Correcting for Endogeneity in Hospitality and Tourism Research

A. George Assaf
Professor
Isenberg School of Management
University of Massachusetts-Amherst
90 Campus Center Way, 204A Flint Lab, Amherst, MA, 01003
Tel :(+1) 4135451492
assaf@isenberg.umass.edu

Mike G. Tsionas
Professor of Econometrics
Lancaster University Management School
m.tsionas@lancaster.ac.uk

Abstract

Purpose

This paper aims to foster a new discussion on endogeneity in hospitality and tourism research.

Design/methodology/approach

This paper elaborates on some of the common sources of endogeneity as well as the methods available to address them.

Findings

We present a variety of methods that can be used to mitigate the endogeneity problem. We provide simulation evidence regarding the risk of incorrectly selecting instrumental variables. We also provide several important practical recommendations for future research.

Research limitations/implications

There are other issues and methods of correcting for endogeneity that are not covered in this paper. However, the paper focuses on issues and methods that can be generalized to most contexts.

Originality/value

The paper provides practical recommendations for more rigorous regression estimation.

1. Introduction

While quantitative modelling in hospitality and tourism research has progressed significantly over the last two decades, an important issue that remains largely ignored is the failure to account for- or address endogeneity-related issues (Assaf and Tsionas, 2019). The danger that endogeneity imposes in the estimation of regression models has been well documented in the literature. Endogeneity arises when the regressors are endogenous in the sense that they are correlated with the error term, leading to biased and inaccurate conclusions about cause and effect relationships. In fields like hospitality and tourism, which rely heavily on regression-related models, the risk of ignoring endogeneity becomes an even greater concern.

This topic is certainly worthy of an in-depth discussion for those in these fields. Authors and reviewers would benefit from such a discussion to better understand the risks associated with endogeneity instead of simply stating that endogeneity is dangerous or even dismissing an otherwise good paper simply due to the fear associated with tackling these challenges (Rutz and Watson, 2019). Our aim in this note is to contribute to this discussion. We revisit the sources of- and problems with endogeneity, and provide simulation evidence regarding the risk of incorrectly selectng instrumental variables. We also discuss what methods are available to deal with endogeneity. We differentiate between methods that require instrumental variables and others that are instrument-free. Finding the appropriate instruments is always challenging. We encourage increased use of instrument-free methods; which, despite their flexibility, remain largely uncommon in these fields.

We emphasize again that the goal with this paper is not to provide a lengthy textbook discussion of endogeneity. Rather, we aim to foster a new discussion on the topic and provide clearer guidelines on how to deal with endogeneity. While there is no perfect solution for endogeneity, revisiting the issue and presenting some of the most robust approaches to deal with it may encourage more serious thinking about endogeneity, both in terms of theoretical design and estimation of regression models.

2. Sources of Endogeneity

Most empirical work in hospitality and tourism studies focuses on testing the relationship between a set of independent variables (X) and a dependent variable of interest (Y). An example would be models for estimating hotel demand that use price as an exogenous variable. Often these models ignore the fact that various events could have taken place in the city where the hotel is located, which if not accounted for, result in the reporting of a biased effect of price on demand. Hotels also belong to various quality categories, which are often not accounted for in the model.

Across many contexts, regression models in the field are vulnerable to endogeneity problems. We rarely see papers comparing between models with endogeneity controls and those without. As recently emphasized by Rutz and Watson (2019, p.482): "the first step to address endogeneity is to understand the potential sources (s) that apply given the research setting, its data, and the modelling approach chosen by the researcher or manager". In the next section, we will elaborate on some of the common sources of endogeneity that are often ignored in hospitality and tourism research.

2.1. Errors in Variables

Quite often, variables are measured with error. Although the regression model accommodates errors in the dependent variable, when errors are present in the regressors, then the observed regressor and the error are correlated. In this example, when we refer to errors in variables, we are not talking about errors in the dependent variable, as this can be captured by the error term in the regression model. Here we are mainly focusing on errors in the independent variables. For example, if a researcher is testing the effect of internationalization on hotel performance, and there is an error in the measurement of the "performance" variable, this will not lead to biased estimates as such error will be captured by the regression error. However, an error in the measurement of the "internationalization" variable will lead to endogeneity problems, and thus to biased regression estimates.

To elaborate further, consider the simple model

$$\begin{aligned} y_t &= \beta x_t^* + u_t, \\ x_t &= x_t^* + v_t, \end{aligned} \tag{1}$$

where v_t is measurement error in the regressor x_t^* which is unobserved. For simplicity, we can assume that u_t and v_t are orthogonal. Substituting the second equation into the first, we have:

$$y_t = \beta x_t + (u_t - \beta v_t). \tag{2}$$

Although this equation involves only observed variables, x_t and the error $e_t = u_t - \beta v_t$ cannot be uncorrelated, as x_t and v_t are correlated. Ordinary Least Squares (OLS) relies on the assumption that x_ts can be taken as given (or orthogonality with the errors can be assumed). This involves the implicit assumption that errors of regression can be measured by fixing x and measuring the difference between actual and predicted ys. Under errors in variables this device fails to deliver consistent parameter estimators. Both variables are subject to error and, therefore, fixing one or the other and measuring distances is problematic, just as suggested by Bartlett's (1949) classic paper.

One of the earliest attempts to derive a consistent estimator in the presence of measurement error was done by Wald (1940) who suggested ordering the observations and then splitting them into two groups. The estimate of the slope is then the difference of average values of ys divided by the difference of average values of xs. Another early remedy for errors in variables is total least squares (TLS). Unlike OLS, TLS suggests that one can use the *orthogonal distance* from a point to the regression line.

The problem of TLS has a long history starting with the work of Adcock (1878) who made the assumption that $\sigma_u^2 = \sigma_v^2$. It can be shown that the objective function of TLS is:

$$\min_{\beta} Q\,(\beta) \equiv \frac{\sum_{t=1}^{n} (y_t - x'_t\beta)^2}{1 + \sum_{j=1}^{k} \beta_j^2}. \tag{3}$$

The OLS estimator is $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$. One can show that the TLS estimator has a closed form given by:

$$\hat{\beta}_{TLS} = (X'X - \omega^2 I_k)^{-1} X'y, \quad (4)$$

where ω^2 is the smallest singular value of $[X\ y]$. Therefore, the TLS estimator is reminiscent of the ridge estimator but the constant ω^2 is subtracted from instead of added to the main diagonal of the cross-products matrix $X'X$. Despite the fact that the asymptotic bias of OLS is removed by subtracting ω^2 from the main diagonal, it is also the case that $X'X - \omega^2 I_k$ is more ill-conditioned compared to $X'X$. For a modern treatment of TLS see Van Huffel and Vandewalle (1991).

2.2. Simultaneity

Another issue that is often ignored in hospitality and tourism research is simultaneity, which occurs when we have the so-called "reverse causality" phenomenon; not only when x_t causes y_t but also if the opposite is true. One can think of many scenarios in tourism and hospitality research where reverse causality can be a potential problem. For example, when testing the impact of tourism spending on economic growth, one can argue that higher economic growth can also lead to higher tourism spending. In such a case, ignoring simultaneity would result in the error term being correlated with the exogenous variables, leading to an endogeneity problem and biased regression estimates.

More specifically, simultaneity takes the following form:

$$\begin{aligned} y_t &= \beta x_t + \gamma_1 z_{t1} + u_t, \\ x_t &= \alpha y_t + \gamma_2 z_{t2} + v_t, \end{aligned} \quad (5)$$

where z_{t1} and z_{t2} are two predetermined variables. The source of the problem is that, in the reduced form, which expresses y_t and x_t in terms of the errors and the predetermined variables, x_t is correlated with both u_t and v_t so estimating the first equation by OLS produces an inconsistent estimator for β. The same is true for the OLS estimator of α in the second equation of (5). To put it differently, the key challenge when estimating a regression model, where simultaneity is a problem, is to "disentangle the temporal order" in which these variables affect each other (Rutz and Watson, 2019). The two variables cause each other, creating this correlation between the error term and the explanatory variables, which violates the OLS assumption.

2.3. Omitted Variable Bias

A final source of endogeneity, and one of the most challenging to test, relates to the omission of explanatory variables. Such omission can lead to endogeneity problems when the omitted variable is correlated with the outcome variable or any of the explanatory variables in the model. In tourism, one can think of many contexts of omitted variable problems. In the estimation of tourism demand models, for instance, researchers often struggle with unavailable data to account for economic and social structures, thus creating bias in the results due to the omitted variable problem (Kuo et al. 2009).

Generally, the omitted variable problem can be expressed with the following. Suppose that the model is given by:

$$y_t = \beta x_t + \gamma z_t + u_t, \quad (6)$$

but we omit z_t and, in effect, we estimate the model $y_t = \alpha x_t + v_t$,where $v_t = \gamma z_t + u_t$. If x_t and z_t are correlated and $\gamma \neq 0$, then the OLS estimator of α is not consistent for β. This is a source of endogeneity in the sense that $\mathbb{E}(x_t v_t) \neq 0$. Of course, omitted variables abound in practical situations and, therefore, there is almost always the risk of inconsistent parameter estimates.

As mentioned above, omitted variable bias is one of the most challenging problems to diagnose. This, of course, excludes the "naive" case where, post regression, one wishes to test whether certain variables have been omitted. The omitted variables problem per se, means that although there are, indeed, omitted variables, it is at the same time unknown which variables these may be. However, given the expression $v_t = \gamma z_t + u_t$ in (6), some remarks can be made. If the omitted variable z_t is autocorrelated, heteroskedastic, or both, then standard OLS residual-based tests will indicate the presence of autocorrelation and / or heteroskedasticity. In this instance, the tests do not support a "knee jerk" reaction like correcting for AR(1) or even worse relying on robust standard errors. Clearly, correcting for AR(1) type of autocorrelation or AR(p) in general if z_t follows an AR(q) process (with $p \geq q$) can mitigate the omitted regressor problem. Things are more involved in the case of heteroskedasticity. However, reliance on robust standard errors is not justified in this case, as OLS estimates are inconsistent to begin with and Heteroskedasticity and Autocorrelation Consistent (HAC) techniques require the correct specification of the model. When the omitted variable z_t is also correlated with the regressor x_t it is natural to suppose that $z_t = f(x_t) + w_t$; where $f(\cdot)$ is an unknown functional term and w_t is an error uncorrelated with u_t. The standard Ramsey Regression Equation Specification Error Test (RESET) for neglected nonlinearity in the regression by regressing OLS residuals on $\hat{y}_t^2, \hat{y}_t^3$ etc., where $\hat{y}_t = x'_t\hat{\beta}_{OLS}$ are the fitted values of the OLS-estimated regression. A Taylor series expansion of $f(\cdot)$ shows, in fact, that the RESET also tests for omitted variables, under the assumption that the omitted variables are functionally related to x_t and the functional form of the dependence is, in itself, unknown, but is smooth enough to be approximated by a Taylor expansion. Thus, again, the RESET test can be interpreted as a diagnostic for misspecification. Therefore, in practice, the empirical researcher has this standard diagnostic test at their disposal to test for misspecification. We again emphasize that these diagnostics are useful to the extent that the omitted regressor, z_t, is autocorrelated, heteroskedastic and / or functionally related to the included regressors up to the measurement errors.

3. Approaches to Control for Endogeneity

We discuss in this section different methods that are available for hospitality and tourism researchers to control for endogeneity. Some of these methods require the availability of instruments, an often challenging issue, which is illustrated through simulation examples below. Other methods do not require the availability of instruments, and hence free the researcher from such challenges. We note that despite their flexibility, these methods remain uncommon in hospitality and tourism research. Also, while we focus in what

follows on linear models, we discuss in Appendix A the context of non-linear moels.

3.1. Instrumental Variables

It is often the case that researchers resort to the use of instrumental variables to control for endogeneity. For example, in studying the effect of air services on tourism demand, Koo et al. (2017) used the degree of air liberalization and total available flights are used as instrumental variables. However, it can be challenging to find instruments. An instrumental variable also needs to meet the strict condition of being correlated with the endogenous variable, while also not being correlated with the error term. Consider the following linear model:

$$y_t = x'_t\beta + u_t,\ t = 1, \dots, n, \tag{7}$$

where x_t is a $K \times 1$ vector of regressors. For the method of OLS to yield consistent estimators, we need the assumption that the regressors and the error term are uncorrelated, meaning $E(x_t u_t) = 0$ for all $t = 1, \dots, n$. In this case, we often write: $x_t \perp u_t$ and we say, alternatively, that the regressors and errors are orthogonal. When $K = 1$, the assumption $E(x_t u_t) = 0$ yields the least square estimator as follows. From (7), if we multiply both sides by x_t and take sample averages we have:

$$n^{-1}\sum_{t=1}^{n} x_t\, y_t = n^{-1}\sum_{t=1}^{n} x_t^2\, \beta + n^{-1}\sum_{t=1}^{n} x_t\, u_t. \tag{8}$$

Since $\text{plim}(n^{-1}\sum_{t=1}^{n} x_t\, v_t) = E(x_t u_t) = 0$, this equation can be solved to yield $\hat{\beta}_{LS} = \frac{\sum_{t=1}^{n} x_t y_t}{\sum_{t=1}^{n} x_t^2}$. Now, if $\sim (x_t \perp u_t)$ but there is a certain instrumental variable z_t for which we have $z_t \perp u_t$, then we can follow the same strategy as in (8) to obtain:

$$n^{-1}\sum_{t=1}^{n} z_t\, y_t = n^{-1}\sum_{t=1}^{n} z_t\, x_t\beta + n^{-1}\sum_{t=1}^{n} z_t\, u_t. \tag{9}$$

Since $\text{plim}(n^{-1}\sum_{t=1}^{n} z_t\, v_t) = E(z_t u_t) = 0$ the IV estimator:

$$\hat{\beta}_{IV} = \frac{n^{-1}\sum_{t=1}^{n} z_t\, y_t}{n^{-1}\sum_{t=1}^{n} z_t\, x_t}, \tag{10}$$

will be consistent. However, a problem arises in that the denominator must be kept away from zero, that is we need:

$$\text{plim} n^{-1}\sum_{t=1}^{n} z_t\, x_t = E(z_t x_t) \neq 0, \tag{11}$$

which means that the instrument z_t must be uncorrelated with the error term but nevertheless it should be strongly correlated with x_t.

In ad hoc models, it is not clear where instruments such as z_t come from unless one exercises his/ her ingenuity to come up with such variables. In simultaneous equations models, however, all exogenous variables can serve as instruments provided the two conditions we mentioned are true. Suppose we have M instrumental variables, $z_{t1}, \dots, z_{tM}$.

In turn, one may be able to write down moment conditions of the form:

$$n^{-1}\sum_{t=1}^{n}(y_t - x'_t\beta)\, z_{tm} = 0,\ m = 1, \ldots, M, \tag{12}$$

where it is quite likely that $M > K$ so that we have $M - K$ over-identifying restrictions. Since we have K elements in β, K instruments would be enough to just identify β. However, it is commonly the case that we have more instruments than parameters. If we use (7), then (12) would imply:

$$Z'y = Z'X\beta + Z'u, \tag{13}$$

where Z is the $n \times M$ matrix of instruments. As by assumption $\text{plim} n^{-1}Z'u \to E(Z'u) = \mathbf{0}$, a simple IV estimator is given by applying LS to (13):

$$\hat{\beta}_{IV} = (X'ZZ'X)^{-1}X'ZZ'y, \text{provided } M \geq K. \tag{14}$$

However, in (13), the errors $e = Z'u$ have zero mean and covariance matrix $\sigma_u^2(Z'Z)^{-1}$, provided $E(uu') = \sigma_u^2 I_n$. In this case, we can use the following estimator, which results from GLS applied to (13):

$$\check{\beta}_{IV} = (X'Z(Z'Z)^{-1}Z'X)X'Z(Z'Z)^{-1}Z'y, \text{if} M > K. \tag{15}$$

Another implication of (13) is that $\text{cov}(Z'u) = E(Z'uu'Z) \equiv \Omega$ has dimension $m \times m$, which does not increase with the sample size. So, if Ω can be estimated in advance, the estimator:

$$\tilde{\beta}_{IV} = (X'Z\Omega^{-1}Z'X)X'Z\Omega^{-1}Z'y \tag{16}$$

would not only be consistent, but efficient as well. Matrix Ω can be estimated as in the HAC procedures available in most software packages. For example, in the case of heteroskedasticity of unknown form, we can estimate Ω by:

$$\widehat{\Omega} = Z'VZ, \tag{17}$$

where $V = \text{diag}[\hat{u}_1^2, \ldots, \hat{u}_n^2]$, where the residuals $\hat{u}_t = y_t - x'_t\hat{\beta}_{IV}$, $t = 1, \ldots, n$.

A regression model can be estimated using the method of Two Stage Least Squares (2SLS), provided instruments z_t are available. The method consists of regressing all endogenous variables in x_t on z_t, obtaining the fitted values, $\hat{x}_t$, and then, either using $\hat{x}_t$ as instruments for x_t or replacing x_t by $\hat{x}_t$. Both techniques result in the same numerical estimates. In that respect, 2SLS is a special case of IV estimation and also a special case of GMM (see Appendix B).

Regardless of the estimation method, one of the main challenges that remain is the availability of suitable instruments. The selection of invalid or weak instruments can result in strong biases and erratic behavior in finite samples. We illustrate this issue further using the following simulation. In line with Kleibergen and van Dijk (1994, 1998) and van Dijk (2002), we consider an Incomplete Simultaneous Equations Model (INSEM):

$$y_1 = \beta y_2 + (\phi v_2 + v_1), \tag{18}$$

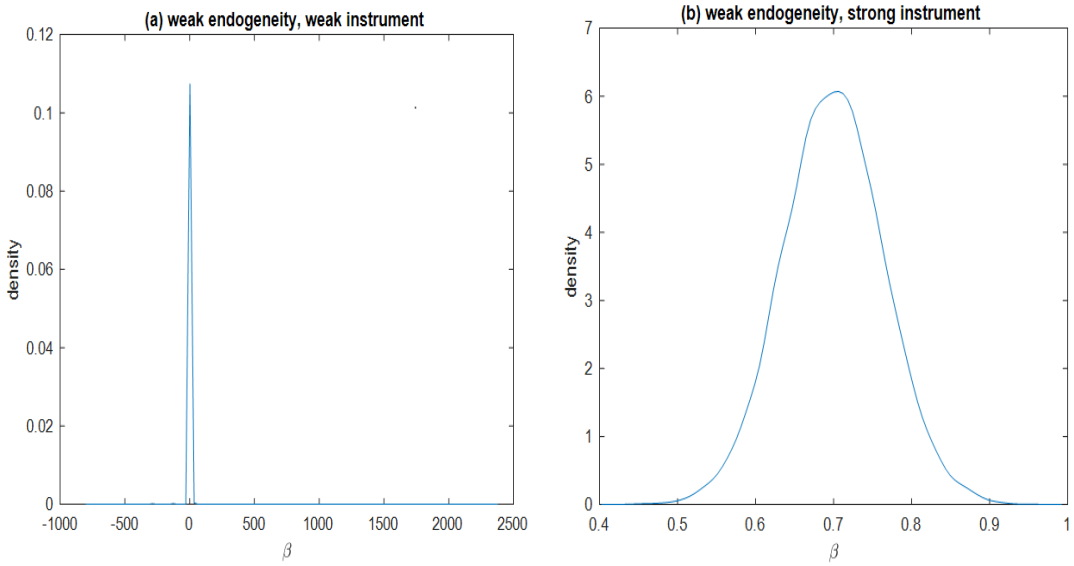
$$y_2 = \pi x + v_2. \tag{19}$$

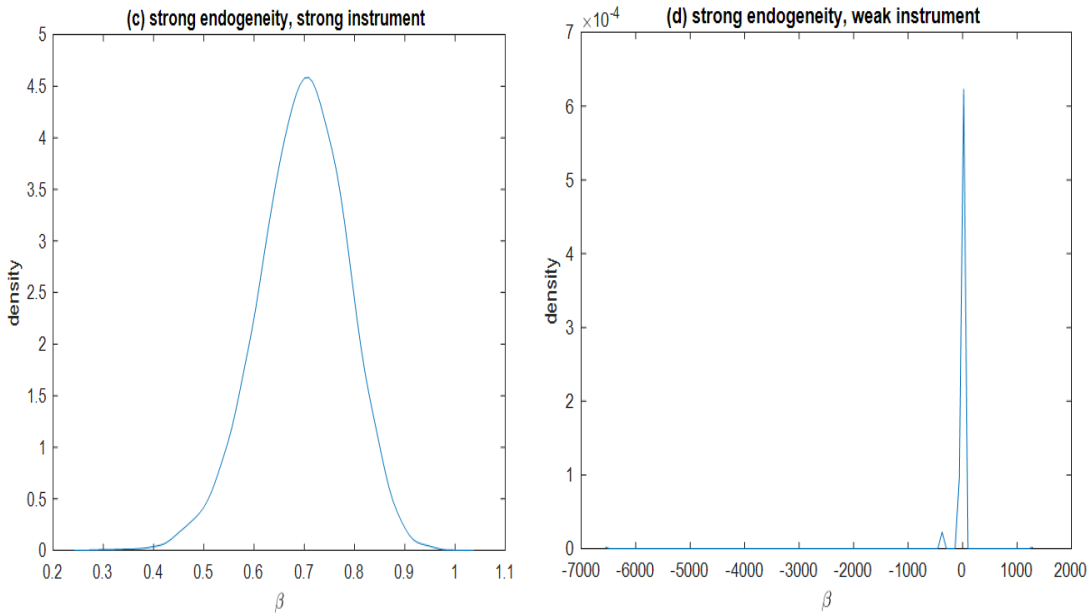
In this model, we have strong endogeneity if $\phi \simeq 1$; assuming we interpret it as a correlation coefficient and instrument x is weak when $\pi \simeq 0$. We use a sample of 250 observations, where v_1 and v_2 have independent standard normal distributions, ϕ, π take two values (0.9 or 0.025) and $\beta = 0.7$. We use 10,000 simulations for the instrumental variables (IV) estimator:

$$\hat{\beta}_{IV} = \frac{x'y_1}{x'y_2}. \tag{20}$$

The sampling distributions are presented in Figure 1.

Figure 1. Sampling distributions of IV estimator





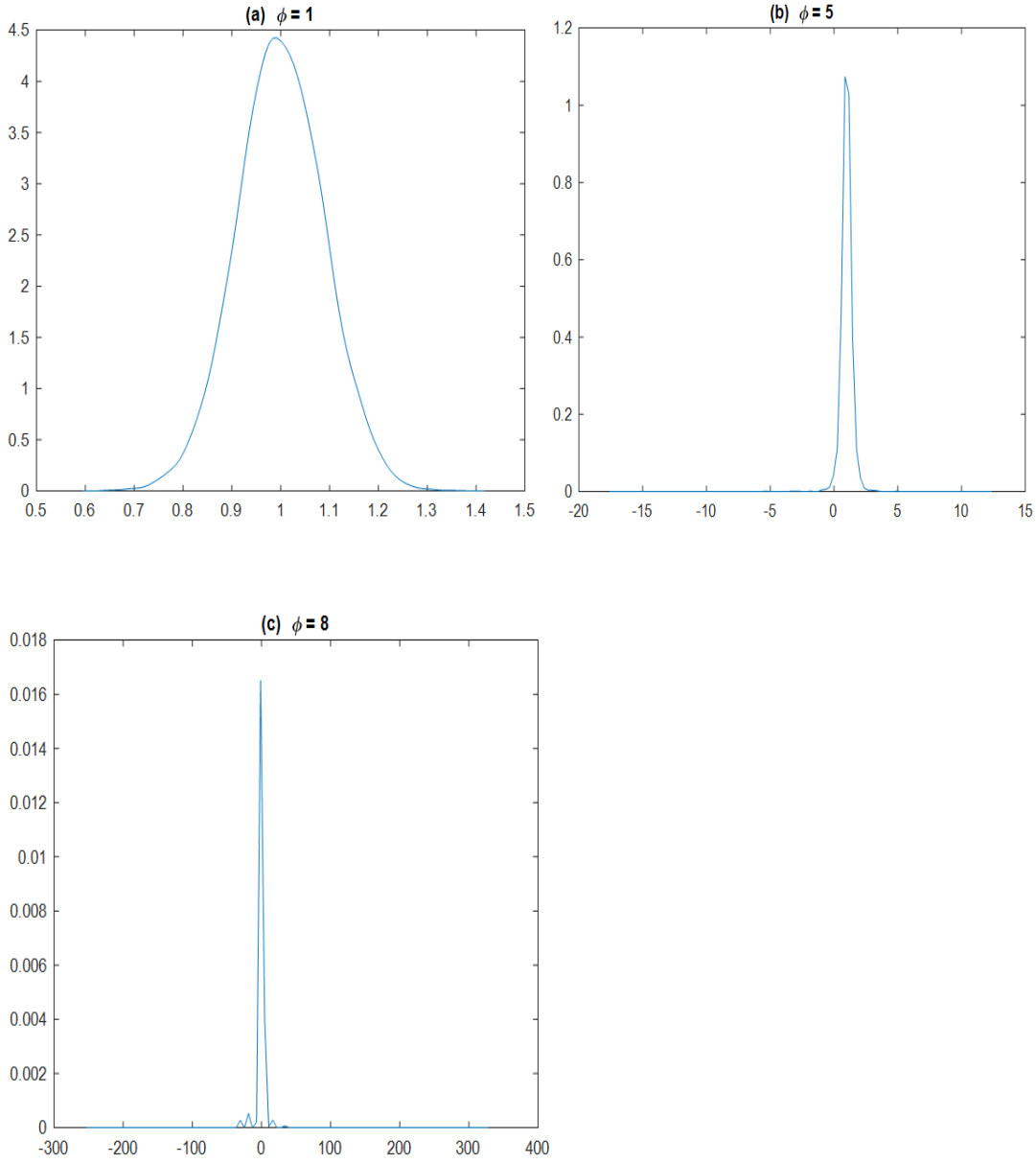
In the case of weak instruments, it is clear that the sampling distribution of the IV estimator has fat tails (case (a)) or diverges to $\pm\infty$ (case (d)). When the instrument is strong as in panels (b) and (c), the sampling distribution shows that the IV estimator performs relatively well.

To establish these properties on firmer ground, suppose we have:

$$y_t = \beta x_t + u_t,\ t = 1, \dots, n, \tag{21}$$

where $u_t \sim iid\, N(0,1)$, $\beta = 1$, $x_t = u_t + e_t$, $e_t \sim iid\, N(0, 0.1^2)$, which implies strong correlation between the regressor and the error term. We generate the instrument as $z_t = x_t + \phi\xi_t$, $\xi_t \sim iid\, N(0,1)$ so, by construction, it is correlated with x_t but this correlation depends on ϕ (as $\phi \to 0$ the correlation is strong). We use three values for ϕ, viz. $\phi = 1$, 5, and 8. The results are shown in Figure 3.

Figure 2. Sampling distributions of IV estimator



In Figure 2, although the sampling distributions are centered around the true value of unity, as φ increases, so that we have less correlation with the included regressor, these distributions have extremely fat tails. This means that in finite samples, it is quite likely to obtain answers that are (very) far from the truth. More details on GMM in nonlinear models are provided in Appendix B.

3.2. Control function (CF) approach

The CF approach is another approach that requires instruments and is similar in spirit to the IV approach. The model is:

$$\begin{aligned} y_t &= \beta x_t + u_t, \\ x_t &= z_t\gamma + v_t, \end{aligned} \tag{22}$$

where z_t is an instrument. Given the linear projection $u_t = \rho v_t + \xi_t$, where ξ_t is an i.i.d error, where $\rho = \frac{\mathbb{E}(u_t v_t)}{\mathbb{E}(v_t^2)}$. Since $\mathbb{E}(v_t \xi_t) = 0$, it follows that we can use the model:

$$y_t = \beta x_t + \rho(x_t - \hat{\gamma} z_t) + \xi_t. \tag{23}$$

Therefore, we can use a two-step approach in which x_t is regressed on z_t and we obtain the OLS residuals. These residuals are included in the model and a consistent estimator of β is obtained via OLS. This method can be understood in a broader sense in nonlinear or limited dependent variable models; specifically, that potentially endogenous variables are regressed on instruments and the residuals are included in the model. This method has probably been reinvented numerous times. For example, see Terza (2018), who used the term "two stage residual inclusion" (2SRI) which has been used in certain nonlinear models (see Tran and Tsionas (2013).

This approach cannot be used when the relationship between x_t and z_t is nonlinear, that is when, for example:

$$\boldsymbol{x}_t = \Pi(\boldsymbol{z}_t; \boldsymbol{\gamma}) + \boldsymbol{v}_t, [u_t, \boldsymbol{v}_t{}']' \sim \mathcal{N}(0, \Sigma), \tag{24}$$

where $\boldsymbol{x}_t \in \mathbb{R}^k, \boldsymbol{z}_t \in \mathbb{R}^m$, and $\boldsymbol{\gamma} \in \mathbb{R}^d$ is a parameter vector, and $\Pi: \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^k$ is a nonlinear functional form (a vector field). Research in this area is rather limited, as specifying $\Pi(\cdot;\cdot)$ is, more often than not, impossible. Although, there have been some papers which are motivated explicitly by economic theory (for example, first-order conditions to cost minimization or profit maximization) and often involve latent variables in a nonlinear way (Atkinson and Tsionas, 2016, Atkinson, Primont, and Tsionas, 2018, Tsionas and Mamatzakis, 2019).

3.3. Latent Instrumental Variables Estimation (LIVE)

As mentioned, one of the challenges with the standard IV method or the Control Function approach is the need for instruments, which are often neither strong nor valid. LIVE is a method that frees the researcher from such challenges. It is an instrument-free approach that addresses the endogeneity problem (see Ebbes et al., 2011, 2016, and Papies et al. 2017) without needing access to instrumental variables. Specifically, LIVE considers the following model:

$$\begin{aligned} y_t &= \beta x_t + u_t, \\ x_t &= z_t^*\gamma + v_t, \end{aligned} \tag{25}$$

where z_t^* are latent or constructed instruments. The second equation is a reduced form.

We assume:

$$\begin{bmatrix} u_t \\ v_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}\right), \tag{26}$$

and also $\mathbb{E}(z_t^* u_t) = \mathbb{E}(z_t^* v_t) = 0$. With LIVE, z_t^* are the unobserved categorical variables arising from a multinomial distribution where the probability of the jth category is, say, π_j with the normalization $\sum_j \pi_j = 1$ and $\pi_j > 0$. More specifically, Ebbes et al. (2005) suggest a latent discrete binary variable for the decomposition of a single distributed endogenous regressor into two components, one uncorrelated with the error term and a and another potentially correlated with the error term. The case of multiple regressors (x_t) is considered in Zhang et al. (2009).

As with the standard IV method or the Control Function approach, the researcher still needs to provide theoretical reasoning for the presence of endogeneity in the models. There are two important caveats that one needs to consider when using the LIVE approach. First, it is important that the endogenous regressor (x_t) does not approximately follow a normal distribution, as the LIVE approach builds on this non-normality to separate the endogenous and exogenous parts of x_t. Second, the error term must be normally distributed, as this is an explicit assumption of LIVE. If these two rules are violated, it is not feasible to use the LIVE approach.

3.4. Copulas

Copulas have also emerged as a practical tool when no external instruments are available. Suppose we have the following model: $y_t = x_t\beta + u_t$, but we suspect that x_t and u_t are correlated. Thus, the idea is that when $\sim (x_t \perp u_t)$ but the joint distribution of x_t and u_t is known and is given by, say, $f(u_t, x_t; \alpha)$ where α are parameters associated with the joint distribution, then one can proceed using the method of maximum likelihood to maximize:

$$L(\beta, \alpha; \boldsymbol{Y}) = \prod_{t=1}^{n} f\,(y_t - x'_t\beta | x_t, \alpha) g(x_t; \alpha), \tag{27}$$

where the joint distribution factorizes identically as $f(u_t, x_t; \alpha) = f(u_t | x_t, \alpha) g(x_t; \alpha)$, and $g(x_t; \alpha)$ is the joint distribution of the regressors. The entire data set is denoted by $\boldsymbol{Y}$. Of course, the problem is that we do not know $f(u_t, x_t; \alpha)$ or $f(u_t | x_t, \alpha)$ or we do not wish to pretend to have such knowledge (for more technical details see Appendix B).

When we are not willing to specify the distribution of the regressors, we can use a copula approach. The copula idea results from Sklar's theorem. Given two random variables, say X_1 and X_2, Sklar's theorem expresses the joint distribution as a product of the marginals times a copula term that depends only on the marginal distribution functions:

$$f(x_1, x_2) = f_1(x_1) f_2(x_2) c\big(F_1(x_1), F_2(x_2)\big).$$

Under independence, the copula term is equal to unity, identically. It can be shown that, in simple linear regressions, the copula approach is equivalent to the following:

$$y_t = x_t\beta + x_t^*\gamma + e_t, \tag{28}$$

where e_t is an error term,

$$x_t^* = \Phi^{-1}(H(x_t)), \tag{29}$$

where $\Phi^{-1}(\cdot)$ is the standard normal inverse cumulative distribution function, and $H(x)$ is the empirical cumulative distribution function (CDF) of the regressor, which can be obtained directly from the data without parametric assumptions. Therefore, the copula approach (using a Gaussian copula) amounts to augmenting the data with the additional regressor x_t^*. In turn, one can use OLS to estimate the parameters β and γ.

The original theorem, attributed to Sklar (1952), says that

$$F(x_1, \dots, x_p) = C\big(F_1(x_1), \dots, F_p(x_p)\big), \tag{30}$$

where $F(x_1, \dots, x_N)$ is the joint CDF of p random variables, and $F_j(x_j)$ are the marginal CDF ($j = 1, \dots, p$). In turn, the joint CDF is the product of marginals (which would be the case under independence) multiplied by a link function, that is:

$$f(x) = \left\{\prod_{j=1}^{p} f_j\,(x_j)\right\} \cdot c(F_1(x_1), \dots, F_p(x_p)), \tag{31}$$

where $F_j(x_j)$ is the CDF corresponding to $f_j(x_j)$. Therefore

$$\frac{f(x)}{f_1(x_1) f_2(x_2) \cdots f_p(x_p)} = c(F_1(x_1), \dots, F_p(x_p)). \tag{32}$$

For example, the multivariate normal copula (Song, 2000) is:

$$C(u_1, \dots, u_p) = F\big(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p); R\big), \tag{33}$$

where $\Phi(\cdot)$ is the standard normal CDF, and F is the standard p-dimensional normal CDF with zero means, unit variances, and correlation matrix R. The density function is given by (Clemen and Reilly, 1999, Nelson, 1999):

$$\begin{aligned} f(x_1, \dots, x_p) = {} & f_1(x_1) \dots f_p(x_p) \\ & \cdot |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2} z'(\Sigma^{-1} - I)z\right\}, \end{aligned} \tag{34}$$

where $z_j = \Phi^{-1}(F_j(x_j))$, $j = 1, \dots, p$, and Σ represents a covariance or correlation matrix.

In the case of more than one endogenous variable, suppose we have:

$$y_t = x'_t\beta + Y_t'\gamma + u_t,\ t = 1, \dots, n, \tag{35}$$

where $Y_t \in \mathbb{R}^m$ is a vector of endogenous variables. The copula approach amounts to correcting this regression as follows:

$$y_t = x'_t\beta + Y_t'\gamma + \widetilde{Y'}_t\delta + \xi_t, \tag{36}$$

where ξ_t is an error term, $\tilde{Y}_{t1} = \Phi^{-1}(F_1(Y_{t1})), \dots, \tilde{Y}_{tm} = \Phi^{-1}(F_m(Y_{tm}))$, and $F_j(Y_{tj})$ represent the empirical CDF of $Y_{tj} (j = 1, \dots, m)$, which can be computed easily from the

data:

$$F_j(y) = n^{-1} \sum_{t=1}^{n} \mathbb{I}\,(Y_{tj} \leq y),\ j = 1, \ldots, m. \tag{37}$$

Here, $\mathbb{I}(\cdot)$ is the indicator function. What is remarkable is that despite the fact that Y_t is multivariate, only univariate CDF estimation is involved. Due to this, the method scales well with the sample size and the number of endogenous regressors[1].

This formulation is similar to Heckman (1978) and Hausman (1978). As x_t^* in (35) is a "generated regressor" (Pagan 1984) the usual standard errors of parameters are incorrect, but a standard bootstrap method to compute the correct standard errors can be used (typically, in MLE the information matrix is used to compute standard errors). In (35) the coefficient of x_t^* is actually $\gamma = \rho\sigma$ and the variance of error is $E(e_t^2) = \sigma^2(1 - \rho^2)$, so we can determine both σ and ρ (the correlation of the regressor x_t and the error u_t). Of course, the same is true for the more general formulation in (34) regarding $\tilde{Y}_t$.

4. Concluding Remarks and Recommendations

As endogeneity continues to remain a serious issue affecting validity of research in our field, the purpose of this paper was to foster a new discussion on the topic and provide more clear guidelines on how to deal with endogeneity. We discuss the various sources of endogeneity and present a variety of methods that can be used to mitigate the problem; distinguishing between those that rely on the use of instrumental variables and those that do not. While we acknowledge that there is no perfect solution for endogeneity, moving forward we recommend the following:

1- More focus on theory to carefully develop the research design and the selection of variables. With a more solid theoretical foundation, one can also identify potential sources of endogeneity and select the appropriate instrumental variables. Although this is easier said than done, standard behavioral assumptions, such as cost minimization or profit maximization, can, in fact, help to suggest what is endogenous and what might be predetermined, under certain conditions.

2- We echo Rutz and Watson (2019) that "when in doubt" collect more data to address any omitted variable problems before quickly thinking about methods to address the problem. The use of panel data with firm and time fixed effects can also help reduce the problem of omitted variable bias. This is so because slowly time-varying endogenous regressors will be absorbed by the firm effects and the problem of endogeneity can be mitigated. If the omitted endogenous regressors

[1] It is also quite important to mention that "while many copula functions have been identified, we believe only two are useful for building a regression model with several covariates. We are aware of only two copula models that allow for this, the normal copula and its generalization, the t-copula (which is based on the multivariate Student's *t* distribution. For example, see the list of copulas in Klugman, Panjer, and Willmot (2008, Chapter 7), where it can be seen that the other copulas do not allow for variations in the association measure" (Parsa and Klugman, 2011, pp. 48–49).

are now slowly varying, panel data do not help and, in fact, create more problems. As a matter of fact, issues of reverse causality can be addressed by using dynamic panel data (DPD) models (which do, however, require instrumental variables).

3- When in doubt, we also recommend estimating two models, one with control for endogeneity and one without to ensure the consistency of the findings.
4- Considering that the IV and Control Function approaches require the use of instrumental variables, we recommend increased use of instrument-free methods such as LIVE and Copulas. Finding instruments requires strong theoretical support and the selection of invalid or weak instrument can result in a strong bias, as we have shown in the above simulations. The use of LIVE and Copulas is not free of challenges. One, for instance, needs to ensure that the non-normality condition in the explanatory variable along with the normality condition in the error terms are met.
5- As there are many sources of endogeneity, we recommend a balanced approach that focuses on using the best method for the task at hand. Engaging with more complicated methods just for the sake of addressing endogeneity is not the best way to proceed. We recommend obtaining a thorough understanding of each of the above methods and weighing their use against the theory and endogeneity problem at hand.

Again, we emphasize that there is no optimal way to address endogeneity. One can also avoid non-experimental data all together and rely on field experiments instead, which "are often presented as the gold standard to create causal insights as they allow for manipulation of variables of interest in controlled settings" (Rutz and Watson, 2019, p. 490). However, if not executed correctly, field experiments can also be subject to endogeneity problems. To sum up, it is difficult to imagine any study, especially those that are not based on experimental data, without any suspicion of endogeneity problems. Our goal with this paper was to revisit the issue, present the various methods, and examine the advantages and disadvantages of each of them. Our hope is to improve the estimation of regression models and the validity of hypothesis testing in our field.

References

Adcock, R. J. (1878). A problem in least squares. *The Analyst*, *5*(1), 53-54.

Amemiya, T. (1973). Generalized Least Squares With an Estimated Autocovariance Matrix. *Econometrica*, 41, 723–732.

Amemiya, T. (1975). The Nonlinear Limited-Information Maximum-Likelihood Estimator and the Modified Nonlinear Two-Stage Least-Squares Estimator. *Journal of Econometrics*, 3, 375–386.

Amemiya, T. (1977). The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model. *Econometrica*, 45, 955–968.

Assaf, A. G., & Tsionas, M. G. (2019). Quantitative research in tourism and hospitality: an agenda for best-practice recommendations. *International Journal of Contemporary Hospitality Management*, in press.

Atkinson, S. E., & Tsionas, M. G. (2016). Directional distance functions: Optimal endogenous directions. *Journal of Econometrics*, *190*(2), 301-314.

Atkinson, S. E., Primont, D., & Tsionas, M. G. (2018). Statistical inference in efficient

production with bad inputs and outputs using latent prices and optimal directions. *Journal of Econometrics*, *204*(2), 131-146.
Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics*, *5*(3), 207-212.
Clemen, R. T., & Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, *45*(2), 208-224.
Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, T. (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics*, *3*(4), 365-392.
Ebbes, P., Papies, D., & Van Heerde, H. J. (2011). The sense and non-sense of holdout sample validation in the presence of endogeneity. *Marketing Science*, *30*(6), 1115-1122.
Ebbes, P., Papies, D., & van Heerde, H. J. (2016). Dealing with endogeneity: A nontechnical guide for marketing researchers. *Handbook of market research*.
Gourieroux, C., A. Monfort, E. Renault, and A. Trognon (1987). Generalised residuals. *Journal of Econometrics,* 34 (1–2), 5–32.
Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, 1251-1271.
Heckman, J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, *46*(4), 931-959.
Kleibergen, F., & Van Dijk, H. K. (1994). On the shape of the likelihood/posterior in cointegration models. *Econometric theory*, *10*(3-4), 514-551.
Kleibergen, F., & Van Dijk, H. K. (1998). Bayesian simultaneous equations analysis using reduced rank structures. *Econometric theory*, *14*(6), 701-743.
Koo, T. T., Lim, C., & Dobruszkes, F. (2017). Causality in direct air services and tourism demand. *Annals of Tourism Research*, *67*, 67-77.
Kuo, H. I., Chang, C. L., Huang, B. W., Chen, C. C., & McAleer, M. (2009). Estimating the impact of avian flu on international tourism demand using panel data. *Tourism Economics*, *15*(3), 501-511.
Liang, W., H. Dai, & S. He (2019). Mean Empirical Likelihood. *Computational Statistics & Data Analysis*, 138, 155–169.
Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 221-247.
Papies, D., Ebbes, P., & Van Heerde, H. J. (2017). Addressing endogeneity in marketing models. In *Advanced methods for modeling markets* (pp. 581-627). Springer, Cham.
Parsa, R. A., & Klugman, S. A. (2011). Copula regression. *Variance Advancing and Science of Risk*, *5*, 45-54.
Qin, J., & L. Lawless (1994). Empirical Likelihood and General Estimating Equations. *The Annals of Statistics* , 22 (1), 300–325.
Robinson, P. (1991). Best Nonlinear Three-Stage Least Squares Estimation of Certain Econometric Models. *Econometrica* , 59 (3), 755–786.
Rutz, O. J., & Watson, G. F. (2019). Endogeneity and marketing strategy research: an overview. *Journal of the Academy of Marketing Science*, *47*(3), 479-498.
Sklar, A. (1952). On the factorization of squarefree integers. *Proceedings of the American Mathematical Society*, *3*(5), 701-705.
Song, P. X. K., Li, M., & Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, *65*(1), 60-68.
Terza, J. V. (2018). Two-stage residual inclusion estimation in health services research and health economics. *Health services research*, *53*(3), 1890-1899.
Tran, K. C., & Tsionas, E. G. (2013). GMM estimation of stochastic frontier model with endogenous regressors. *Economics Letters*, *118*(1), 233-236.

Tsionas, M. G., & Mamatzakis, E. (2019). Further results on estimating inefficiency effects in stochastic frontier models. *European Journal of Operational Research*, *275*(3), 1157-1164.
van Dijk, H. (2002). *On Bayesian structural inference in a simultaneous equation model* (No. EI 2002-10).
Van Huffel, S., & Vandewalle, J. (1991). *The total least squares problem: computational aspects and analysis* (Vol. 9). Siam.
Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, *11*(3), 284-300.
Zhang, J., Wedel, M., & Pieters, R. (2009). Sales effects of attention to feature advertisements: a Bayesian mediation analysis. *Journal of Marketing Research*, *46*(5), 669-681.

Appendix A

Nonlinear models

With nonlinear models, endogeneity becomes a more difficult problem. A classical case is one equation out of a possibly larger model which has the form:

$$y_t = f(Y_t, x_t; \beta) + u_t, t = 1, \ldots, T, \tag{A.1}$$

where Y_t denotes the endogenous variables other then y_t, β is a parameter vector, and x_ts are independent of the error term u_t. In the limited information maximum likelihood (LIML) approach the system is completed with a reduced form for the remaining endogenous variables as follows:

$$Y_t = \Pi x_t + U_t, \tag{A.2}$$

and, typically, it is assumed that $[u_t, U'_t]' \sim^{iid} \mathcal{N}(0, \Sigma)$. Then a likelihood function can be formulated and maximized jointly with respect to β and parameters in Π. How one can define a nonlinear two-stage or three-stage least squares estimator in this context is part of an older but interesting literature (Amemiya, 1973, 1975, 1977, Robinson, 1991).
Let us consider a specific nonlinear model, for example Poisson regressor where the dependent variable y_t assumes integer values only (including zero). We assume a Poisson distribution:

$$y_t|\lambda_t \sim \mathcal{P}(\lambda_i), t = 1, \ldots, n, \tag{A.3}$$

whose probability mass function is, by definition:

$$p(y_t|\lambda_t) = e^{-\lambda_t} \frac{\lambda_t^{y_t}}{y_t!}, t = 1, \ldots, n. \tag{A.4}$$

A common assumption is that the mean (and variance) of the Poisson parameter depend on a vector of explanatory variables x_t with parameters β:

$$\lambda_t = e^{x'_t\beta} \Rightarrow \log\, \lambda_t = x'_t\beta, t = 1, \ldots, n. \tag{A.5}$$

The likelihood function of this Poisson regression model is:

$$L(\beta; y, X) \propto \prod_{t=1}^{n} e^{-e^{x'_t\beta}} \left(e^{x'_t\beta}\right)^{y_t}, \tag{A.6}$$

where y, X denote the data. Therefore, the log-likelihood function is:

$$l(\beta) \equiv \log\, L\,(\beta; y, X) = y_t(x'_t\beta) - e^{x'_t\beta}. \tag{A.7}$$

The first derivatives provide the score vector

$$\nabla l(\beta) = \sum_{t=1}^{n} x_t \left(y_t - e^{x'_t\beta}\right). \tag{A.8}$$

The Hessian matrix of the log-likelihood function is given by

$$\nabla^2 l(\beta) = -\sum_{t=1}^{n} x_t\, x_t' e^{x'_t\beta}. \tag{A.9}$$

Setting the score vector equal to zero, gives the nonlinear equations:

$$\sum_{t=1}^{n} x_t \left(y_t - e^{x'_t\beta}\right) = 0 \Rightarrow X'\hat{u}(\hat{\beta}) = \mathbf{0}, \tag{A.10}$$

where $\hat{\beta}$ is the ML estimator, and $\hat{u}(\beta) \equiv y - e^{X\beta}$ where the exponential is taken component-wise for a vector. The system of equations can be solved in a standard way and asymptotic standard errors may be computed from the inverse negative Hessian, viz. $\mathrm{cov}(\hat{\beta}) = -\left[\nabla^2 l(\hat{\beta})\right]^{-1} = \left(\sum_{t=1}^{n} x_t\, x_t' e^{x'_t\hat{\beta}}\right)^{-1}$. As the log-likelihood is globally concave, the maximum is unique. From (10) it is clear that for the ML to be consistent it must be the case that x_t is orthogonal to the vector of generalized residuals $\hat{u}(\beta)$ (Gourieroux et al., 1987). If this is not the case but there is a vector z_t that satisfied this condition, then we can use the following estimating equations:

$$\sum_{t=1}^{n} z_t \left(y_t - e^{x'_t\beta}\right) = \mathbf{0}, \tag{A.11}$$

provided the dimensionality of z_t weakly exceeds the number of parameters in β.

Another prominent example is the probit model. Suppose we have n observations the first n_o of which have $y_t = 1$ and the remaining have $y_t = 0$. The likelihood function is:

$$L(\beta; y, X) = \prod_{t=1}^{n_o} \Phi\,(x'_t\beta) \prod_{t=n_o}^{n} \Phi\,(-x'_t\beta), \tag{A12}$$

where $x_t \in \mathbb{R}^k$ are the explanatory variables, and $\beta \in \mathbb{R}^k$ is the parameter vector. The score vector is

$$\nabla l(\beta) = \sum_{t=1}^{n_o} x_t\, \Lambda(x'_t\beta) - \sum_{t=n_o+1}^{n} x_t\, \Lambda(x'_t\beta), \tag{A.13}$$

where $\Lambda(q) = \frac{\varphi(q)}{\Phi(q)}, q \in \mathbb{R}$. These equations can be written as

$$\sum_{t=1}^{n} x_t\, \mathbb{I}_t \Lambda_t(\beta) = \mathbf{0}, \tag{A.14}$$

where

$$\mathbb{I}_t = \begin{cases} 1, \text{if} y_t = 1 \\ -1, \text{otherwise.} \end{cases} \tag{A.15}$$

In effect, the generalized residuals of a probit model can be defined as $u_t(\beta) = I_t \Lambda_t(\beta)$ and the requirement for consistency of ML estimation is that the x_ts are orthogonal to $u_t(\beta)$. If this requirement is not satisfied but there is a vector z_t and $dim(z_t) \geq k$ then we can define an alternative estimator based on the following estimating equations:

$$\sum_{t=1}^{n} z_t \, \mathbb{I}_t \Lambda_t(\beta) = \mathbf{0}. \tag{A.16}$$

These so-called "likelihood instrumental variable" estimators extend the scope of instrumental variables in nonlinear and / or likelihood-based models. Despite the use of generalized residuals in Gourieroux et al. (1987) the "likelihood instrumental variable" estimator proposed here appears novel. In fact, based on (A.11) or (A.14) it is possible to remove the distributional assumptions and use empirical likelihood methods, see Qin and Lawless (1994), and Liang et al. (2019) for a recent reformulation.

Appendix B

Generalized method of moments (GMM)

Another popular estimator of the IV method is GMM. In the case of linear models, it reduces to (15) and there are many ways to obtain Ω so that estimates are "robust" to heteroskedasticity or autocorrelation of unknown form. To explain the method, suppose we have the linear model in (6) and an $M \times 1$ vector of instruments z_t so that, for all $t = 1, \dots, n$ it is true that $z_t \perp u_t$. Since $u_t = y_t - x'_t\beta$ this condition can be given a sample interpretation as follows:

$$n^{-1} \sum_{t=1}^{n} (y_t - x'_t\beta) \otimes z_t = \mathbf{0}_M, \tag{A.17}$$

where $\otimes$ is a Kronecker product, and $\mathbf{0}_M$ is an $M \times 1$ vector of zeros. The Kronecker product may seem strange, but it yields a rich class of models, possibly nonlinear, in a system context. Suppose $y_t = [x_t', z'_t]'$ and the system has the form:

$$\begin{aligned} f_1(y_t; \beta) &= u_{t1}, \\ f_2(y_t; \beta) &= u_{t2}, \\ &\dots \\ f_N(y_t; \beta) &= u_{tN}, \end{aligned} \tag{A.18}$$

where $f = [f_1, \dots, f_N]'$ is a vector of functions, and $u_t = [u_{t1}, \dots, u_{tN}]'$ is a vector of error terms. The notation $f_1(x_t, z_t; \beta) = u_{t1}$ means that different x_ts and z_ts may enter into different equations. To clarify this notation, the following example may be helpful:

$$\begin{aligned} x_{t1} &= \beta_1 x_{t2} + \beta_2 z_{t1} + \beta_3 z_{t2} + u_{t1} \\ x_{t2} &= \beta_4 x_{t1} + \beta_5 z_{t3} + \beta_6 z_{t4} + u_{t2}. \end{aligned} \tag{A.3}$$

This system can be put into the form (A.2). There are four predetermined variables (z_{t1}

through z_{t4}) and it is possible that there are linear or nonlinear constraints among the parameters, for example $\beta_6 = 1 - \beta_1 - \beta_2$ or $\beta_5 = \beta_4^2 + \frac{1}{1-\beta_3}$.

If we define $\boldsymbol{f}(y_t;\beta) = [f_1(y_t;\beta), \dots, f_N(y_t;\beta)]'$ then the moment conditions can be written in the form:

$$n^{-1}\sum_{t=1}^{n} \boldsymbol{f}\,(y_t;\beta) \otimes z_t = \mathbf{0}_{NM}. \tag{A.19}$$

We have NM moment conditions in total to estimate the $K \times 1$ vector of parameters β. Alternatively, we may write the moment conditions as follows:

$$\boldsymbol{G}(\beta) \equiv n^{-1}\sum_{t=1}^{n} \boldsymbol{g}\,(y_t;\beta) = \mathbf{0}_{NM}, \tag{A.20}$$

where $\boldsymbol{g}(y_t;\beta) \equiv \boldsymbol{f}(y_t;\beta) \otimes z_t$, where $\boldsymbol{G}(\beta)$ contains NM elements. Typically, $NM > K$ so there are more equations than unknown parameters; the extra equations being known as over-identifying restrictions. As in Least Squares (LS), one proceeds to minimize the criterion:

$$S(\beta) = \boldsymbol{G}(\beta)'W\boldsymbol{G}(\beta), \tag{A.21}$$

for some weighting matrix W. The optimal weighting matrix is $W = E[\boldsymbol{g}(y_t;\beta)\boldsymbol{g}(y_t;\beta)']$ and can be estimated using $\widehat{W} = n^{-1}\sum_{t=1}^{n} \boldsymbol{g}\,(y_t;\beta)\boldsymbol{g}(y_t;\beta)'$. One can implement the GMM estimator in two ways. The first way consists of two stages. In the first stage one sets $W = I$ to obtain a consistent estimator. In the second stage, given the estimates, $\widehat{W}$ is computed and the optimization is repeated. The second version of GMM is known as Continuously Updated Estimator (CUE) and, basically, directly incorporates $\widehat{W}$ into the problem to minimize:

$$S(\beta) = \left[n^{-1}\sum_{t=1}^{n} \boldsymbol{g}\,(y_t;\beta)\right]' \left[n^{-1}\sum_{t=1}^{n} \boldsymbol{g}\,(y_t;\beta)\boldsymbol{g}(y_t;\beta)'\right]^{-1} \left[n^{-1}\sum_{t=1}^{n} \boldsymbol{g}\,(y_t;\beta)\right] \tag{A.22}$$