Linear regression: Endogeneity

Alex Enrique Amaguaya Pacalla - ID: 242833

January 12, 2024

Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

IV	Instrumental Variable
OLS	Ordinary Least Square
T2LS	Two-Stage Least Square

Nomenclature

ε	Residual
β	True parameters
$\hat{\beta}$	Estimated parameters
$\mathbf{\Sigma}$	Variance-Covariance Matrix
$\mathbf{x}_1$	Endogenous Regressor
$\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	Exogenous Regressors
$\mathbf{z}_1, \mathbf{z}_2$	Instrumental Variables
y_1	Dependent Variable

1 Introduction

In the framework of linear regression, the exogeneity assumption plays an important role because it states that the residuals are not related to the regressors and, therefore, allows us to make accurate inferences about the parameters. In other words, this assumption is one of the strongest because it states that the variable of interest is uniquely explained by a set of regressors. However, this assumption is often not satisfied in practice and therefore produces problems in the estimation of the parameters. This type of problem in the field of linear regression is known as endogeneity.

The endogeneity problem occurs when one or some of the regressors are correlated with the error term and leads to a biased estimation of the estimators and also affects the statistical inference of the model. In models where the inference of the estimators is essential, this problem does not allow conclusions to be drawn about the causal relationships between the variables used.

As mentioned above, endogeneity causes biased estimators, and in addition to that this problem produces inefficient standard errors and erroneous hypothesis testing. If this problem is not mitigated, it can lead to misleading conclusions and this would lead to inadequate decision making. This is why some methods have been developed to tackle this problem.

One of the methods proposed to address the endogeneity problem is the instrumental variables (IV) method. This approach is characterized by the use of external variables, which are called instruments, that have some characteristics such as that they must be correlated with the endogenous variables and that they must not be correlated with the residuals of the model. The instrumental variables allow to extract the exogenous variation of the variables that have endogeneity problems and then use that variation to explain the dependent variable of the model. This process allows to calculate unbiased estimates of the parameters through the external variables known as instruments.

This paper covers the endogeneity problem in the following sections. In the literature review section some theoretical and applied research papers on this topic are presented, then in the model description section it is explained in detail how the IV method works to combat endogeneity. On the other hand, in the experiments and results section two scenarios are set up and simulations are performed to compare the estimates with and without IV. A more detailed explanation of the simulations is given in this section. And finally, the results of the simulations will be discussed in the conclusion section.

2 Literature Review

The idea of IV was first introduced by Philip G. Wright and he sought to determine the supply and demand for butter using the data of prices and quantities from the American market. Then, he noticed that the price affected the demand and supply functions at the same time, therefore it was impossible to construct immediately a function that explained one of them with the available data. Based on this problem, Philip determined that he required an additional variable that helped to correct this issue and it had to be associated only with one of them. The regional rainfall was used as an IV of the supply and so the functions were able to be estimated. After this, the IV concept was used in other problems and some examples are listed in this section.

Angrist and Krueger (1991) study the causal relationship between education and wages, and used some additional control variables. The paper mentions that education is probably correlated with the error term, and for that reason, they used a dummy variable of the quarter of birth date as IV for the education regressor. After that, they use the exogenous variation of education, regressing education with date of birth and some controls, in the wage equation. The results of the wage equation estimated by Ordinary Least Square(OLS) and Two-Stage Least Square (TSLS using season of birth as an instrument for education) are very similar. And the positive effect of education on wages persiste in all specifications but at different levels.

Fonseca, Michaud, and Zheng (2019) examine the effect of education on some health outcomes using a probit model and an IV. Compulsory schooling laws are used as an IV and represent the minimum years of education required for an individual. In the first stage, education is regressed on the IV and some individual and country control variables. Then, in the second stage, a model is estimated between the health outcome variables and education and a set of control variables such as gender, income, marital status, and others. Based on the estimated results, the authors conclude that there is a high reduction in reporting poor health for those affected after the reform year.

Angrist and Lavy (1999) analyze the influence of class size on school performance (math and reading skills tests) and use a rule called Maimonides imposed by the 12th-century rabbis as an IV for class size. This rule consists of having a maximum class size of 40 students. The rule is then used to calculate the instrumental variable that is used in the first stage of estimating the effects of class size on test scores. In the second stage, a regression is estimated between test scores, class size (calculated in the first stage), and some school features as control variables. The results of this study show that reducing class size increases test scores for fourth and fifth graders.

Rouse (1995) studies the impact of community colleges on educational attainment. Educational attainment in this paper is measured based on years of education or the attainment of a bachelor's degree. Access to community colleges is used as an IV in order to deal with the problem of self-selection into types of colleges. In the first stage, community college attendance (2 or 4 years) is regressed on family and state background variables, distance, and community college fees. The second stage estimates the impact of community college attendance (obtained from the first stage) on educational attainment and additionally includes control variables. The results of this study show that community colleges could apparently increase the total years of education. And in the case of bachelor's degree attainment, community college attendance does not seem to change the probability of attainment.

Puhani and Weber (2006) estimate the effect of school entry age on educational attainment using three different data sets for Germany. These data sets are merged and contain information on children at different stages, such as elementary school, secondary school, and several years after finishing secondary school. Due to the endogenous problem of school entry age, an IV is used and is calculated as a function of the month of birth and the start of the school year. This research uses different versions of the IV following the Hamburg Accord. The results show a positive influence (when the IV is used in the model) on the school performance of students who enter school at age seven rather than at age six.

3 Model Description

This section describes the approach to tackling the endogeneity problem using IV. Figure 1 shows a diagram of this issue with the relationships between the independent variables, error, and target variable. The variables $x_{(1)}$ are exogenous and meanwhile $x_{(2)}$ are endogenous. The IV approach to endogeneity is to attempt to break the relationship between endogeneity variables ($x_{(2)}$) and the error (ε).

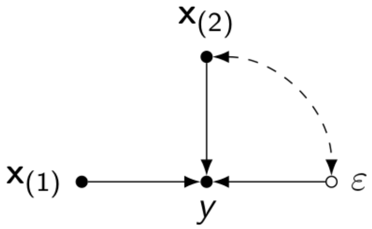

Figure 1: Endogeneity problem

The potential solution in this diagram is to find a variable w that affects the target variable y through the exogenous part of $x_{(2)}$. Figure 2 shows that the variable w satisfies three conditions. The first is that w influences the independent variable y through $x_{(2)}$. The second is that w has to be exogenous, and it means that w is not related to the error. And third, w has to be correlated with the endogenous variables.

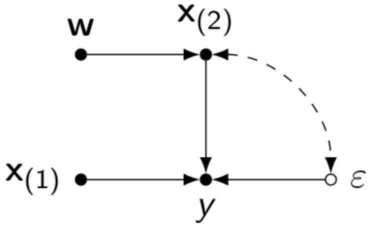

Figure 2: Endogeneity solution using IV

To explain the solution to the endogeneity issue via instrumental variables, some assumptions must be mentioned below:

Assumption 1 (Linearity).

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i; \qquad i = 1, ..., N; \qquad \beta = (\beta_1, ..., \beta_K)'$$

The first assumption is linearity, and it means that the independent variable is linearity explained by the set of regressors $\mathbf{x}$.

Assumption 2 (Independent and identically distributed - I.I.D)

The sequence $(y_i, \mathbf{x}_i, \mathbf{w}_i)_{i \in \mathbb{N}}$ is i.i.d. The second assumption means that each observation i is independent and identically distributed concerning the other observations. This assumption uses the initial variables $(y_i, \mathbf{x}_i)$ and a set of variables that help to combat the endogeneity problem ($\mathbf{w}_i \in \mathbb{R}^M$). This new set of variables is observable and the variables are not considered explanatory features in the linear model equation.

Assumption 3 (Orthogonality from Instrumental Variables)

Now, we assume that L of the set of $(K + M)$ variables in $(\mathbf{x}_i', \mathbf{w}_i')'$ are orthogonal to the model's error. The L variables are gathered in the vector $\mathbf{z}_i \in \mathbb{R}^L$, and this assumption holds:

$$\mathbb{E}[\mathbf{z}_i \varepsilon_i] = 0$$

The stricter conditional mean assumption ($\mathbb{E}[\varepsilon_i | \mathbf{z}_i] = 0$) would imply this assumption ($\mathbb{E}[\mathbf{z}_i \varepsilon_i] = 0$), and this means that the $\mathbf{z}_i$ variables are not related with the error term ε_i. In this context, the z_i variables are called instruments and represent the exogenous regressors of the set $(\mathbf{x}_i, \mathbf{w}_i)$.

Assumption 4 (Rank Condition)

This assumption is composed of two items:

- The population matrix: $\mathbb{E}[\mathbf{z}_i \mathbf{z}_i'] \equiv Q_{ZZ}$ isn't singular. The elements of $\mathbf{z}_i \mathbf{z}_i'$ will turn out to be i.i.d and using the Law of Large Numbers (LLN), we get $\frac{1}{N} Z'Z = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i' \xrightarrow{P} Q_{ZZ}$. This means that the average estimation of $Z'Z$ converges in probability to the expected value of $\mathbb{E}[\mathbf{z}_i \mathbf{z}_i']$.

- The population matrix: $\mathbb{E}[\mathbf{z}_i \mathbf{x}_i'] \equiv Q_{ZX}$ has full column rank. And similarly to the previous matrix, each element is i.i.d and using the LLN, we obtain $\frac{1}{N} Z'X = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i' \xrightarrow{P} Q_{ZX}$. It means that the instrumental variables are adequately linearly related to the regressors from the model equation $(\mathbf{x}_i)$. It also means that the average estimation of $Z'X$ converges in probability to the expected value of $\mathbb{E}[\mathbf{z}_i \mathbf{x}_i']$. The rank condition ($\mathbb{E}[\mathbf{z}_i \mathbf{x}_i'] = K$) is satisfied if and only if $L \geq K$. Therefore, it is necessary to have many instrumental variables as regressors $(\mathbf{x}_i)$, and is also important to remember that the exogenous features can be instrumented by themselves.

Identification Process:

The first condition of the previous assumption lets to estimate the coefficient vector, $\pi_k \in \mathbb{R}^L$, of the best predictor for each variable $\mathbf{x}_{ik}$ where $k = 1, ..., K$. And, the following equation shows how the previous is used for the estimation.

$$\pi_k = (\mathbb{E}[\mathbf{z}_i \mathbf{z}_i'])^{-1} \mathbb{E}[\mathbf{z}_i \mathbf{x}_{ik}']$$

Using the information it is possible to estimate the best predictor of the regressor based on the instrumental variable, the regressors are possible to decompose into two terms, the exogenous and (potentially) endogenous components, and considering that the error term $\mathbf{v}_{i,k}$ is not related with the instrumental variable $\mathbf{z}_i$ (or in other words, $\mathbb{E}[\mathbf{z}_i \mathbf{v}_{i,k}] = 0$). Now, the term π_k of the equation represents the parameter vector of the best linear predictor of $\mathbf{x}_{ik}$, and the term $\mathbf{v}_{i,k}$ is the error term of this equation (endogenous component). In addition, in this equation, the linear combination between the instrumental variable and the vector of the best parameters produces the best prediction for the endogenous variable.

$$\mathbf{x}_{ik} = \mathbf{z}_i' \pi_k + \mathbf{v}_{i,k}$$

The best prediction can be rewritten as $\mathbf{x}_{ik}^* = \mathbf{z}_i' \pi_k$. Now, using the previous equation and the assumption that instrumental variables are not related to the error term of that equation ($\mathbb{E}[\mathbf{z}_i \mathbf{v}_{i,k}] = 0$), the following equation can be rewritten as follows. The term $\mathbf{x}_{ik}^*$ constitutes the exogenous component and $\mathbf{v}_{i,k}$ is the endogenous component of the variable $\mathbf{x}_{ik}$.

$$\mathbf{x}_{ik} = \mathbf{z}_i' \pi_k + \mathbf{v}_{i,k} = \mathbf{x}_{ik}^* + \mathbf{v}_{i,k}$$

The best prediction of the $\mathbf{x}_{ik}$ regressors can also be expressed in matrix notation and it can rewritten as $\mathbf{x}_i^* = \Pi' \mathbf{z}_i$. The Π matrix represents the parameters of the best prediction for all regressors, it can be estimated by Ordinary Least Square (OLS) using this formula: $\Pi = (\mathbb{E}[\mathbf{z}_i \mathbf{z}_i'])^{-1} \mathbb{E}[\mathbf{z}_i \mathbf{x}_i']$.
On the other hand, the term $\mathbf{x}_i^*$ contains the exogenous part (related to ε_i) of the regressors due to the assumption of the orthogonality of the instruments ($\mathbb{E}[\mathbf{z}_i \varepsilon_i] = 0$), and therefore it holds that the $\mathbb{E}[\mathbf{x}_i^* \varepsilon_i] = 0$. This means that the new estimated regressors $\mathbf{x}_i^*$ of the previous equation are not related to the error term ε_i because the components of these are also not related to the error.

$$\begin{gathered} x_i^* = \Pi' z_i \\ \mathbb{E}[\mathbf{x}_i^* \varepsilon_i] = \mathbb{E}[\Pi' \mathbf{z}_i \varepsilon_i] \\ \mathbb{E}[\mathbf{x}_i^* \varepsilon_i] = \Pi' \mathbb{E}[\mathbf{z}_i \varepsilon_i];\ \textit{and use}\ \mathbb{E}[\mathbf{z}_i \varepsilon_i] = 0 \\ \mathbb{E}[\mathbf{x}_i^* \varepsilon_i] = 0 \end{gathered}$$

Based on the identification process, the instrument assumptions can be summarized in the following two items:

1. Instrument Exogeneity: $\mathbb{E}[\mathbf{z}_i \varepsilon_i] = 0$
2. Instrument relevance: $\mathbb{E}[\mathbf{z}_i \mathbf{x}_i'] = K$

The first item means that the instruments are not correlated with the error term (ε_i), and it allows isolating the exogenous ($\mathbf{x}_i^*$) and endogenous ($\mathbf{v}_i$) part of the regressors. On the other hand, the second assump-

tion guarantees that the inverse matrix of the term $\mathbb{E}[\mathbf{x}_i^*\mathbf{x}_i']$ or $\mathbb{E}[\mathbf{x}_i^*(\mathbf{x}_i^*)']$ exists, and therefore the parameter vector β is identified. In other words, this means that the existence of the inverse makes it possible to calculate the parameter vector. The equality between the terms $\mathbb{E}[\mathbf{x}_i^*\mathbf{x}_i'] = \mathbb{E}[\mathbf{x}_i^*(\mathbf{x}_i^*)']$ is shown below:
Using the terms $\mathbf{x}_i^* = \Pi'\mathbf{z}_i$ and $\Pi = (\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'])^{-1}\mathbb{E}[\mathbf{z}_i\mathbf{x}_i']$.

$$\mathbb{E}[\mathbf{x}_i^*\mathbf{x}_i'] = \mathbb{E}[\Pi'\mathbf{z}_i\mathbf{x}_i'] = \Pi'\mathbb{E}[\mathbf{z}_i\mathbf{x}_i']$$
$$\mathbb{E}[\mathbf{x}_i^*\mathbf{x}_i'] = ((\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'])^{-1}\mathbb{E}[\mathbf{z}_i\mathbf{x}_i'])'\mathbb{E}[\mathbf{z}_i\mathbf{x}_i']$$
$$\mathbb{E}[\mathbf{x}_i^*\mathbf{x}_i'] = \mathbb{E}[\mathbf{x}_i\mathbf{z}_i'](\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'])^{-1}\mathbb{E}[\mathbf{z}_i\mathbf{x}_i']$$

Now, using the right-hand side of the first equality, we obtain:

$$\mathbb{E}[\mathbf{x}_i^*(\mathbf{x}_i^*)'] = \mathbb{E}[\Pi'\mathbf{z}_i(\Pi'\mathbf{z}_i)'] = \mathbb{E}[\Pi'\mathbf{z}_i\mathbf{z}_i'\Pi]$$
$$\mathbb{E}[\mathbf{x}_i^*(\mathbf{x}_i^*)'] = \Pi'\mathbb{E}[\mathbf{z}_i\mathbf{z}_i']\Pi$$
$$\mathbb{E}[\mathbf{x}_i^*(\mathbf{x}_i^*)'] = \mathbb{E}[\mathbf{x}_i\mathbf{z}_i'](\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'])^{-1}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'](\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'])^{-1}\mathbb{E}[\mathbf{z}_i\mathbf{x}_i'], and\ we\ know\ that\ (\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'])^{-1}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'] = I$$
$$\mathbb{E}[\mathbf{x}_i^*(\mathbf{x}_i^*)'] = \mathbb{E}[\mathbf{x}_i\mathbf{z}_i'](\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'])^{-1}\mathbb{E}[\mathbf{z}_i\mathbf{x}_i']$$

First and second estimation stages

The identification process provides information that it is possible to extract the exogenous part of the regressors. In the first stage, the parameters of the best predictions of the instruments on the regressors are estimated. In other words, for each variable a regression model is estimated and the parameters are obtained.

The exogenous vector component $\mathbf{x}_i^* = \Pi'\mathbf{z}_i$ can be transformed to the matrix notation for all regressors $X^* = Z\Pi$, and the following equation explains how the prediction of each regressor is estimated.
Using $\pi_k = (\mathbb{E}[\mathbf{z}_i\mathbf{z}_i'])^{-1}\mathbb{E}[\mathbf{z}_i\mathbf{x}_{ik}']$. Then, estimate π_k via OLS, we get:

$$\hat{\pi_k} = (Z'Z)^{-1}Z'\mathbf{x}_{(k)}, where\ \mathbf{x}_{(k)} = (\mathbf{x}_{1k}, ..., \mathbf{x}_{Nk})'$$

The subscript of the above equation represents the k initial regressors, and the $\hat{\pi_k}$ is the least squares estimator of the following equation.

$$\mathbf{x}_{ik} = \mathbf{z}_i'\pi_k + \mathbf{v}_{i,k}$$

Now, the exogenous matrix component X^* can be calculated using the above equations, and the matrix of predictions $\hat{X^*}$ of the instruments on the regressors is obtained.

$$\hat{X^*} = Z\hat{\Pi},\ \textit{we know that}\ \hat{\Pi} = (Z'Z)^{-1}Z'X$$
$$\hat{X^*} = Z(Z'Z)^{-1}Z'X$$
$$\hat{X^*} = P_Z X$$

In the first stage, it is possible to detect whether the instruments are good or bad predictors of the endogenous regressors, in other words, whether the instruments explain much or little variation on the endogenous variable. Instruments that provide little information on the regressors are known as weak instruments and they can provide inaccurate coefficient estimates on the endogenous variables. In addition, this problem influences the distribution of the estimator and causes the distribution to deviate significantly.
Once the estimation of the first stage of the estimation is finished, the new model for the second stage is established. In the second stage, the model keeps the same dependent variable, but the matrix of initial regressors is exchanged for the predictions of the exogenous part of them. The following equation shows how the new model is proposed for this second stage of estimation.

$$y = \hat{X}^*\beta + \epsilon$$

Using the above equation, the vector of parameters is estimated by OLS. The estimation process is detailed below.

$$\hat{\beta} = (\hat{X}^{*'}\hat{X}^*)^{-1}\hat{X}^{*'}y$$
$$\hat{\beta} = (X'P_Z'P_ZX)^{-1}X'P_Z'y,\ \textit{we know that}\ P_Z'P_Z = P_Z\ \textit{because this matrix is idempotent.}$$
$$\hat{\beta} = (X'P_ZX)^{-1}X'P_Z'y$$
$$\hat{\beta} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y = \hat{\beta}_{2SLS}$$

The above equations show the importance of the existence of the inverse of the term $(\hat{X}^{*'}\hat{X}^*)$, and it is because the inverse of this term allows estimating the model parameters in the second stage of the estimation process.

4 Simulation details

This section sets out some scenarios in which instrumental variables are used to estimate the parameters of the regression model. The scenarios proposed for the simulations are as follows:

1. The first scenario is a model with an endogenous, exogenous, and instrumental variable.

2. The second scenario is a model with an endogenous regressor, some exogenous variables, and an in-

strumental variable.

In both cases, an instrumental variable will be considered to deal with the endogeneity problem of the models. For the Monte Carlo simulations of the above scenarios, it is necessary to define some details of this process such as the number of observations, the number of Monte Carlo runs, the data generation setup, the design of the variance-covariance matrix of the variables, the type of instruments to be used, and the true parameters. The number of observations (n) for the simulation considers two values $n = 100$ and $n = 1000$. On the other hand, the Monte Carlo runs (N) proposed for the simulations also have two values $N = 100$ and $N = 1000$. The data generation setup used for regressors, instruments, and errors is a zero mean and one variance normal distribution. For the values of the variance-covariance array, the matrix considers the correlations that may exist between the variables used and also the type of instruments used for each of the scenarios. The types of instruments used will be weak and strong instruments, and these are differentiated by whether they have a high or low correlation with the endogenous variable. For the simulations, two correlation values are considered $\rho_{\mathbf{z}_i,\mathbf{x}_i} = 0.7$ and $\rho_{\mathbf{z}_i,\mathbf{x}_i} = 0.1$. The algorithm for the estimation process is detailed in more detail below.

Algorithm 1: Estimation process

```
Data: Scarlars: $n > 0$ and $N > 0$; Mean vector: $\vec{\mu}$; Var-cov matrix: $\mathbf{\Sigma}$; True values: $\vec{\beta}$; Error
      Variable name: $\vec{e}$; Endog. and Exog. Variables Names: $\vec{x} \in \mathbb{R}^{1\times K}$; Instrumental Variables
      Names: $\vec{z} \in \mathbb{R}^{1\times L}$
Result: $\vec{\hat{\beta}}_{OLS}, \vec{\hat{\beta}}_{IV}$
for $t \leftarrow 1$ to $N$ do
    $Sample \leftarrow f(n, \vec{\mu}, \mathbf{\Sigma})$ ;        /* data generation setup with multiv.normal distrib.  */
    $X \leftarrow Sample[\vec{x}]$;
    $\vec{\epsilon} \leftarrow Sample[\vec{e}]$;
    $\vec{y} \leftarrow f(\vec{\beta}, X, \vec{\epsilon})$;
    $Z \leftarrow Sample[\vec{z}]$;
    $\vec{\hat{\beta}}_{OLS} \leftarrow f(X, \vec{y})$ ;                              /* Estimation without IV */
    $\vec{\hat{\beta}}_{IV} \leftarrow f(X, Z, \vec{y})$ ;                               /* Estimation with IV */
end
```

First, the proposed algorithm for the simulations establishes some hyperparameters such as the number of observations, the number of runs, the vector of means, the variance-covariance matrix, the vector of true parameters, and the name of the regressors (exogenous and endogenous variables), instrumental variables and error term. After this, the number of runs is repeated N times. For each run, a sample is created using the vector of means, variance-covariance matrix, the number of observations n, and the type of distribution selected for the variables. This sample is a matrix that contains the data for the regressors, instrumental variables, and the error. The simulations use a normal distribution for all variables during the estimation process.

Once the variable data has been generated, the regressor matrix, instrumental variables, and error term can be constructed. The matrix or vectors are obtained from the sample matrix and then are used to estimate the dependent variable vector. The dependent variable uses the vector of true parameters, the matrix of regressors, and the error term for its estimation. Once the dependent variable has been computed, it is possible to estimate the vector of parameters using the regressors matrix, the instrumental variables, and the vector of the dependent variable. Two cases are considered for the estimation of the parameter vector, when instrumental variables are used and when they are not used. For these two cases, it has been decided to add a subindex to the vector of parameters in order to be able to differentiate them. The first is $\vec{\hat{\beta}}_{OLS}$ and uses only the regressors and estimates the coefficient parameter by OLS, while the second is called $\vec{\hat{\beta}}_{IV}$ and uses the instruments for the estimation of the coefficients by T2LS.

This algorithm is used for the estimation of the proposed scenarios together with the combination of the hyperparameters $(n, N, \rho_{textbfz_i, textbfx_i})$. The table below shows a summary of the values of the possible combinations of the hyperparameters for each of the scenarios proposed at the beginning.

hyperparameter/scenario	Value
First scenario	Variables: 1 exogenous, 1 endogenous, 1 instrument
Second scenario	Variables: 3 exogenous, 1 endogenous, 1 instrument
Number of observations (n)	100, 1000
Number of runs (N)	100, 1000
Instrument strength $(\rho_{\mathbf{z}_i, \mathbf{x}_i})$	0.7, 0.1

Table 1: Hyperparameters/scenarios from the simulations

5 Experiments and Results

This section describes the simulations for the two proposed scenarios and these use the hyperparameters discussed in the previous section $(n, N, \rho_{\mathbf{z}_i, \mathbf{x}_i})$. Before the simulations, it is necessary to create the vector of means and the variance-covariance matrix of all the variables. The figure below shows the structure of the matrix and also includes the correlation with the error term. In addition, it shows that the variance of the variables is equal to one, and the endogenous variable $(\mathbf{x}_1)$ has a strong correlation with the residual and is equal to 0.6.

On the other hand, the instrumental variables in the experiments used the notation $\mathbf{z}_1$ and $\mathbf{z}_2$ (strong and weak instrument, respectively), the exogenous variables are $\mathbf{x}_2, \mathbf{x}_3$ and $\mathbf{x}_4$, and the error term is *eps*. All variables employed in the experiments have a normal distribution with mean zero and variance one. In the simulations, an intercept is used which is a vector of ones and is associated with the parameter $\hat{\beta}_0$. In addition, the variance-covariance matrix shows the relation between the endogenous regressor and the weak and strong instrument.

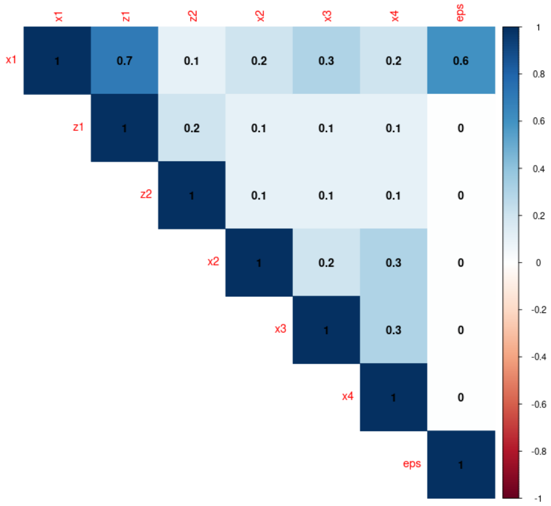

Figure 3: Variance-covariance matrix

The first simulation uses an endogenous, exogenous, and instrumental variable and then it estimates the model. This scenario uses a first combination of hyperparameters equal $n = 100, N = 100$ and $\rho_{z_i,x_i} = 0.7$. In other words, this first simulation uses the algorithm described in the previous section with a small sample, a small number of runs, and a strong correlation between the endogenous and the instrumental variable.

The table below shows the simulation results of this first experiment and shows the average parameter estimate for the 100 simulations performed. The parameter $\hat{\beta}_1$ corresponds to the endogenous regressor and the results show that the average OLS estimator without the instrumental variable is highly overfitted concerning the true value of β_1 and the results of β_0 are relatively close to the true value one. Meanwhile, the mean IV estimator of $\hat{\beta}_1$ is very close to the true value and the underfitting is relatively small. On the other hand, the IV estimation of the parameter associated with the exogenous regressor $(\hat{\beta}_2)$ is slightly overestimated and the results of the OLS method underestimate it. In addition, the table exposes the standard deviations of the parameters and this is computed by calculating the standard deviation of the simulation results. The standard deviations of the IV results are larger than the OLS method.

Parameter	True Value	OLS	IV (z_1)	OLS std.	IV (z_1) std.
$\hat{\beta}_0$	1	0.999004	1.004062	0.08052477	0.1080805
$\hat{\beta}_1$	3	3.631951	2.983357	0.07444358	0.1552467
$\hat{\beta}_2$	5	4.883493	5.009892	0.07929164	0.1151390

Table 2: Case 1 - Parameter results with $N = 100$, $n = 100$ and z_1

The figure below shows the distribution of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ during this first simulation. The vertical dashed lines represent the true values and are equal to 1 for $\hat{\beta}_0$, 3 for $\hat{\beta}_1$, and 5 for $\hat{\beta}_2$. The estimated values of $\hat{\beta}_0$ using OLS and IV are around the true value, but the situation is very different for the case of $\hat{\beta}_1$. The estimated parameters using OLS are overestimated and far away from the true value, whereas, the estimated coefficients by IV are around the true value. On the other hand, the computed results via OLS of $\hat{\beta}_2$ are underestimated from the value five, whereas, the estimated coefficients by IV are around the true coefficient.

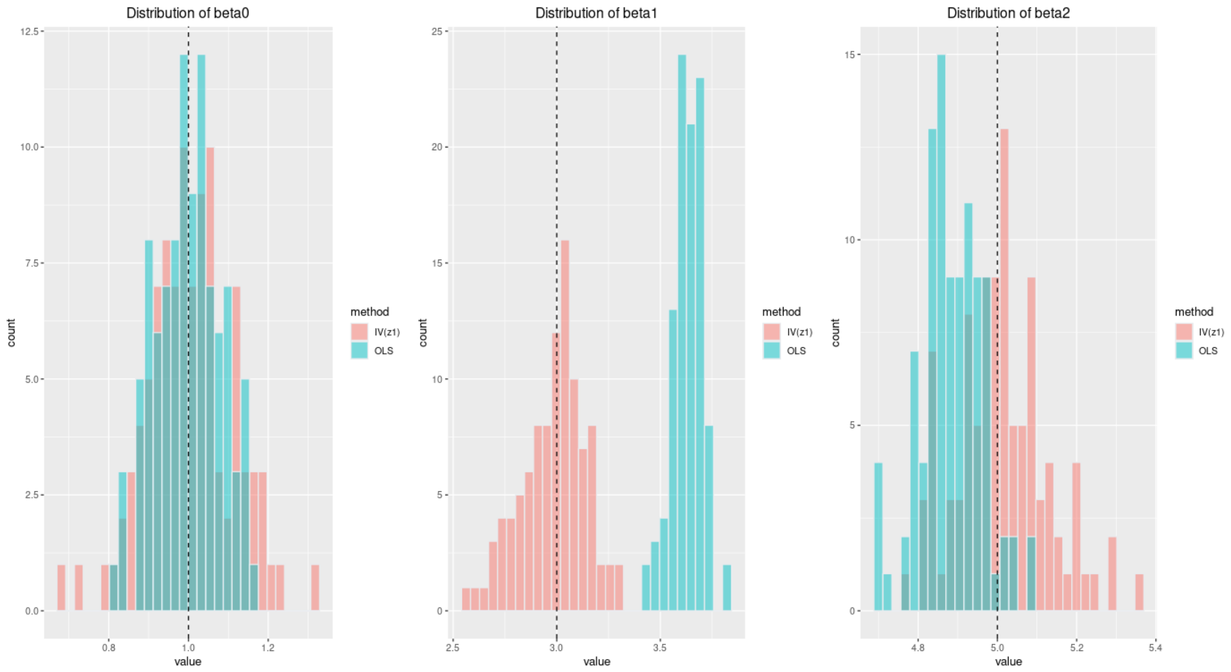

Figure 4: Case 1 - histograms with $N = 100$, $n = 100$ and z_1

The second combination of hyperparameters of the first scenario uses $n = 100, N = 100$ and $\rho_{z_i,x_i} = 0.1$. This means that this simulation employs the previous algorithm with a small sample, a small number of runs, and a weak correlation between the endogenous and the instrumental variable.

The following table shows the results of this second simulation and shows the mean parameter estimate of the 100 simulations performed but now using a weak instrument. The results show that the mean OLS estimator of $\hat{\beta}_1$ without the instrumental variable is highly overfitted concerning the true value three. Meanwhile, the mean IV estimator overestimates the true value, but the overfitting of the IV results is smaller relative to the OLS method. On the other hand, the IV and OLS results of the parameter associated with the exogenous regressor ($\hat{\beta}_2$) underestimate the value of five. Furthermore, the table shows that the standard deviations of the estimated IV parameters increase significantly.

Parameter	True Value	OLS	IV (z_1)	OLS std.	IV (z_1) std.
$\hat{\beta}_0$	1	0.999004	1.038017	0.08052477	0.3458671
$\hat{\beta}_1$	3	3.631951	3.311240	0.07444358	4.3276262
$\hat{\beta}_2$	5	4.883493	4.894441	0.07929164	1.1874252

Table 3: Case 1 - Parameter results with $N = 100$, $n = 100$ and z_2

The following figure shows the distribution of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ for the second simulation and exposes that the dispersion of the estimated coefficients using the IV method is very large compared to the OLS method. This fact occurs for all coefficients, $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$, using the results of the 100 simulations. Based on the two simulations performed so far, the estimates with the z_1 instrument are better than the estimates with the z_2 instrument because most of the estimated values are around the true value and have low dispersion. The last figures demonstrate how the strength of the instrument influences the estimation of the parameters and this can cause a small or large bias in the coefficients.

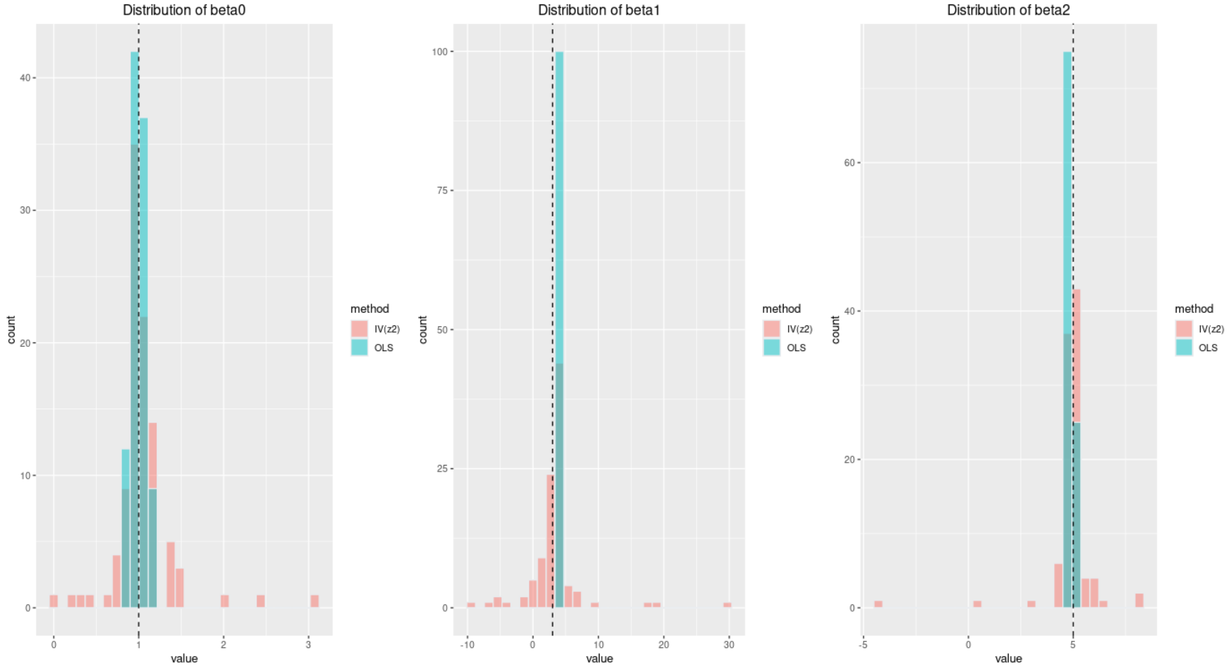

Figure 5: Case 1 - histograms with $N = 100$, $n = 100$ and z_2

The third combination of hyperparameters of the first scenario uses $n = 1000, N = 1000$ and $\rho_{z_i,x_i} = 0.7$. Now, the simulation utilizes the algorithm with a large sample, a large number of runs, and a strong correlation between the endogenous and the instrumental variable.

The table below exposes the results of this third simulation and shows the mean parameter estimate of the 1000 simulations performed. The results show that the mean OLS estimator without the instrumental variable for β_0 is slightly underestimated, whereas, the conclusions about the β_1 parameter remain similar to previous simulations, highly overfitted relative to the true value. On the other hand, the mean IV estimator of β_0 is slightly overestimated and for β_1 the result is slightly underestimated. On the other hand, the IV and OLS estimations of the parameter associated with the exogenous regressor $(\hat{\beta}_2)$ are underestimated. Concerning the standard deviation of the estimated parameters, the table shows that these were reduced for the IV and OLS methods by more than 50% compared to the first simulations when a small sample and runs were used.

The figure below shows the results for the third simulation of the estimation of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$. The graph

Parameter	True Value	OLS	IV (z_1)	OLS std.	IV (z_1) std.
$\hat{\beta}_0$	1	0.9995065	1.001826	0.02522978	0.03349296
$\hat{\beta}_1$	3	3.6251739	2.999041	0.02542158	0.04628324
$\hat{\beta}_2$	5	4.8736141	4.998840	0.02534909	0.03301900

Table 4: Case 1 - Parameter results with $N = 1000$, $n = 1000$ and z_1

on the left side of the figure shows that the parameters estimated by the OLS and IV methods are placed around the true value and their histograms are bell-shaped (normal distribution). The graph in the middle shows that the IV estimated coefficients with a strong instrument are around three, whereas, the OLS results without any instrument have values between 3.5 and 3.75. This means that although the number of observations increases, OLS estimates of $\hat{\beta}_1$ are affected by the endogeneity problem and lead to an overestimation. On the other hand, the computed results via OLS of $\hat{\beta}_2$ are underestimated from the true value, whereas, the estimated coefficients by IV are close to the true coefficient. In addition, there is a low dispersion in the estimated results based on the histogram charts.

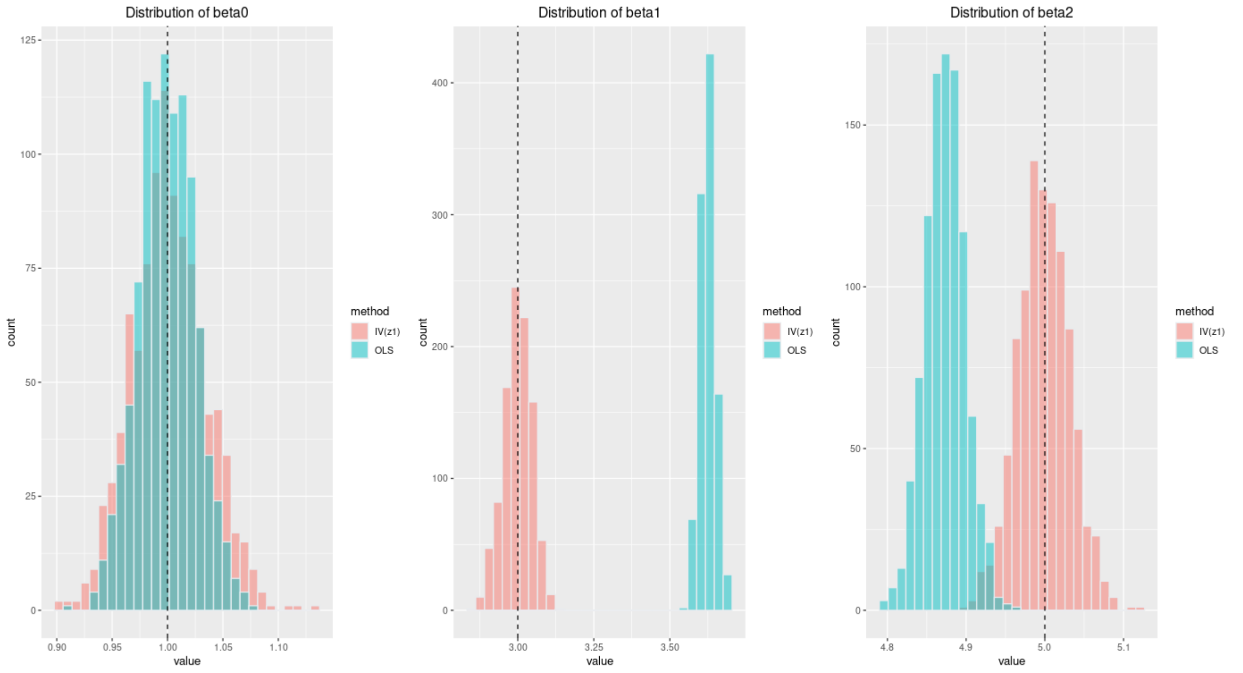

Figure 6: Case 1 - histograms with $N = 1000$, $n = 1000$ and z_1

In the fourth simulation of the first scenario, it uses a combination of hyperparameters $n = 1000, N = 1000$ and $\rho_{z_i,x_i} = 0.1$. Now, the simulation employs the algorithm with a large sample size, a large number of runs, and a low correlation between the endogenous and instrumental variable (weak instrument).

The table below presents the results of this fourth simulation and shows the mean parameter estimate after the 1000 simulations performed with a weak instrument. The mean IV estimator of β_0 is slightly underestimated relative to the previous simulations (3rd simulation: 1.001826 vs 4th simulation: 0.9970257). Meanwhile, the mean IV estimator of β_1 is further away from the true value compared to the third simulation (3rd simulation: 2.99 vs 4th simulation: 2.908). On the other hand, the IV and OLS estimations of the parameter associated with the exogenous regressor ($\hat{\beta_2}$) are overestimated and underestimated, respectively.

Parameter	True Value	OLS	IV (z_1)	OLS std.	IV (z_1) std.
$\hat{\beta}_0$	1	0.9995065	0.9970257	0.02522978	0.07616998
$\hat{\beta}_1$	3	3.6251739	2.9080908	0.02542158	2.57983331
$\hat{\beta}_2$	5	4.8736141	5.0146690	0.02534909	0.60103920

Table 5: Case 1 - Parameter results with $N = 1000$, $n = 1000$ and z_1

Regarding the standard deviation of the estimated parameters, the table exposes that the parameter β_1 estimated through a weak instrument shows a substantial increase in the dispersion concerning the third simulation. In the third simulation, the standard deviation β_1 was 0.04 and now it is 2.57, and for the other parameter, the increase of the standard deviation is minimal.

The figure below shows the results for the fourth simulation of the estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$. The graph on the left side of the figure shows that the distribution of the IV results is affected by the outliers. A similar situation occurs in the results of the other coefficients.

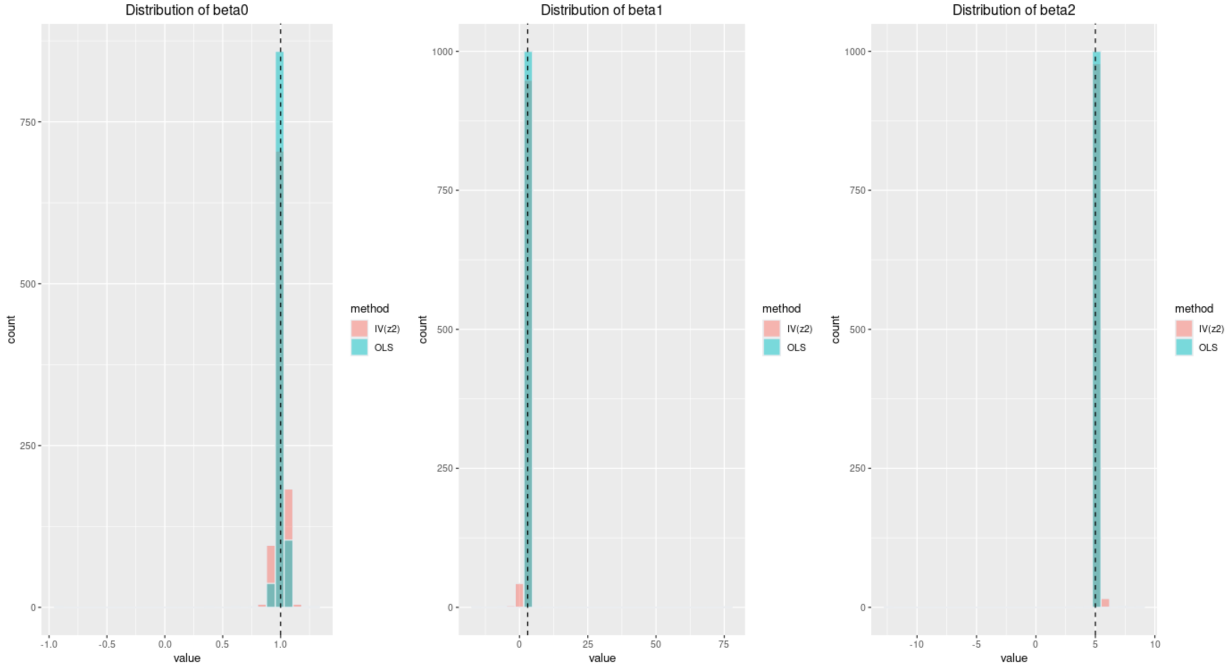

Figure 7: Case 1 - histograms with $N = 1000$, $n = 1000$ and z_2

So far only simulations for the first scenario have been performed, and now simulations for the second case are performed following the same combination of hyperparameters of the previous cases. The simulations of the second scenario are composed of one endogenous variable, one instrument and three exogenous variables. The first simulation of the second scenario employs a combination of hyperparameters equal $n = 100, N = 100$ and $\rho_{z_i,x_i} = 0.7$. This means that the simulation uses the algorithm described in the previous section with a small sample, a small number of runs, and a strong correlation between the endogenous and instrumental variable. The objective of this simulation is to evaluate the influence of the instrument on the estimation of multiple exogenous regressors.

The following table presents the simulation results of this first experiment for the second proposed scenario and shows the mean parameter estimate for the 100 simulations performed. The parameter $\hat{\beta}_1$ corresponds to the endogenous regressor, and the results show that the mean OLS estimator without the instrumental variable is highly overfitted (3.66) with respect to the true value of β_1.

Parameter	True Value	OLS	IV (z_1)	OLS std.	IV (z_1) std.
$\hat{\beta}_0$	1	0.9995567	1.005444	0.07412054	0.09616025
$\hat{\beta}_1$	3	3.6677685	2.995771	0.08082512	0.17222778
$\hat{\beta}_2$	5	4.9226266	5.000546	0.08048370	0.11440846
$\hat{\beta}_3$	2	1.8289724	2.008557	0.08450497	0.10888159
$\hat{\beta}_4$	4	3.9305894	3.982290	0.09283184	0.10740479

Table 6: Case 2 - Parameter results with $N = 100$, $n = 100$ and z_1

And for the β_0 estimate, the mean OLS estimate is relatively close to one. On the other hand, the mean of the IV estimator for (2.99) is very close to the true value and the underfit is relatively small. The mean OLS estimates of $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$ are underfitted, whereas, the mean results of the IV estimator are slightly underfitted ($\hat{\beta}_4$) or overfitted ($\hat{\beta}_2$ and $\hat{\beta}_3$). In addition, the table shows that the standard deviations from all estimated parameters in the IV simulations are larger than the OLS method.

The figure below shows the simulation results for all parameters. This figure exposes that for the parameter $\hat{\beta}_0$ the estimation results of the IV and OLS methods are around the true value, whereas, the results of the coefficient estimation $\hat{\beta}_1$ have a very different behavior concerning the other parameters. The IV method estimates are close to the value three and the OLS results are too far away from the true value. This occurs because the OLS method uses an endogenous regressor that has a high correlation with the error term and this affects the calculations.

For the results of $\hat{\beta}_2$, the figure shows that the histogram of the OLS method estimates has a positive skew, and the IV method results are around the true value, but these do not have a well-defined bell shape. On the other hand, the results of $\hat{\beta}_3$ are quite similar to those of the previous parameter. Finally, the results of the OLS method for the coefficient $\hat{\beta}_4$ are a little distant from the true value, whereas, the results of the IV do not exhibit a well-defined bell shape. In conclusion, these results demonstrate that although the IV estimates use a strong instrument, the estimation of the other parameters is affected but not significantly. Furthermore, these plots demonstrate the importance of using a strong instrument in the estimation process.

The second combination of hyperparameters of the second scenario uses $n = 100, N = 100$ and $\rho_{z_i,x_i} = 0.1$.

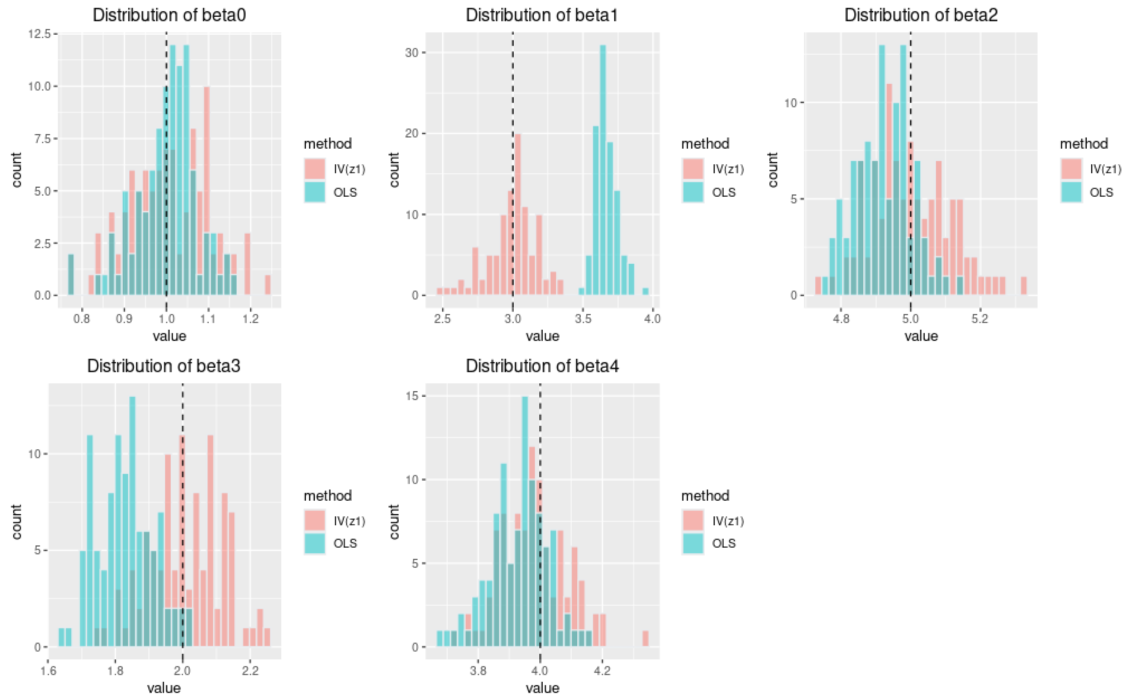

Figure 8: Case 2 - histograms with $N = 100$, $n = 100$ and z_1

This simulation utilizes the algorithm with a small sample, a small number of runs, and a weak correlation between the endogenous and the instrumental variable.

The table below presents the simulation results of this second experiment for the second proposed scenario and shows the mean parameter estimate for the 100 simulations performed with a weak instrument. The mean IV estimate of $\hat{\beta}_1$ (4.27) is very distant from the true value and the overfit is very large relative to the mean OLS estimate. On the other hand, the mean of the IV estimates of $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$ are greatly overestimated compared to the results obtained from the OLS method. The IV estimation results for $\hat{\beta}_1$ do not seem to be affected and are slightly overfitted. Furthermore, the table shows that the standard deviations of all the parameters estimated in the IV simulations are extremely large relative to the OLS method, especially for the coefficient belonging to the endogenous regressor.

Parameter	True Value	OLS	IV (z_2)	OLS std.	IV (z_2) std.
$\hat{\beta}_0$	1	0.9995567	1.017821	0.07412054	0.888053
$\hat{\beta}_1$	3	3.6677685	4.271304	0.08082512	8.490247
$\hat{\beta}_2$	5	4.9226266	4.882442	0.08048370	1.139105
$\hat{\beta}_3$	2	1.8289724	1.621703	0.08450497	2.419381
$\hat{\beta}_4$	4	3.9305894	3.919274	0.09283184	1.095849

Table 7: Case 2 - Parameter results with $N = 100$, $n = 100$ and z_2

The figure below shows the simulation results for all parameters. This figure shows that for the parameter

$\hat{\beta}_0$ the estimation results of the IV and OLS methods are around the true value, but the IV method has some extreme values and they can be seen in the histogram.

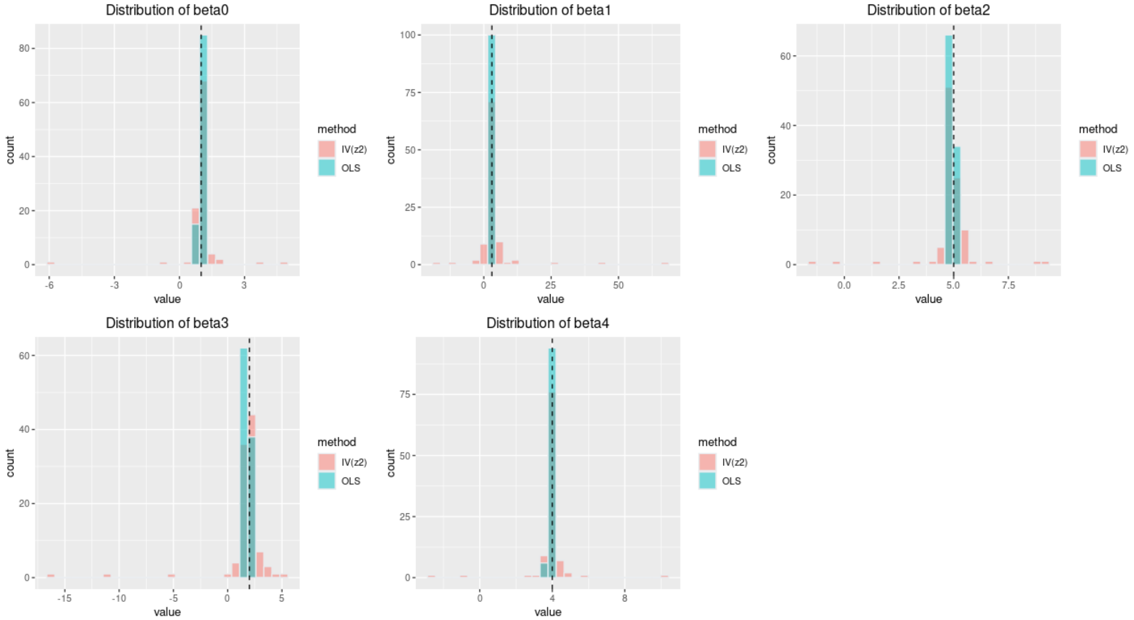

Figure 9: Case 2 - histograms with $N = 100$, $n = 100$ and z_2

A similar situation occurs for the rest of the parameters and their plots show some outliers in the histogram charts. This fact occurs because the simulation uses a weak instrument and affects the calculation of the other parameters causing these extreme values. In addition, the charts show that the intervals of the histograms expand significantly and this causes a distortion in the distribution of the results.

Now, the following simulations use a large number of observations and runs for the second proposed scenario. The third hyperparameter combination of the second scenario employs $n = 1000, N = 1000$ and $\rho_{z_i,x_i} = 0.7$. In other words, this simulation uses the algorithm with a large sample size, a large number of runs, and a strong correlation between the endogenous and instrumental variable. The instrumental variable z_1 was used in this simulation to cope with the endogeneity problem of the x_1 variable.
The following table exposes the simulation results of the third experiment for the second proposed scenario and shows the mean parameter estimate for the 1000 simulations performed with the strong instrumental variable z_1. For the parameter $\hat{\beta}_1$, the mean IV estimate of z_1 (3.001) is close to the value three, whereas, the OLS estimate (3.68) of this parameter is far away from the true value. This result states that the OLS estimate overestimates the true parameter and the overestimation is very large relative to the mean IV estimate.
On the other hand, the OLS and IV results of $\hat{\beta}_0$, $\hat{\beta}_2$ and $\hat{\beta}_4$ behave similarly. The OLS estimates are underestimated and the IV estimates are slightly overestimated. For the $\hat{\beta}_3$ results, the OLS and IV estimates are underestimated, but the underestimation of OLS (1.83) is larger than that of IV (1.99). In addition,

Parameter	True Value	OLS	IV (z_1)	OLS std.	IV (z_1) std.
$\hat{\beta}_0$	1	0.9999692	1.000323	0.02520373	0.03218942
$\hat{\beta}_1$	3	3.6801902	3.001364	0.02633809	0.04886188
$\hat{\beta}_2$	5	4.9152139	5.000412	0.02582169	0.03364120
$\hat{\beta}_3$	2	1.8304835	1.997732	0.02676377	0.03374283
$\hat{\beta}_4$	4	3.9399188	4.001327	0.02695798	0.03369826

Table 8: Case 2 - Parameter results with $N = 1000$, $n = 1000$ and z_1

the table shows that the standard deviations of all estimated parameters in the IV simulations are slightly large relative to the OLS method. These IV and OLS standard deviation results are lower than those of the previous two simulations of this scenario.

The following figure shows the histograms of the simulation results for all parameters in this third simulation. This figure shows that for the parameter $\hat{\beta}_0$ the estimation results of the IV and OLS methods are around one (true parameter), whereas, the results of the coefficient estimation $\hat{\beta}_1$ have opposite conclusions.

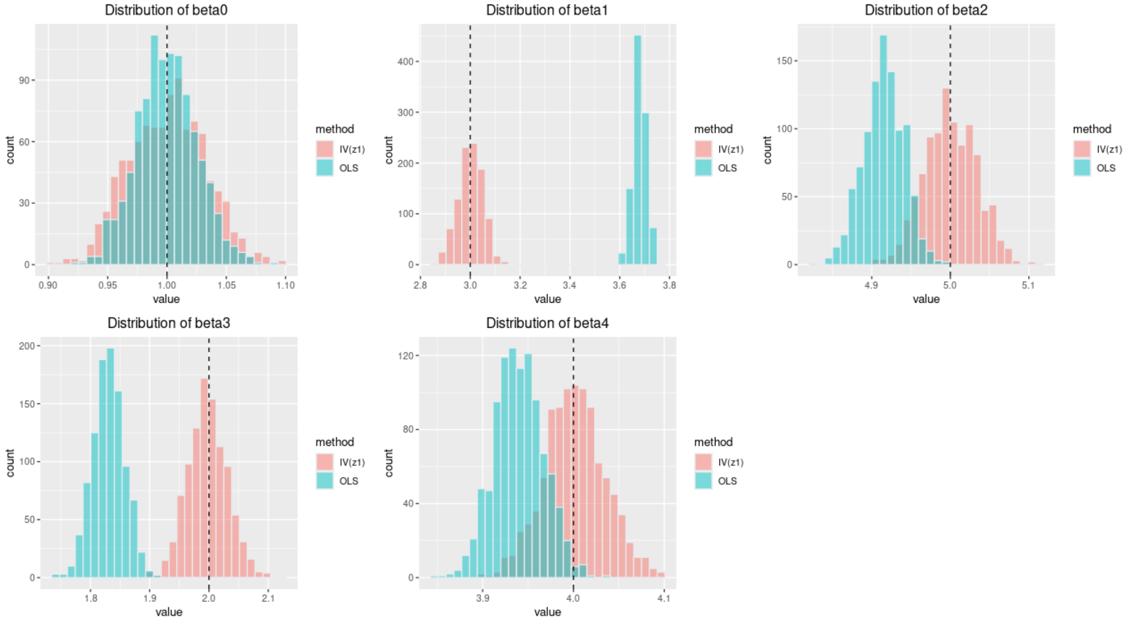

Figure 10: Case 2 - histograms with $N = 1000$, $n = 1000$ and z_1

The IV estimation results for this parameter are around the value three and the histogram has a well-defined bell shape, and the OLS method results are far from the true parameter and have no bell shape. For the results of $\hat{\beta}_2$, the figure shows that the histograms of the OLS and IV methods have a bell shape. The OLS method estimates do not approximate the actual coefficient, whereas, the IV method results for this parameter approximate the value five. The plots for the $\hat{\beta}_3$ and $\hat{\beta}_4$ parameters show similar conclusions to those for the $\hat{\beta}_2$ parameter. The plots for these parameters expose that the histograms have a well-defined bell shape. These results demonstrate that when the simulation does not use the instrumental variable it affects the calculation of the other exogenous regressors.

The fourth hyperparameter combination of the second scenario employs $n = 1000, N = 1000$ and $\rho_{z_i,x_i} = 0.1$. Now, this simulation employs the algorithm with a large sample size, a large number of runs, and a weak correlation between the endogenous and instrumental variable. The instrumental variable z_2 was used in this simulation to tackle the endogeneity problem of the x_1 variable.

The table below exposes the simulation results of the last experiment for the second proposed scenario and shows the mean parameter estimate for the 1000 simulations performed with the weak instrumental variable z_2. For the coefficient $\hat{\beta}_1$, the mean IV estimate of z_2 (2.92) underestimates the true parameter value, whereas, the OLS estimate (3.68) of this parameter is far away from the true value. This result states that the OLS estimate overestimates the true parameter and the IV method slightly underestimates the value three. On the other hand, the OLS and IV results of $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$ behave similarly. This means the OLS estimates are underestimated and the IV estimates are slightly overestimated. For the $\hat{\beta}_0$ results, the OLS and IV estimates are underestimated, but the underestimation of IV (0.995) is slightly larger than that of OLS (0.999). In addition, the table shows that the standard deviations of all estimated parameters in the IV simulations are extremely larger relative to the OLS method. These IV and OLS standard deviation results are larger than the previous simulation with a strong instrumental variable.

Parameter	True Value	OLS	IV (z_2)	OLS std.	IV (z_2) std.
$\hat{\beta}_0$	1	0.9999692	0.9951424	0.02520373	0.2278063
$\hat{\beta}_1$	3	3.6801902	2.9243560	0.02633809	3.9051344
$\hat{\beta}_2$	5	4.9152139	5.0053875	0.02582169	0.4850085
$\hat{\beta}_3$	2	1.8304835	2.0285536	0.02676377	0.9821463
$\hat{\beta}_4$	4	3.9399188	4.0050607	0.02695798	0.3413574

Table 9: Case 2 - Parameter results with $N = 1000$, $n = 1000$ and z_2

The figure below shows the results from all parameters in this last simulation. This figure exposes that for the parameter $\hat{\beta}_0$ the estimation results of the IV and OLS methods are around the true value, but the IV method has many extreme values and they can be seen in the histogram. These extreme values distort the plot and cause a considerable increase in the standard deviation of the simulations. A similar situation occurs for the rest of the parameters and their plots expose some atypical values in the histogram charts. In addition, this fact occurs because this simulation uses a weak instrument and affects the calculation of the other parameters causing these extreme values. A similar graph was obtained in the simulation with a reduced number of runs, a reduced number of observations, and a weak instrument.

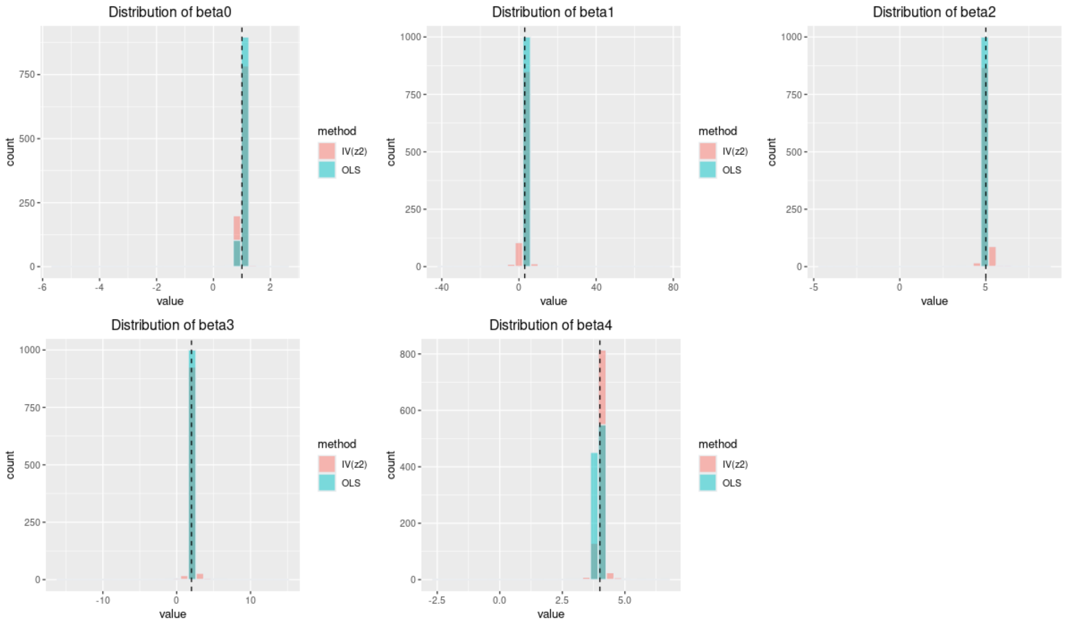

Figure 11: Case 2 - histograms with $N = 1000$, $n = 1000$ and z_2

6 Conclusions

The results of the previous section demonstrated the importance of using an instrumental variable and the strength of this. The different combinations of the hyperparameters provide insight into how the number of runs and observations influence the distribution of the calculated parameters. According to the simulation results, an increase in the number of runs, observations, and a strong instrument helps to reduce the variance of the distribution of the results, and the distribution of the results has a well-defined bell shape. In addition, the results of the second scenario show that despite using a large number of observations and runs if a strong instrument is not used it causes the parameter estimates to be biased. In the case of using a weak instrument concerning the previous scenario, it causes greater dispersion in the calculation of the parameters, and outliers appear in the distribution of the results. The results of the second scenario show some insights into how the endogeneity problem affects the other exogenous variables. When a strong instrument is not used to help deal with endogeneity the distributions of these parameters tend to move away from the true value of the coefficients. In conclusion, the increase in the number of observations allows us to have a bell-shaped distribution of results when using a strong instrument. But in the case where a weak instrument is used, this can lead to a bias in the distribution of the parameters. In addition, the standard deviation of the estimated parameters is affected by the strength of the instrument, since a weak instrument produces an increase in these values in relation to the cases in which a strong instrument is used.

References

[1] Angrist, J. D., Keueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? The Quarterly Journal of Economics, 106(4), 979–1014. doi:10.2307/2937954

[2] Fonseca, R., Michaud, P.-C., Zheng, Y. (2019). The effect of education on health: evidence from national compulsory schooling reforms. SERIEs, 11(1), 83–103. doi:10.1007/s13209-019-0201-0

[3] Angrist, J. D., Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. The Quarterly Journal of Economics, 114(2), 533–575. doi:10.1162/003355399556061

[4] Rouse, C. E. (1995). Democratization or Diversion? The Effect of Community Colleges on Educational Attainment. Journal of Business Economic Statistics, 13(2), 217–224. doi:10.1080/07350015.1995.10524596

[5] Puhani, Patrick A. and Weber, Andrea Maria and Weber, Andrea Maria, Does the Early Bird Catch the Worm? Instrumental Variable Estimates of Educational Effects of Age of School Entry in Germany (January 2006). U of St.Gallen Department of Economics Discussion Paper No. 2006-02, Available at SSRN: https://ssrn.com/abstract=880016