS estimate of the coefficient on x_1 (the endogenous variable) is significantly different than 3, the value of the population parameter.[4] Another way to see this is to look at the row titled "Bias in x_1 coefficient" – this is the difference of the estimated coefficient on x_1 from 3. A *t*-test to see if this coefficient is significantly different from 0 is the same as a *t*-test comparing the coefficient on x_1 to 3. They show that the OLS estimate is indeed biased. The goal with estimations labeled ZKHL, 2SLS, and Control Function are to show how each methodology addresses this bias.

The bias using the ZKHL estimate is worse than OLS at 1.3416. The estimated coefficient on x_1 is significantly different from 3 and actually moves further away to 4.3416.[5] In contrast, bias is essentially addressed using the appropriate 2SLS estimate. Here, the bias is not significantly different than 0[6] and the coefficient on x_1 is closest to 3 at 2.6805.

We note that the standard errors presented in the ZKHL and 2SLS estimations are incorrect and too small. This is because both estimations are done step-by-step in two stages. In taking this approach, we include a "generated regressor" in the second stage of the estimation (Pagan, 1984). Anytime you include an independent variable that was estimated from a previous regression it introduces more sampling variation into the estimation – something that is not accounted for in straightforward OLS standard error calculation (Wooldridge, 2010). This is typically not an issue with most standard econometric packages (Stata, Eviews, etc.) that have built-in procedures for 2SLS and other IV estimations which use the appropriate formulas to obtain the 'correct' standard errors. Further, these econometric packages can usually obtain standard errors that are robust to heteroskedasticity, clustering, autocorrelation, etc.

We bring up the problem of generated regressors because there may be sometimes when a researcher wants to estimate 2SLS or a related IV estimation by-hand like we do here. In these cases, it may be necessary to "bootstrap" to obtain asymptotically valid standard errors, t-statistics, F-statistics, and the like.[7]

For instance, while ZKHL incorrectly state that the residuals from the first-stage estimation are included *in lieu* of the endogenous variable, 2SLS coefficient estimates are also obtained when the residuals from the first-stage estimation are included *in addition* to the endogenous variable. Recall that the basic idea of the first-stage estimation is to separate the endogenous variable into the likely exogenous part, the predicted values, and the likely endogenous part, the residuals. Also, recall that endogeneity is caused by correlation with unobservables that are captured in the error term. The endogeneity problem is addressed if those unobservables can be observed and entered directly into the main estimation. This is exactly what is done by including the residuals from the first-stage estimation along with the endogenous variable in the main estimation. This approach is called the "control function" approach in that endogeneity bias is "controlled for" by including an estimation of the endogenous part of the offending variable separately from the variable itself in the main estimation.[8] In Table 1, the Control Function estimate is also presented where the row titled "Control function correction" is the estimated coefficient on the residuals from the first-stage estimation of x_1. The same coefficient estimates as 2SLS are obtained when the residuals from the first-stage estimation are included along with the endogenous variable in the Control Function column.

There are times when a researcher might want to consider the control function estimation in addition to traditional 2SLS. First, there is a readily available statistical test to see if endogeneity is actually a problem. The regression form of the Durbin-Wu-Hausman test for significant endogeneity bias in OLS estimates (Cameron & Trivedi, 2010; Wooldridge, 2010, 2015) is simply an F-test of the significance of the estimated coefficients on the residuals from the first-stage estimations. We do not go through all the specifics of the test here for brevity, but the intuition is that if the F-test indicates the coefficients on the residuals are significantly different from zero then endogeneity matters and 2SLS is preferred to OLS. However, if the F-test indicates that the residuals are not significantly different from zero then endogeneity bias

[4] A standard *t*-test gives a t-statistic = 3.78 with a *p*-value < .01.
[5] A standard *t*-test gives a t-statistic = 5.357 with a p-value < .01.
[6] A standard *t*-test gives a t-statistic = −0.87 with a p-value = .384.

[7] Cameron and Trivedi (2010) provide a useful and applied discussion of bootstrapping techniques along with code to implement in Stata and we direct the interested reader there.

[8] Heckman's (1979) two-step correction is a popular control function approach for addressing selection bias. We do not include a full discussion of selection bias or Heckman's (1979) correction here. However, we note that the inverse mills ratio (IMR) derived from the first-stage probit is the "generalized residual" (Gourieroux, Monfort, Renault, & Trognon, 1987) of the probit estimation. This generalized residual is included in the second-stage (i.e. the main model of interest) to address selection bias. Also, using IMRs obtained from a first-stage probit in the main model of interest is another common method to address endogenous treatment effects (Wooldridge, 2010) (see for example Kupfer, Pähler vor der Holte, Kübler, and Hennig-Thurau (2018)).

is not large enough to outweigh the other benefits of traditional OLS (e.g. "efficiency" of the OLS estimator). This test offers an advantage over the traditional Hausman test for endogeneity in that the test statistic can be made robust to heteroskedasticity. Also, we note that it is not necessary to bootstrap the second-stage estimation to obtain a valid Durbin-Wu-Hausman statistic (Wooldridge, 2010, 2015). In our example, the Durbin-Wu-Hausman test yields and F-stat of 17.177 ($p < .01$) which indicates that endogeneity of x_1 is a concern.

Second, the control function approach may be useful when a researcher is interested in the impact of the endogenous variable and how this impact may be moderated by other exogenous variables (see for example Saboo, Sharma, Chakravarty, and Kumar (2017)). Traditional 2SLS requires additional instruments for each interaction term involving the endogenous variable – the key idea being that the interaction terms are also endogenous because they include the endogenous part of the endogenous variable by construction. In applied work it can be quite difficult to find good instruments (i.e. instruments that are significantly correlated with the endogenous variable yet uncorrelated with the error term) for each endogenous variable. The control function approach offers a way to address endogeneity of interaction terms between several exogenous variables and an endogenous variable by including the residuals from a first-stage estimation of the non-interacted endogenous variable in the main estimation. This approach is less robust than 2SLS in that it imposes additional distribution assumptions on the first-stage estimation for consistency (Wooldridge, 2015), but can be a useful alternative when it is difficult to find instruments to separately identify interaction terms.

It is possible that the empirical results above are a function of the single random sample that we generated. After all, it could be that the random sample we used contained values that by pure happenstance resulted in traditional 2SLS outperforming ZKHL's methodology. To address this, we generated 1000 simulated dataset using the same parameters and distribution specifications described above, then ran OLS, ZKHL, and 2SLS estimations for each of these datasets. Fig. 1 shows the histograms of the estimated coefficients on x_1 from these estimations.

For a complete description of the simulations please see the online appendix (see the section titled "Simulations"). However, we note that only 2SLS estimates of the coefficient on x_1 center around 3, the population value, while OLS and ZKHL are biased and the bias for ZKHL is largest. We now move on to a discussion of 3SLS.

4. 3-stage least squares (3SLS)

We include a discussion of the control function approach above for two reasons. First, to highlight when a researcher would correctly include the residuals from a first-stage estimation (as opposed to the predicted values of the first-stage estimation). Second, to highlight that the discussion of 3SLS in ZKHL is incorrect and ultimately missing.

The 3SLS environment ZKHL describe is one where standard 2SLS (if applied appropriately) would address the endogeneity issues – in section 3.1.2 on page 42, the authors describe a situation with 2 endogenous variables: (1) "Trust" and (2) "Trust" interacted with another exogenous variable. As we discuss above, the interaction term is endogenous because it includes the endogenous part of "Trust". Both endogeneity issues could be addressed with appropriate instruments for each offending endogenous variable via standard 2SLS or with the control function approach we describe above.

For completeness, we include a description of the appropriate setting for 3SLS that is missing in ZKHL. 3SLS is typically used in a simultaneous equation framework where the dependent variable in one equation is an independent variable in another and vice versa. For example, let's say our goal is to recover the parameters on the variables in the following equations:

$$y_1 = 1 + 0.5y_2 + x_1 + 0.5z_1 + \varepsilon_1 \quad (3)$$

$$y_2 = 1 + 0.8y_1 + x_2 + 0.5z_2 + \varepsilon_2 \quad (4)$$

We do not include a full discussion of all the conditions necessary to identify the coefficients in the structural model shown in Eqs. (3) and (4) (see Greene (2018) for a full treatment), however typically at least one exogenous variable has to satisfy the exclusion restriction for each endogenous variable and the entire model has to be overidentified (i.e. more instruments than endogenous variables). In Eq. (3), x_2 and z_2 do not directly impact y_1 and therefore can be used as instruments to identify the impact y_2 has on y_1; similarly in Eq. (4), x_1 and z_1 do not directly impact y_2 and therefore can be used as an instrument to identify the impact y_1 has on y_2. Also, the model is overidentified in that there are more instruments for the entire model (four) than endogenous variables (two).

One way to approach this estimation is equation-by-equation 2SLS. That is, for Eq. (3) run a first-stage estimation of x_1, x_2, z_1, and z_2, on y_2 and obtain predicted values. Then regress y_1 on the predicted values from the first-stage estimation, x_1, and z_1. Similarly, for Eq. (4) – regress y_2 on the predicted values of y_1 from the first-stage estimation, x_2, and z_2. While this approach yields consistent estimates of the coefficients, a more "efficient" estimator (i.e. an estimator with a smaller sampling variance) is available that considers correlation of ε_1 and ε_2.

The 3SLS estimator has an additional step to equation-by-equation 2SLS. After 2SLS, estimate the correlation between the error terms in each equation, then use this information to compute a feasible generalized least squares (FGLS) estimator.[9] The FGLS estimator is more efficient than equation-by-equation 2SLS. Typically, equation-by-equation 2SLS is referred to as a "limited information" estimation while 3SLS is referred to as a "full information" estimation because it takes into account the additional information contained in the correlation of the error terms in each equation (Greene, 2018). We present examples of both estimations along with OLS estimates in Table 2. This is based on simulated data using the DGP described in Eqs. (3) and (4). We put restrictions on the distributions of x_1, x_2, z_1, and z_2 such that all of these are $N(0,1)$. Additionally, we let $\varepsilon_1 = 3\theta + 10r_1$ and $\varepsilon_2 = 3\theta + 10r_2$ where the distributions of θ, r_1, and r_2 are all also $N(0,1)$. The simulated dataset is available from the authors upon request.

Note that the coefficients on y_1 and y_2 in OLS estimation are both significantly biased. This is exactly the endogeneity issue that arrises from simultaneous causality – since y_1 impacts y_2 at the same time that y_2 impacts y_1, we can not obtain an unbiased and consistent estimate of the coefficients on either using simple equation-by-equation OLS. However, both 2SLS and 3SLS[10] lead to consistent estimates of the True Parameters. There is no significant bias in the coefficients on y_1 and y_2 using either equation-by-equation 2SLS or 3SLS. However 3SLS offers and advantage over 2SLS in that the estimates are more efficient – standard erros are a bit smaller for 3SLS compared to 2SLS. We emphasize that the amount of efficiency gain from 3SLS is dependent on a variety of factors (Greene, 2018) including the amount of correlation between the econometric error terms of the separate equations and the (lack of) correlation between the regressors of each equation. In practice, these efficiency gains can be quite significant (Cameron & Trivedi, 2010;Greene, 2018; Wooldridge, 2010). However, in the simulated setting here we focus on a simple example to show implementation of the method rather than conditions under which efficiency gains would be most dramatic.

5. Additional issues

We highlight two additional issues with ZKHL. These are not as significant as the incorrect description of standard 2SLS methodology,

[9] FGLS is a type of weighted least squares estimator. See Cameron and Trivedi (2005), Greene (2018) and Wooldridge (2010) for full discussions and applications to simultaneous equation models.

[10] 3SLS estimates are produced using the System Fit package in R.

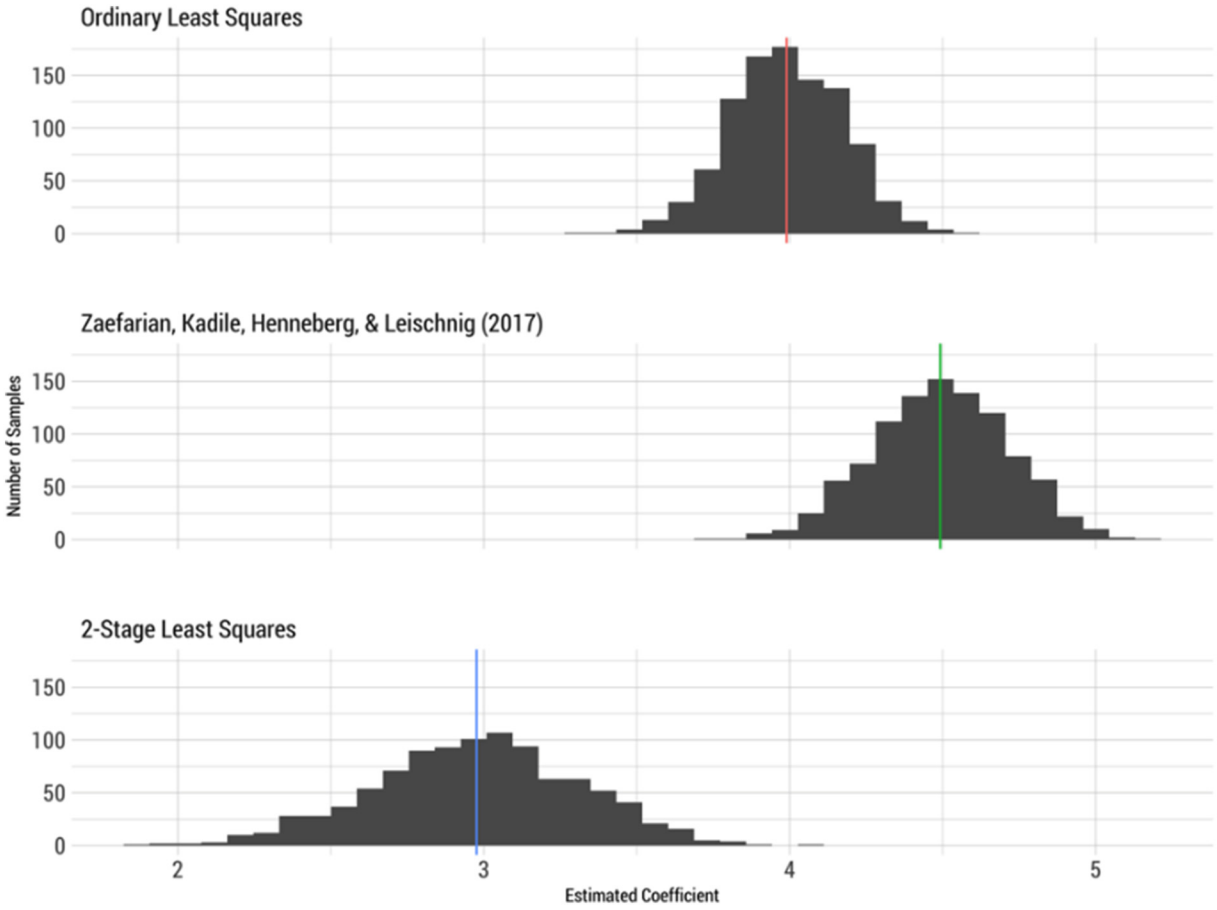

Fig. 1. Histograms of estimated coefficients on using OLS, ZKHL methodology, and 2SLS on 1000 simulated datasets.

Table 2
OLS, 2SLS, and 3SLS Estimation of Simultaneous Equation System (3) and (4).

Estimated parameters	True parameter	Estimation Method					
		OLS		2SLS		3SLS	
		Eq. (3)	Eq. (4)	Eq. (3)	Eq. (4)	Eq. (3)	Eq. (4)
Intercept	1.0	0.2227 (0.2452)	0.1617 (0.2888)	2.1980 (1.4700)	1.6866[c] (0.9389)	2.1985 (1.4700)	1.6863[c] (0.9388)
y_1	0.8		1.1120[a] (0.0121)		0.6249[b] (0.2607)		0.6249[b] (0.2607)
y_2	0.5	0.8018[a] (0.0087)		0.2469 (0.3833)		0.2466 (0.3832)	
x_1	1.0	0.5484[b] (0.2424)		0.9253 (0.6040)		0.9150[c] (0.5269)	
x_2	1.0		0.9442[a] (0.2813)		1.1142[b] (0.4652)		1.1172[b] (0.4516)
z_1	0.5	0.6928[a] (0.2425)		1.2112[c] (0.6522)		1.2194[b] (0.6080)	
z_2	0.5		0.3514 (0.2871)		0.4313 (0.4676)		0.4231 (0.3539)
Bias in y_1			0.3120[a] (0.0121)		−0.1751 (0.2607)		−0.1751 (0.2607)
Bias in y_2		0.3018[a] (0.0087)		−0.2531 (0.3833)		−0.2534 (0.3832)	
Unadjusted R^2		0.8950	0.8950	0.4691	0.7237	0.4686	0.7237
Adjusted R^2		0.8947	0.8947	0.4675	0.7229	0.4670	0.7229
N		1000	1000	1000	1000	1000	1000

Note: Estimates without superscript are nonsignificant at p = .1. Standard errors in parentheses.
[a] $p \leq 0.01$.
[b] $p \leq 0.05$.
[c] $p \leq 0.1$.

or the ultimately missing description of 3SLS, but they are important none the less. First, the description of generalized method of moments (GMM) estimation in ZKHL is rather limited and appears to apply only to dynamic panel data techniques. While the use of GMM in dynamic panel analysis certainly is popular (see for example Chung (2017) and Oberholzer-Gee and Strumpf (2007)), it is important to note that GMM encompasses a class of estimators of which 2SLS and 3SLS are special cases (Greene, 2018). Single equation GMM estimates and GMM estimates of simultaneous equations may offer an advantage over 2SLS and 3SLS counterparts in that they are more efficient in the presence of arbitrary heteroskedasticity (Greene, 2018; Wooldridge, 2010). Also, it is typically quite easy to implement GMM counterparts to 2SLS and 3SLS with modern econometric packages (Cameron & Trivedi, 2010).

Second, the Hansen's *J*-statistic (Hansen, 1982) mentioned in ZKHL

is not a direct test of instrument orthogonality to the econometric error term. Instrument validity is not empirically verifiable (Rossi, 2014), rather the applied researcher must rely on theoretical arguments as to why an instrument is likely not correlated with the econometric error term. Indeed, the Hansen's *J*-statistic, and other related statistics, can only be obtained if a model is "overidentified" in that more instruments exist than endogenous variables.[11] In this case, the Hansen's *J*-statistic is basically a test to see if "extra instruments" are correlated with the error term. A nonsignificant test statistic is merely additional support of the theoretical arguments, but ultimately the validity of the test relies on the model being appropriately specified (Cameron & Trivedi, 2010). In other words, the test is susceptible to false negatives and a researcher should be wary of taking an insignificant *J*-statistic as strong evidence of instrument validity.

6. Conclusion

Endogeneity is a serious issue that can impact causal interpretation in applied work. As ZKHL note, marketing scholars are taking notice and attempting to address endogeneity issues more and more. While we disagree with ZKHL's statement that "the reported estimates are rarely statistically significant (p. 41)" when referring to the use of 2SLS in applied work – a cursory look at recent issues of *Journal of Marketing*, *Journal of Marketing Research*, *Marketing Science*, *Management Science* suggests this is not the case – this is a subjective assessment and we would not dispute their subjective interpretation of the literature. However, their description of 2SLS, the work-horse model used to deal with endogeneity in applied work, as well as 3SLS, are objectively incorrect and, if followed, would lead an applied researcher down the wrong path, exacerbating any endogeneity issue. We recommend that researchers follow the procedures outlined here or in the many econometrics text books cited in this piece rather than the procedure suggested in ZKHL.

Acknowledgement

This paper was reviewed by the IMM editor and did not undergo the regular reviewing process.

[11] We provide an example of an overidentified 2SLS model in the online appendix (see the section titled "What if we have more than One Instrument for x_1?").

References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An Empiricist's companion.* Princeton and Cambridge: Princeton University Press.

Bascle, G. (2008). *Controlling for endogeneity with instrumental variables in strategic management research. 6 (3)*, Strategic Organization285–327. https://doi.org/10.1177/1476127008094339.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications.* Cambridge: Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics Using Stata (Revised Edition).* College Station, TX: Stata Press.

Chung, D. J. (2017). How much is a win worth? An application to intercollegiate athletics. *Management Science, 63*(2), 548–565. https://doi.org/10.1287/mnsc.2015.2337.

Gourieroux, C., Monfort, A., Renault, E., & Trognon, A. (1987). Generalised residuals. *Journal of Econometrics, 34*(1–2), 5–32. https://doi.org/10.1016/0304-4076(87)90065-0.

Greene, W. H. (2018). *Econometric analysis* (8th ed.). New York, NY: Pearson.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50*(4), 1029–1054. https://doi.org/10.2307/1912775.

Heckman, J. J. (1979). Sample selection Bias as a specification error. *Econometrica, 47*(1), 153. https://doi.org/10.2307/1912352.

Kupfer, A.-K., Pähler vor der Holte, N., Kübler, R. V., & Hennig-Thurau, T. (2018). The role of the partner brand's social media power in brand alliances. *Journal of Marketing, 82*(3), 25–44. https://doi.org/10.1509/jm.15.0536.

Oberholzer-Gee, F., & Strumpf, K. (2007). The effect of file sharing on record sales: An empirical analysis. *Journal of Political Economy, 115*(1), 1–42. https://doi.org/10.1086/511995.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review, 25*(1), 221. https://doi.org/10.2307/2648877.

Rossi, P. E. (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.

Saboo, A. R., Sharma, A., Chakravarty, A., & Kumar, V. (2017). Influencing acquisition performance in high-technology industries: The role of innovation and relational overlap. *Journal of Marketing Research, 54*(2), 219–238. https://doi.org/10.1509/jmr.15.0556.

Semadeni, M., Withers, M. C., & Trevis Certo, S. (2014). The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations: Research notes and commentaries. *Strategic Management Journal, 35*(7), 1070–1079. https://doi.org/10.1002/smj.2136.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western, Cengage Learning.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources, 50*(2), 420–445. https://doi.org/10.3368/jhr.50.2.420.

Zaefarian, G., Kadile, V., Henneberg, S. C., & Leischnig, A. (2017). Endogeneity bias in marketing research: Problem, causes and remedies. *Industrial Marketing Management, 65*, 39–46. https://doi.org/10.1016/j.indmarman.2017.05.006.