
Munich Personal RePEc Archive

Synthetic data: an endogeneity simulation

Carbajal De Nova, Carolina

Department of Economics, Autonomous Metropolitan University

13 February 2014

Online at https://mpra.ub.uni-muenchen.de/79067/
MPRA Paper No. 79067, posted 12 May 2017 05:59 UTC

SYNTHETIC DATA: AN ENDOGENEITY SIMULATION

Working paper, February 13, 2014

Carolina Carbajal De Nova[1]

Department of Economics, Autonomous Metropolitan University

enova@xanum.uam.mx

Abstract

This paper uses synthetic data and different scenarios to test treatments for endogeneity problems under different parameter settings. The model uses initial conditions and provides the solution for a hypothetical equation system with an embedded endogeneity problem. The behavioral and statistical assumptions are underlined as they are used through this research. A methodology is proposed for constructing and computing simulation scenarios. The econometric modeling of the scenarios is developed accordingly with the feedback obtained from previous scenarios. The inputs for these scenarios are synthetic data, which are constructed using random number machines and/or Monte Carlo simulations. The outputs of the scenarios are the model estimators. The research results demonstrated that a treatment for endogeneity can be developed as the sample size increases.

Keywords: synthetic data, endogeneity problems, scenarios, Monte Carlo simulations.

[1] Economics Professor (on leave). Av. San Rafael Atlixco, No. 186, Col. Vicentina, Del. Iztapalapa, Mexico City, Mexico, tel. 001 52 55 5804 4768, fax 001 52 55 5894 4769.

Introduction

This study develops a statistical application for providing an adequate treatment for endogeneity problems. The application is a hypothetical equation system that has an embedded endogeneity problem. This specification is a general form for the well-known endogeneity problem in economics. To characterize the evolution of the system of equations under different sample sizes and simulation scenarios, it is first necessary to create synthetic data as the equation system inputs. The synthetic data are constructed using Monte Carlo simulations based on the use of random number machines. In this respect, the author takes computational advantage of the already random machine modules installed in Matlab.

The main idea behind synthetic data is the creation of customized data and to feed a particular system of equations with them. The variation of key parameter values in the synthetic data and the variation of the sample size allow the implementation of a methodology to treat the endogeneity problem. Consequently, feasible tests are generated for demonstrating how the endogeneity problem decreases substantially under controlled conditions and without restrictions on data access. Thus, the proposed method adjusts the econometric modeling accordingly with the feedback obtained from previous scenarios. These different scenarios change when synthetic data and the system of equation parameter values take on different initial values. These changes respond to changes in assumptions. These scenarios represent behavioral experiments, which could have been very expensive had they been performed in real life. However, taking advantage of software developments and the increase of computational capacity in recent years, this research provides low-cost alternatives for performing these economic experiments alongside their corresponding econometric tests and analyses.

This paper is organized as follows. The first section presents the model based on a system of equations, with an embedded endogeneity problem. The second section briefly describes the construction of synthetic data. The third section implements

the simulation scenarios and reports the corresponding results. The last section concludes.

1. Model

Consider the following equation in which income is a function of different demographics:

$$Income = \alpha_0 + \alpha_1 MS + \alpha_2 Age + \alpha_3 Ability + \epsilon_I$$

where $Income$ stands for the returns of gaining a Master of Science degree; MS stands for a Master of Science; Age stands for age; and $Ability$ stands for an unobservable variable that is linked with performance. The estimators are as follows: α_0 stands for the constant; α_1 stands for the MS estimator; α_2 stands for Age estimator; α_3 stands for $Ability$ estimator; and ϵ_I stands for the error term associated with this equation. This last term is assumed to be normal, independent and identically distributed, i.e., $\epsilon_I \sim N(0, \sigma_I^2)$.

For the moment, suppose that you cannot observe the variable $Ability$. Hence, you must estimate the following equation given data availability:

$$Income = \alpha_0 + \alpha_1 MS + \alpha_2 Age + \tilde{\epsilon}_I$$

where $\tilde{\epsilon}_I$ stands for the error term associated with the $Income$ variable, which contains the unobservable $Ability$ variable.

Additionally, assume that MS has a Bernoulli distribution:

$$\Pr(MS) = [\phi(\gamma_0 + \gamma_1 Ability)]^{MS} + [1 - \phi(\gamma_0 + \gamma_1 Ability)]^{1-MS}$$

where $\Pr(MS)$ is the probability of observing the variable MS; $Ability$ is the same variable that appeared in the first equation; and ϕ stands for the standard normal cumulative distribution function (cdf).

Furthermore, assume that the next three assumptions hold:

Statistical assumptions

a) In deriving the method of moments estimator, the covariance expected value between regressors and the error term is zero, i.e., $E(X'\epsilon) = 0$;
b) The variable MS is uncorrelated with the error term of equation (1). Then its correlation is $\rho_{MS,\,\epsilon_I} = 0$.

Behavioral assumption

c) Negative quantities and prices only exhibit positive values.

Placing numbers to each of the equations above, the following system can be written:

$$(1) \qquad Income = \alpha_0 + \alpha_1 MS + \alpha_2 Age + \alpha_3 Ability + \epsilon_I;$$

$$(2) \qquad Income = \alpha_0 + \alpha_1 MS + \alpha_2 Age + \tilde{\epsilon}_I;$$

$$(3) \qquad \Pr(MS) = [\phi(\gamma_0 + \gamma_1 Ability)]^{MS} + [1 - \phi(\gamma_0 + \gamma_1 Ability)]^{1-MS}.$$

A variable is said to be endogenous when there is a correlation between it and the error term.[2] In the equation system above, the source of endogeneity seems to be related to the omitted variable. This omitted variable could be $Ability$ if it is not observable. In fact, this should be true if the correct model is the one in equation (1), and the one that is feasible, given data availability, is the model depicted by equation

[2] The use of a system of equations with a priory determination of exogenous and endogenous variables is not rare in the economics profession, as expressed in Hart, Mills and Whitaker (1964).

(2). Since equation (3) establishes a relationship between $Ability$ and MS, then the endogeneity problem could be revealed by means of the equation system algebraic manipulation.

The first step in demonstrating the existence of the endogeneity problem in the above equation system is to show that equation (2) contains the omitted variable $Ability$. The second step consists of demonstrating that MS is correlated with the error term of equation (2): $\rho_{MS,\tilde{\epsilon}_I} \neq 0$.

Consider that this hypothetical system has embedded an endogeneity problem. This problem is evidenced on equation (3). If equation (3) holds, then the correlation between MS and $Ability$ cannot be zero. Equation (3) links $Ability$ and MS through a non-linear relationship. For a moment, assume that $Ability$ is observable and that equation (1) can be estimated. This means that the estimator α_3 can be computed. In this case, the specification of a correct statistical model to commensurate income determinants is equation (1).

Now, consider that the variable $Ability$ cannot be observed. Equation (2) is then the only feasible model for estimation purposes. This implies that equation (2) replaces the econometric model of equation (1). However, if the correct model is the one set on equation (1), it implies that a measurement error associated with the omitted variable $Ability$ exists in equation (2). This last idea can be represented as follows:

$$\tilde{\epsilon}_I = \alpha_3 Ability + \epsilon_I. \quad (4)$$

Equation (4) shows that the omitted variable $Ability$ is absorbed by the error term of equation (2). Since the estimator α_3 is the one associated with $Ability$, it appears in equation (4). Some partial weight of ϵ_I should also be present on $\tilde{\epsilon}_I$. For easiness, consider that ϵ_I is so small, that some partial weight of it and itself have almost the same magnitude. In this case and for the sake of simplicity, equation (4) contains ϵ_I.

The second step consists of showing that the correlation between MS and $\tilde{\epsilon}_I$ is not zero if equation (3) holds. Consider the following cross moment or conditional mean of MS given $\tilde{\epsilon}$:

$$E[MS|\tilde{\epsilon}] = \beta + (X'X)^{-1}X'Z\delta$$

Additionally, consider that the Gauss-Markov theorem holds for this case:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

where Y stands for the variable *Income* and X represents a set of independent variables. Hence, the last equation can be rewritten as follows:

(5) $$\hat{\beta} = (X'X)^{-1}X'(\alpha_1 MS + \alpha_2 Age + \alpha_3 Ability + \epsilon_I).$$

where Y is written in terms of equation (1). Suppose that $\alpha_1 MS + \alpha_2 Age = X\beta$; $\alpha_3 Ability = \delta Z$; Z stands for the instrumental variable (IV); δ stands for the estimator of the IV; and $\epsilon_I = U$; U represents a generalized error term. Making substitutions and distributing terms in this last expression, it can be written as follows:

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + Z\delta + U)$$

$$\hat{\beta} = (X'X)^{-1}X'(X\beta) + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'U$$

$$\hat{\beta} = \beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'U$$

Suppose that $U = 0$, and considering expectations, the last expression simplifies to:

$$E\left[\hat{\beta}\middle|X\right] = \beta + (X'X)^{-1}X'Z\delta$$

$$E\left[\hat{\beta}\middle|X\right] = \beta + bias$$

The generalized estimator $\hat{\beta}$ is biased if $X'Z \neq 0$ or $\delta \neq 0$.[3] In this case, the bias is represented by the term $(X'X)^{-1}X'Z\delta$. This well-known result is the omitted variable formula.

In the last equation, $bias$ is proportional to a weighted portion of $Ability$, which depends on a non-linear association with MS. This association is set on equation (3). Please note that $Bias$ does not depend on Age, since this hypothetical equation system does not provide an equation where Age and $Ability$ are related.[4]

Another way to express $bias$ is as follows:

(6) $E\left[\hat{\beta}\middle|X\right] = \beta + \frac{cov[MS,Ability]}{Var[MS]}\alpha_3$.

where α_3 is the $Ability$ estimator of equation (1).

From equation (4), it is known that $\tilde{\epsilon}_I = \alpha_3 Ability + \epsilon_I$. Thus, $Ability$ can be expressed as follows:

$$Ability = \frac{\tilde{\epsilon}_I - \epsilon_I}{\alpha_3}$$

Plugging the last expression for $Ability$ into equation (6) results in the following:

(6)' $E\left[\hat{\beta}\middle|X\right] = \beta + \frac{cov\left[MS,\frac{\tilde{\epsilon}_I - \epsilon_I}{\alpha_3}\right]}{Var[MS]}\alpha_3$.

[3] For an example containing endogeneity bias, see Hayashi (2000, p. 187).
[4] According to Greene (2012), bias is caused by the omission of relevant variables.

If $\hat{\beta}$ is $biased$ as explained previously, then it should follow that $\frac{cov\left[MS,\frac{\tilde{\epsilon}_I-\epsilon_I}{\alpha_3}\right]}{Var[MS]}\alpha_3 \neq 0$ and thus $cov\left[MS,\frac{\tilde{\epsilon}_I-\epsilon_I}{\alpha_3}\right] \neq 0$.

The correlation coefficient for the MS and $Ability$ variables is as follows:

$$\rho_{MS,Ability} = corr(MS,\ Ability) = \frac{cov(MS,Ability)}{\sigma_{MS}\,\sigma_{Ability}}, \text{ or}$$

$$\rho_{MS,\frac{\tilde{\epsilon}_I-\epsilon_I}{\alpha_3}} = corr\left(MS,\frac{\tilde{\epsilon}_I-\epsilon_I}{\alpha_3}\right) = \frac{cov(MS,\frac{\tilde{\epsilon}_I-\epsilon_I}{\alpha_3})}{\sigma_{MS}\,\sigma_{\frac{\tilde{\epsilon}_I-\epsilon_I}{\alpha_3}}}$$

since $cov\left[MS,\frac{\tilde{\epsilon}_I-\epsilon_I}{\alpha_3}\right] \neq 0$, and $\epsilon_I \sim N(0,\sigma_I^2)$ it follows that $\rho_{MS,\frac{\tilde{\epsilon}_I-\epsilon_I}{\alpha_3}} \neq 0$. Thus, there is indeed a correlation between MS and $\tilde{\epsilon}_I$. With these two steps, the endogeneity problem embedded in the hypothetical equation system has been made evident, since it matched the definition of an endogenous variable: an endogenous variable exists when there is a correlation between it and the error term. This applies specifically to MS in equation (2).

Once the endogeneity problem has been made evident in the equation system, what follows is the specification for synthetic data construction and simulation scenarios. These procedures are presented in the next section. The construction of synthetic data can be done in different ways. One is by means of Monte Carlo simulations, which are widely used. It uses a recursive algorithm and a fix population. The seed fixes the population to a given size. Therefore, its seed limits Monte Carlo sample sizes. The use of random machines, which are modules preinstalled on Matlab, allows samples to be drawn from an infinite population. These samples have asymptotic properties that are absent in samples derived from Monte Carlo

simulations.[5] The author opts for this last method, since it results in samples with asymptotic properties. In the next section, the synthetic data and construction of scenarios are explained.

2. Synthetic data construction and simulation scenarios

Consider the next four stages for describing synthetic data construction and scenario estimations:

1. Take initial parameter values for the variables that composed equations (1); (2) and (3): $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \gamma_0, \gamma_1, \sigma_I^2$;

2. With the help of the random number generators, which are modules preinstalled on Matlab, generate 20 observations for each variable in this research equation system;

3. Proceed to estimate the income equation with different sample sizes, i.e., 50 and 100 observations. Compare these equation estimators in terms of percentage deviations from initial parameter values. Register the results properly in the corresponding tables;

4. Changing initial parameters values reproduces alternative scenarios. Let steps 2 and 3 adjust automatically to these new values and take note.

It is important to underline that the variables created in steps one and two represent synthetic data. They are generated with random number machines, and every time the run button is hit in the Matlab software, the algorithm is updated. This allows omitting the step of taking draws from the determined size of one

[5] "Since the variables were generated with random number [generators], every time the code is run, the variables get updated. This allows the analyst to omit an additional step of taking draws from the population in the Monte Carlo simulation. So, the population from which the samples are being taken is infinite, instead of being bounded to 100 or 1,000 population size." Carbajal (2013, p. 4 *bracket added*).

population, as when the Monte Carlo method is used. This is the case because this procedure restricts the draws that could be taken given a certain population size. Regardless of whether Monte Carlo or the random machines method is used, both allow for the creation of synthetic data.

Equation (2) was solved using MS synthetic data. This is important because it allows it to have a determined distribution and asymptotic properties. That is, the population moments are shared by the synthetic data samples. The variable MS is key to estimating the system of equations and to showing an endogeneity problem treatment.

For instance, assume that $Ability$ is not available. Thus, equation (2) nests the variable $Ability$ in its error term as described in equation (4). Furthermore, assume that $Ability$ is available and denote it with the name $Abilitybis$ to distinguish it from the case when it is not observable. The Pr(MS) on equation (3) can then be computed given that the variable $Abilitybis$ is used as an input. Using these two types of variables, $Ability$ and $Abilitybis$, the differences of having an endogenous problem without and with treatment will become obvious in the next section. The treatment implements the use of IV. Instrumental variables are a common method in economics to correct endogeneity problems.

3. Scenarios and results

3.a. First scenario

n=20 observations case

Follow steps 1-3 from the previous section. Set the number of repetitions to 20, for each of the variables on equations one to three. Initial parameter values are assigned as follows:

$\alpha_0 = 10;$
$\alpha_1 = 20;$

α_2 =30.

After running a regression for equation (2) using the method of ordinary least squares, the following estimators are obtained:

$\hat{\alpha}_0$ =41.7398;
$\hat{\alpha}_1$ =19.9312;
$\hat{\alpha}_2$ =30.1700.

The estimator for α_2 is closed to its assigned value. The difference between them is (30.17-30)/30*100=0.56% (overestimation). Something similar occurs with α_1, where the difference between the true value and its estimator is close to (19.9312-20)/20*100=-0.344% (underestimation). Regarding the coefficient estimate for α_0, it is very far from its true value.

The following 95% confidence intervals are obtained after running the corresponding regressions:

For $\hat{\alpha}_0$ =41.7398, [13.3659 70.1137], with a true value of α_0 =10;
for $\hat{\alpha}_1$ =19.9312, [19.4721 20.3904], with a true value of α_1 =20;
for $\hat{\alpha}_2$ =30.1700, [29.4188 30.9213], with a true value of α_2 =30.

As we can see from the information given above, the confidence interval contained the truth-value. One exception is represented by α_0. For instance, $\alpha_1 = 20$ is contained in the confidence interval [19.4721 20.3904]. Likewise, $\alpha_2 = 30$ is contained in [29.4188 30.9213].

The coefficient estimate for MS ($\hat{\alpha}_1$) is close to the lower bound of its 95% confidence interval. The coefficient estimate for Age ($\hat{\alpha}_2$) is near to the upper bound of its 95% confidence interval. The coefficient for the constant ($\hat{\alpha}_0$) represents

merely the average of the observations, so it is around the middle of its 95% confidence interval.

n= 50 observations case, results

$\hat{\alpha}_0$ =75.8929;
$\hat{\alpha}_1$ =19.8502;
$\hat{\alpha}_2$ =30.0290.

Remember that the following values for the above parameters are assigned as follows:

α_0 =10;
α_1 =20;
α_2 =30.

The estimate for α_2 is closed to its assigned value. The difference between them is (30.0290-30)/30*100=0.09% (overestimation). Something similar occurs with α_1, where the difference between the estimate and its true value is close to (19.8502-20)/20*100=-0.749% (underestimation). Regarding the coefficient estimate for α_0, it is very far from its true value.

n= 100 observations case, results

$\hat{\alpha}_0$ = 129.2581;
$\hat{\alpha}_1$ = 19.9555;
$\hat{\alpha}_2$ = 29.9135.

Remember that the following values for the above parameters are assigned as follows:

$\alpha_0 = 10$;

$\alpha_1 = 20$;

$\alpha_2 = 30$.

The estimate for α_2 is close to its assigned value. The difference between them is (29.9135-30)/30*100=-0.288% (underestimation). Something similar occurs with α_1, where the difference between the estimate and its true value is close to (19.9555-20)/20*100=-0.222% (underestimation). Regarding the coefficient estimate for α_0, it is very far from its true value.

The differences of the estimates with 100 observations and the previous estimates with 20 and 50 observations, are compared in Table 1. This comparison aims to verify whether there is an improvement in the estimator convergence with its initial value, given an increase in the sample size.[6]

Table 1. Estimator convergence. Sample sizes 20, 50 and 100.

Estimator	Differences (%) 20 observations	Differences (%) 50 observations	Differences (%) 100 observations	Improvement?
$\hat{\alpha}_1$	-0.344	-0.749	-0.222	Yes
$\hat{\alpha}_2$	0.560	0.090	-0.288	Yes

From Table 1, the differences measured in percentage terms have decreased as the number of observations increased in the sample. For instance, the difference between the estimator and its true value for α_1 is -0.344% with 20 observations, and it decreases to -0.749% with 50 observations; thus, this percentage decreases. It decreases even more with 100 observations, i.e., -0.222% (in absolute terms); thus, the percentage decreases. A similar improvement is seen in α_2 because it passes

[6] According to Neese and Hollinger (1985), the process of varying the number of observations affects the confidence intervals' sensitivities.

from 0.56% with 20 observations to 0.09% with 50 observations and -0.288 with 100 observations; thus, the percentage decreases.

From Table 1, it can be confirmed empirically that as the number of observations increases, the coefficient estimates become closer and closer to their true values. This is simply a confirmation of the Law of Large Numbers, the Central Limit Theorem and the Asymptotic Distribution of the OLS estimators.[7] The intuition is that as the number of observations increases to infinity, the estimators will converge with their initial value. Remember that this research equation system has embedded an endogeneity problem. Convergence is possibly achieved, because the increase in sample size diminishes the estimator bias, and the aforementioned asymptotic properties hold. If this intuition is true, then this research results are aligned with those of Alvarez and Arellano (1988) and Anderson and Hsiao (1981).[8]

3.b. Second scenario

In the scenario above, the variable $Ability$ was unobservable. This assumption has now changed and $Ability$ is observed. This observable variable is called $Abilitybis$. Equation (3) can be rewritten in terms of this observable variable.

(3)' $$\Pr(MS) = [\phi(\gamma_0 + \gamma_1 Abilitybis)]^{MS} + [1 - \phi(\gamma_0 + \gamma_1 Abilitybis)]^{1-MS}.$$

The initial parameter values for this scenario are:

α_0=10;

[7] For a clear demonstration of the Central Limit Theorem, see Hogg, McKean and Craig (2013, p. 307). More information is available on Hamilton (1994) regarding the Law of Large Numbers and Asymptotic Distributions.

[8] When Alvarez and Arellano (1988) derive the asymptotic properties of different kind of estimators (GMM, LIML and WG) when N tends to infinity, they find that these estimators are asymptotically equivalent to the WG estimators. In this study, this could imply that if the initial parameter values are set equal to the WG estimators, then other types of estimators could reach asymptotic convergence to WG estimators as the sample size increases and bias diminishes. In the case of Anderson and Hsiao (1981), when N tends to infinity, the MLE delivers consistent estimators depending on the initial conditions. The author understands as consistent estimators those, which are efficient and unbiased.

α_1=50;

α_2=30;

γ_0=40;

γ_1=50;

σ_I^2=1.

where σ_I^2 represents the variance of the error term; γ_0 and γ_1 represent the estimators of equation (3).

In this scenario, equation (2) includes equation (3)'. This implies that MS has been computed in a previous stage and is later entered as input in equation (2).

n= 20 observations case

The estimators for equation (2) are:

$\hat{\alpha}_0 = 126.2643$;

$\hat{\alpha}_1 = 20.2350$;

$\hat{\alpha}_2 = 30.0320$.

The associated 95% confidence intervals are:

For $\hat{\alpha}_0 = 126.2643$, [117.7658 134.7627], with a true value of α_0 =10;
for $\hat{\alpha}_1$ =20.2350, [20.0225 20.4475], with a true value of α_1 =50;
for $\hat{\alpha}_2$ =30.0320, [29.7946 30.2694], with a true value of α_2 =30.

From the information above, the confidence intervals do not contain the true initial value for α_0 and α_1. That is, α_0 =10 is not inside the range of the following interval [117.7658 134.7627]; α_1 =50 is not contained in [20.0225 20.4475]. Reversing this behavior trend, we see that α_2 =30 is contained in [29.4188 30.9213]. This behavior is likely observed, because now equation (2) nests equation (3)'.

n= 50 observations case

Once the sample size is increased from 20 to 50 observations, the following estimators are computed:

$\hat{\alpha}_0 = 128.6414;$
$\hat{\alpha}_1 = 20.0608;$
$\hat{\alpha}_2 = 29.9382.$

n= 100 observations case

$\hat{\alpha}_0 = 23.1880;$
$\hat{\alpha}_1 = 49.9987;$
$\hat{\alpha}_2 = 30.1105.$

Next, in Table 2, the information for this scenario is summarized:

Table 2. Estimator convergence. Sample sizes 20, 50 and 100.

Estimator	Differences (%) 20 observations	Differences (%) 50 observations	Differences (%) 100 observations	Improvement?
$\hat{\alpha}_1$	-59.530	-59.878	-0.002	Yes
$\hat{\alpha}_2$	0.106	-0.002	0.368	No

From Table 2, it can be observed that the differences measured in percentage terms have decreased as the number of observations increased from 20 to 50. For instance, the difference between the estimator α_1 and its true value is -59.530% with 20 observations, and it decreases to -59.878% with 50 observations; it decreases even more with 100 observations, i.e., -0.002 (in absolute terms). The

difference for the estimator α_2 and its true value decreases when the sample size increases from 20 to 50 observations, with values of 0.106 and -0.002, when the number of observations increases from 50 to 100, there is no improvement in its difference.

In general, Table 2 shows empirically that as the number of observations increases, α_1 estimators become closer and closer to its true value. This is simply a confirmation of the Law of Large Numbers, the Central Limit Theorem and the Asymptotic Distribution of the OLS estimators, as explained previously. In the case of α_2, convergency is not achieved, since the endogeneity problem is now included in the estimation of equation (2).

3.c. Third scenario

Proceed to find a new method of moments for equation (2). This scenario includes the use of IV to correct the endogeneity problem of the second scenario. The variable Z is referred as the instrument or IV.

To determine whether $Abilitybis$ could be used as IV, it is necessary to compute the correlation coefficient between $Abilitybis$ and $\tilde{\epsilon}_I$. Note that in the model section, it was found that $\rho_{MS,\frac{\tilde{\epsilon}_I-\epsilon_I}{\alpha_3}} \neq 0$. In this scenario, the correlation needed is $\rho_{Abilitybis,\tilde{\epsilon}_I}$. It can be computed with the aid of synthetic data: $\rho_{Abilitybis,\tilde{\epsilon}_I} = -0.0050$. This value is close to zero. It is decided that this empirical correlation coefficient value is almost zero and that $Abilitybis$ and $\tilde{\epsilon}_I$ are not correlated.

The theoretical justification for the use of instrumental variables can be found in Greene (2012). Remember that $Abilitybis$ is observable to the econometrician.

The IV assumptions are as follows:

- Exogeneity. They are uncorrelated with the error term $E(Z'\tilde{\epsilon}_I) = 0$, where Z' indicates IV transpose;
- Relevance. They are correlated with the independent variables.

The asymptotic properties of the IV estimator indicate in general, that if Z has a finite variance, then the remaining regression equation components also have a finite variance:

$$plim\left(\frac{Z'\tilde{\epsilon}_I}{n}\right) = plim\left(\frac{Z'Y}{n}\right) - plim\left(\frac{Z'X\beta}{n}\right) = 0$$

Rearranging,

$$plim\left(\frac{Z'Y}{n}\right) = \left[plim\left(\frac{Z'X}{n}\right)\right]\beta + plim\left(\frac{Z'\tilde{\epsilon}_I}{n}\right)$$

If $plim\left(\frac{Z'\tilde{\epsilon}_I}{n}\right) = 0$, then

$$plim\left(\frac{Z'Y}{n}\right) = \left[plim\left(\frac{Z'X}{n}\right)\right]\beta + 0$$

$$plim\left(\frac{Z'Y}{n}\right) = \left[plim\left(\frac{Z'X}{n}\right)\right]\beta$$

If Z has the same number of variables as X, then the rank of $Z'X$ is k. This implies that $Z'X$ is a square matrix and its inverse exists:

$$\beta = \left[plim\left(\frac{Z'X}{n}\right)\right]^{-1} plim\left(\frac{Z'Y}{n}\right)$$

This last expression in closed-form leads to the IV estimator:

$$b_{IV} = (Z'X)^{-1}Z'Y$$

Consider $E(Z\tilde{\epsilon}_I) = 0$ moment condition. When this moment condition holds, it replaces the standard orthogonal condition and the derivation of a new general method of moment estimator can be found, for the income equation:

(3)' $\Pr(MS) = [\phi(\gamma_0 + \gamma_1 Abilitybis)]^{MS} + [1 - \phi(\gamma_0 + \gamma_1 Abilitybis)]^{1-MS}$.

Taking the log in both sides of equation (3)' produces the following expression:[9]

$$ln\Pr(MS) = ln[\phi(\gamma_0 + \gamma_1 Abilitybis)]^{MS} + ln[1 - \phi(\gamma_0 + \gamma_1 Abilitybis)]^{1-MS}$$

$$ln\Pr(MS) = \ln\{[\phi(\gamma_0 + \gamma_1 Abilitybis)]^{MS}[1 - \phi(\gamma_0 + \gamma_1 Abilitybis)]^{1-MS}\}$$

Applying the exponential to the last line yields the following:

(7) $\Pr(MS) = [\phi(\gamma_0 + \gamma_1 Abilitybis)]^{MS}[1 - \phi(\gamma_0 + \gamma_1 Abilitybis)]^{1-MS}$.

This last equation has a similar form as the Bernoulli probability density function (pdf):[10]

(8) $p(x) = p^x[1-p]^{1-x}, x = 0,1.$

When x takes on values of 0 and 1, an indicator function is produced. The population moments for a Bernoulli pdf are as follows:

$$\mu = p$$

[9] With this operation the corresponding equation is linearized. The corresponding estimators from the equation with and without log should be equivalent, because they differ only in scale. Kantz and Schereiber (2003, p. 109) provide further insight regarding this equivalence.
[10] Casella and Berger (1990) provide a full Bernoulli pdf description.

$$\sigma^2 = p[1-p]$$

Equalizing equations (7) and (8) yields the following:

$$p = \phi(\gamma_0 + \gamma_1 Abilitybis)$$

Thus, the moments of the population are as follows:

$$\mu = \phi(\gamma_0 + \gamma_1 Abilitybis)$$

$$\sigma^2 = \phi(\gamma_0 + \gamma_1 Abilitybis)[1 - \phi(\gamma_0 + \gamma_1 Abilitybis]$$

The sample moments for equation (8) are as follows:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = s^2$$

where $\bar{x}$ is the sample mean and s^2 is the sample variance.

To link the theory to the data, take the population moments (theory moments) and sample moments (empiric moments) and equalized them. Solve the following system of two equations:

$$p = \bar{x}$$

$$p(1-p) = s^2$$

Solving this system yields the following:

$$p = \bar{x}$$

$$p(1-p) = \bar{x} - \bar{x}^2$$

Thus, a synthetic data can be constructed for $Abilitybis$, considering the theory and empirical moment values of $\mu = p = \bar{x}$, and $\sigma^2 = p(1-p) = \bar{x} - \bar{x}^2$. In this way, the reliability of inference based on its statistical adequacy is secured. Additionally, these results are backed by the analytical model, which gauges the same functional relationships, information and optima among the relevant variables.[11]

Once constructed, $Abilitybis$ generates a standard normal cdf from the random uniform variable $Abilitybis$ with the values of $\mu = 0$ and $\sigma^2 = 1$. This computation will produce the variable $p = \phi(\gamma_0 + \gamma_1 Abilitybis)$. This procedure assures that the previous computation is correct.

Equation (7) can represent a probit model. Its log likelihood function is as follows:

$$lnPr(MS) = MSln[\phi(\gamma_0 + \gamma_1 Abilitybis)] + (1 - MS)\ln\,[1 - \phi(\gamma_0 + \gamma_1 Abilitybis)]$$

The maximum likelihood estimator for the variable $MSln[\phi(\gamma_0 + \gamma_1 Abilitybis)]$ is 2.887, with a standard error of 0.927. These figures for $(1 - MS)\ln[1 - \phi(\gamma_0 + \gamma_1 Abilitybis)]$ are -1.875 and 0.916, respectively.

Thus, plugging the above estimators values into equation (7) a value for MS can be obtained. Table 3 reports the results after estimating the corresponding equation system for different sample sizes.

[11] Spanos (2011).

n= 20 observations case

The initial parameter values for this scenario are as follows:

$\alpha_0 = 10$;
$\alpha_1 = 20$;
$\alpha_2 = 30$.

After running a regression for equation (2) using the method of ordinary least squares, the following estimators are obtained:

$\hat{\alpha}_0 = 32.9122$;
$\hat{\alpha}_1 = 0.0000$;
$\hat{\alpha}_2 = 27.6321$.

n= 50 observations case

$\hat{\alpha}_0 = 109.3834$;
$\hat{\alpha}_1 = 0.0000$;
$\hat{\alpha}_2 = 29.2389$.

n= 100 observations

$\hat{\alpha}_0 = 81.0565$;
$\hat{\alpha}_1 = -0.0000$;
$\hat{\alpha}_2 = 30.2680$.

Table 3. Estimator convergence. Sample sizes 20, 50 and 100.

Estimator	Differences (%) 20 observations	Differences (%) 50 observations	Differences (%) 100 observations	Improvement?
$\hat{\alpha}_1$	0.000	0.000	0.000	No
$\hat{\alpha}_2$	-7.893	-2.5735	0.8933	Yes

From the results reported in Table 3, it can be noted that the estimate coefficient for α_1 is 0 for 20; 50 and 100 observations. This means that once the variable MS is controlled by an observable random event, such as $Abilitybis$, it has an impact close to zero in the determination of income.

3.d. Five scenario

In this case, consider the following values for σ_I^2; γ_1 and α_1:

$\alpha_1 = 50;$
$\gamma_1 = 60;$
$\sigma_I^2 = 2.$

The corresponding estimators and their differences with respect to the initial parameters values are reported in Table 4.

n= 20 observations

$\hat{\alpha}_0 = 17.4690;$
$\hat{\alpha}_1 = 0.0000;$
$\hat{\alpha}_2 = 31.4482.$

n= 50 observations

$\hat{\alpha}_0$ =109.3834;

$\hat{\alpha}_1$ =0.0000;

$\hat{\alpha}_2$ =29.2389.

n= 100 observations

$\hat{\alpha}_0$ =137.6044;

$\hat{\alpha}_1$ = −0.0000;

$\hat{\alpha}_2$ =29.7237.

Table 4. Estimator convergence. Sample sizes 20, 50 and 100.

Estimator	Differences (%) 20 observations	Differences (%) 50 observations	Differences (%) 100 observations	Improvement?
$\hat{\alpha}_1$	0.000	0.000	0.000	No
$\hat{\alpha}_2$	4.827	-2.537	-0.921	Yes

After changing the parameters values for σ_I^2; γ_1 and α_1, it is found that the variable *MS* continues to exhibit a virtually null impact over the determination of income as in Table 3, under different initial parameter values. This implies that *MS* was only significant for the econometric model when it incorporates the omitted variable *Ability*. Once an endogeneity treatment based on the IV *Abilitybis* is implemented, the variable MS stops being a relevant factor in the determination of income.

Conclusion

By increasing the sample size, the income estimators begin converging with the assigned initial parameter values. Perhaps these estimators become consistent, implying efficiency and a decreasing bias as an asymptotic result. Once the

endogeneity problem embedded in a hypothetical equation system is controlled by the observable IV *Abilitybis*, it is found that *MS* has virtually a nil impact in the determination of income. This result suggests that *MS* has nothing to do with the determination of income and that the statistical inference from scenarios 3.a. and 3.b. is incorrect, since they have the endogeneity problem without treatment. Scenario 3.c. applies GMM to estimate equation (7) assuring a correct equation system computation. Scenario 3.d. treats the endogeneity problem by means of IV and its results are consistent under different initial parameter values. The recommendation derived from the use of synthetic data and simulation scenarios is to find an exogenous determinant of income to avoid incorrect statistical inference induced by the endogeneity problem. In case that exogenous determinants of income are not available, a feasible treatment for endogeneity by means of synthetic data consist in increasing the sample size. To see this consider scenario 3.c. asymptotic results, which are similar to scenario 3.d. IV endogeneity correction results.

References

Alvarez, J., and M. Arellano. 1998. "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators." *Working Paper no. 9808* CEMFI, pp. 1-64.

Anderson, T. W., and C. Hsiao. 1981. "Estimation of Dynamic Models with Error Components." *Journal of the American Statistical Association* 76(375), pp. 598-606.

Carbajal, C. 2013. "How to use Matlab in a Statistical Application." Unpublished mimeo.

Casella, G. and R. L. Berger. 1990. *Statistical Inference*, Belmont: Duxbury Press.

Greene, W. H. 2012. *Econometric Analysis*, New Jersey: Pearson Education, Inc.

Hamilton, J. D. 1994. *Time Series Analysis*, New Jersey: Princeton University Press.

Hayashi, Fumio. 2000. *Econometrics*, New Jersey: Princeton University Press.

Hart, P. E., G. Mills and J. K. Whitaker. 1964. *Econometric Analysis for National Economic Planning*, London: Butterworths.

Hogg, R. V., J. W. McKean and A. T. Craig. 2013. *Introduction to Mathematical Statistics*, New Jersey: Pearson Education, Inc.

Kantz, H. and T. Schereiber. 2003. *Nonlinear Time Series Analysis*, Cambridge: Cambridge University Press.

Kuh, E., J. W. Neese and P. Hollinger. 1985. *Structural Sensitivity in Econometric Models*, New York: John Wiley and Sons.

Spanos, A. 2011. "Foundational Issues in Statistical Modeling: Statistical Model Specification and Validation." *Rationality, Markets and Morals* 2, pp. 146-178.