Instrumental Variable: A Remedy for Endogeneity

Minhwa Lee

1 Introduction

Regression is a powerful method of analysis that provides information about the impact of an explanatory variable on the outcome variable. With only a few lines of codes, users of statistical software packages can perform regression analysis relatively easily. Due to this convenience, however, the conditions under which regression analysis holds valid are oftentimes overlooked. The goal of this paper is to specifically address the endogeneity condition and introduce a statistical method that remedies the violation of this condition.

In order to communicate the usefulness of the method, we will perform the following tasks. First, we will delineate the underlying logic of Ordinary Least Squares (OLS), an estimator most commonly used in linear regression. Second, we will state the endogeneity condition and explain why it is a requirement for OLS to be an appropriate estimator. Third, we will describe how the instrumental variable method, also known as Two-Staged Least Squares (2SLS), can resolve simultaneity bias. Lastly, we will apply 2SLS to a data set and compare the results with that of OLS.

2 Ordinary Least Squares (OLS)

In this section, we define the key components of linear regression analysis. **Ordinary Least Squares (OLS)** is an estimator that enables one to derive a value that is close to the true value of a parameter. In regression analysis, the parameters of interest are the estimated coefficients of our explanatory variables in the following theoretical equation:

$$Y_i = \beta_0 + \beta_1 \, X_{1i} + \beta_2 \, X_{2i} + \cdots + \beta_K \, X_{Ki} + \epsilon_i \tag{1}$$

where K is the number of explanatory variables and i is the observation number. The **error term**, ϵ, accounts for the stochastic variation in Y by absorbing the randomness that cannot be encapsulated in our model. OLS is a function that minimizes the sum of squared residuals, where **residual**, e_i, refers to the difference between the actual and estimated values of Y, in other words $e_i = Y_i - \widehat{Y}_i$. Thus, the error term and the residuals are similar in that they both represent the variation in Y that cannot be explained by the X variables. The difference is that error term comes from the theoretical regression equation whereas residuals can actually be calculated using data.

3 The Endogeneity Condition for OLS

Among the seven conditions required for OLS to be the best estimator, we specifically pay heed to the **endogeneity condition**. OLS requires that there is no endogeneity in the model. Endogeneity refers to the presence of correlation between an explanatory variable and the error term. In Equation (1), X_j is said to be **endogenous** if the correlation between X_j and ϵ_j is non-zero and **exogenous** if the correlation is zero. If endogeneity is present in the model, OLS produces biased coefficient estimates, which means that the expected value of an estimated coefficient is not equal to the unknown true value of the coefficient. Put differently, endogeneity is problematic because the **bias** in our coefficient estimates becomes non-zero, or equivalently, $E(\widehat{\beta_j}) - \beta_j \neq 0$. This is not a desirable result since we want our estimates to be as close to the true value of the parameter as possible.

3.1 Simultaneous Models

The endogeneity problem commonly occurs in simultaneous models in which variables are jointly determined. To give an example, education level and income may have a two-way causal relationship. On one hand, those with a higher level of education may be more likely to find a higher-paying job. On the other hand, higher income may enable one to afford longer years of education without having to hurriedly get a job. Following Studenmund (2017), we present a generalized form of simultaneous equations that contain such jointly determined variables:

$$Y_{1t} = \alpha_0 + \alpha_1 Y_{2t} + \alpha_2 X_{1t} + \epsilon_{1t} \tag{2}$$

$$Y_{2t} = \beta_0 + \beta_1 Y_{1t} + \beta_2 X_{2t} + \epsilon_{2t}. \tag{3}$$

Equations (2) and (3) are **structural equations** that are formulated on the basis of theoretical speculation. Similar to the example of education level and income, Y_{1t} and Y_{2t} are jointly determined and are thus endogenous variables. Meanwhile, the X_1 and X_2 variables do not possess this simultaneous nature and are thus exogenous.

3.2 Endogeneity in Simultaneous Equations

Simultaneous equations with jointly determined variables violate the endogeneity condition. We can show this violation rather intuitively by the following train of thoughts. If ϵ_{1t} increases, then Y_{1t} increases in Equation (2); increased Y_{1t} leads to increased Y_{2t} in Equation (3); and due to the simultaneous nature, increased Y_{2t} also increases Y_{1t} in Equation (2). Thus, we see that an increase in ϵ_{1t} leads to an increase in Y_{2t}, which is an independent variable in Equation (2). Due to this correlation between the error term and an explanatory variable, we claim that the model expressed in Equation (2) violates the endogeneity condition of the OLS. Without loss of generality, the same conclusion can be made about Y_{1t} and ϵ_{2t}.

Endogeneity in simultaneous equations produces **simultaneity bias** in the coefficient estimates. Mathematically proving the existence of simultaneity bias is beyond the scope of this paper. Alternatively, we will show the presence of simultaneity bias by performing a computer simulation on a data set in Section 5.2.2.

4 Instrumental Variable: A remedy for endogeneity

The OLS estimator produces biased coefficient estimates when used in simultaneous models. To cope with this problem, the researcher may choose to use an alternative estimation procedure called **Two-Stage Least Squares (2SLS)**. As the name implies, this method is also based upon the idea of minimizing squared residuals. In fact, 2SLS performs the OLS procedure across two stages which involves the selection of an instrumental variable. An **instrumental variable** is one that is highly correlated with the endogenous variable and is uncorrelated with the error term.

4.1 Reduced-Form Equations

In Equation (2), Y_{2t} is an independent variable that is correlated with the error term, ϵ_{1t}. To reduce (or, ideally, eliminate) this correlation and resolve the endogeneity problem, we construct another regression equation to express Y_{2t} as a function of some instrumental variables that are uncorrelated with the error term, ϵ_{1t}. Doing the same for Equation (3), we establish a set of **reduced-form equations** that express our endogenous variables, Y_{1t} and Y_{2t}, in terms of the exogenous variables:

$$Y_{1t} = \gamma_0 + \gamma_1 X_{1t} + \gamma_2 X_{2t} + \gamma_3 X_{3t} + \gamma_4 X_{4t} + v_{1t} \tag{4}$$

$$Y_{2t} = \delta_0 + \delta_1 X_{1t} + \delta_2 X_{2t} + \delta_3 X_{3t} + \delta_4 X_{4t} + v_{2t}. \tag{5}$$

Note that we include *all* exogenous variables within the system into the reduced-form equations. This is because every exogenous variable in the simultaneous system is a "candidate" to be an instrumental variable. By choosing only one instrumental variable, we would be "throwing away" information (Studenmund, 2017). The γs and δs are the **reduced-form coefficients**, whereas vs are the error terms for the new regression equations. The reduced-form equations do not have inherent simultaneity, so we have successfully avoided the violation of the endogeneity condition. In particular, the variables X_3 and X_4 have been additionally selected as instrumental variables for Y_1 and Y_2, respectively. Now, we proceed to describe the two stages of 2SLS.

4.2 2SLS: Stage One

The first step of the 2SLS procedure is to estimate the reduced-form equations using OLS. We estimate Equations (4) and (5) and write:

$$\widehat{Y_{1t}} = \widehat{\gamma_0} + \widehat{\gamma_1} X_{1t} + \gamma_2 X_{2t} + \widehat{\gamma_3} X_{3t} + \widehat{\gamma_4} X_{4t} \tag{6}$$

$$\widehat{Y_{2t}} = \widehat{\delta_0} + \widehat{\delta_1} X_{1t} + \widehat{\delta_2} X_{2t} + \widehat{\delta_3} X_{3t} + \widehat{\delta_4} X_{4t}. \tag{7}$$

Ideally, the researcher would identify all instrumental variables so that endogeneity is completely eliminated. If this is the case, the X variables would be exogenous and uncorrelated with the error terms, vs. Also, the OLS estimates of the reduced-form coefficients, $\widehat{\delta}$s and $\widehat{\gamma}$s, would be unbiased.

4.3 2SLS: Stage Two

In order to perform the second-stage estimation, we must first replace the endogenous variables with the estimated reduced-form equations. That is, we substitute the Ys with the $\widehat{Y}$s in order to express the structural equations only in terms of the exogenous X variables:

$$Y_{1t} = \alpha_0 + \alpha_1 \widehat{Y_{2t}} + \alpha_2 X_{1t} + \epsilon_{1t} \tag{8}$$

$$Y_{2t} = \beta_0 + \beta_1 \widehat{Y_{1t}} + \beta_2 X_{2t} + \epsilon_{2t} \tag{9}$$

In essence, we are rewriting Equations (2) and (3) by plugging Equations (6) and (7) into the $\widehat{Y_{1t}}$ and $\widehat{Y_{2t}}$ terms. After algebraic simplification, we obtain:

$$Y_{1t} = \alpha_0' + \alpha_1' X_{1t} + \alpha_2' X_{2t} + \alpha_3' X_{3t} + \alpha_4' X_{4t} + \epsilon_{1t}' \tag{10}$$

$$Y_{2t} = \beta_0' + \beta_1' X_{1t} + \beta_2' X_{2t} + \beta_3' X_{3t} + \beta_4' X_{4t} + \epsilon_{2t}'. \tag{11}$$

Now, we perform the second-stage of 2SLS by estimating Equations (10) and (11). Recall that we estimated the first-stage equations with OLS. It must be noted that the second-stage equations are *not* estimated with OLS. Estimating the second-stage equations with OLS will produce incorrect standard errors in our coefficient estimates, $SE(\widehat{\alpha})$ and $SE(\widehat{\beta})$. Therefore, it is important that we use the computer's 2SLS procedure when performing the second stage. In Section 5.X, we will specifically demonstrate this discrepancy in standard errors by estimating the second stage "by hand" (in other words, with OLS) and then with a statistical package.

4.4 Properties of 2SLS

There are two properties to note about 2SLS estimation. First, 2SLS estimates are still biased. Simultaneity bias cannot be fully eliminated due to any remaining correlation between the $\widehat{Y}$s produced by the first-stage estimations and the ϵs. With larger sample size, the 2SLS bias will be reduced but will remain non-zero. However, the expected bias due to 2SLS will be smaller than the expected bias due to OLS. This is certainly an advantage that 2SLS has over OLS. The second property of 2SLS is that the coefficient estimates have increased variances and $SE(\widehat{\beta})$s compared

to OLS estimates (Wooldridge, 2016). These properties will be observed and discussed further in Section 5.2.

5 Application to Real-World Data Set

In the following sections, we will test a real-world data set using both OLS and 2SLS procedures. The data set is sourced from the 2009 annual report of the U.S. government, *The Economic Report of the President* (**original source to be put**). The data set contains information about seven economic variables, as described in Table 1 below, over the course of 33 years. All variables except for the interest rate, r_t, are adjusted for inflation and are measured in billions of 2,000 dollars. For the sake of simplicity, we will not consider the units of measurement in our interpretations of results. The following equations will be used to build a macroeconomic model of the U.S. economy (Studenmund, 2017):

$$Y_t = CO_t + I_t + G_t + NX_t \tag{12}$$

$$CO_t = \beta_0 + \beta_1 YD_t + \beta_2 CO_{t-1} + \epsilon_{1t} \tag{13}$$

$$YD_t = Y_t - T_t \tag{14}$$

$$I_t = \beta_3 + \beta_4 Y_t + \beta_5 r_{t-1} + \epsilon_{2t}. \tag{15}$$

Table 1: Variables List in the Four Equations (Studenmund, 2017)

Y_t	Gross Domestic Product (GDP) in year t
CO_t	Total personal consumption in year t
I_t	Total private domestic investment in year t
G_t	Government purchases of goods and services in year t
NX_t	Net exports of goods and services (exports minus imports) in year t
r_t	Interest rate in year t
YD_t	Disposable income in year t

Equations 12 and 14 do not have an error term because they each *define* the composition of Gross Domestic Product (GDP) and disposable income. On the other hand, Equations 13 and 15 have the stochastic error terms because they *describe* the relationship between variables. Following Studenmund (2017), we first identify the following variables as endogenous: Y_t, CO_t, YD_t, and I_t.

It is clear that if I_t increases, then Y_t increases as well (Equation 12), which also increases in I_t (through Equation 15) in the long run. We also observe that a change in YD_t results in a change in CO_t (Equation 13). Then, we see that CO_t changes Y_t as shown in Equation 12, which will again have an impact on YD_t (Equation 14). Having identified the endogenous variables, we claim that the remaining variables, G_t, NX_t, T_t, CO_{t-1}, and r_{t-1}, are exogenous.

We will now consider two research questions as we apply different estimation methods to the data set. First, what is the impact of GDP and interest rate on investment? Second, does 2SLS successfully reduce simultaneity bias and resolve the endogeneity concern, compared to OLS?

5.1 Estimation of the Investment Function

To answer our first research question, we will use both OLS and 2SLS procedures. In particular, we will carry out the 2SLS procedure both "by hand" and with computer.

5.1.1 OLS Estimation

Using OLS, we estimate the regression equation of the investment function, as shown in Equation 15:

$$I_t = \beta_3 + \beta_4 Y_t + \beta_5 r_{t-1} + \epsilon_{2t}.$$

We estimate this equation using the *lm* function in the statistical software R, which is commonly used for linear regression analysis. The estimated results can be summarized as:

$$\widehat{I_t} = -267.166 + 0.193 Y_t - 9.258 r_{t-1}. \tag{16}$$

5.1.2 2SLS Estimation "By Hand"

This time, we estimate the investment function using the 2SLS procedure. Recall that we defined Y_t as an endogenous variable. The first stage of 2SLS involves expressing the endogenous variables in terms of the exogenous ones. As a result, we obtain the reduced-form equations. This process is

Table 2: Regression Results Using OLS

	Dependent variable:
	I_t
Y_t	0.193^{***}
	(0.012)
r_{t-1}	-9.258
	(11.193)
Constant	-267.166
	(179.222)
Observations	32
R^2	0.958
Adjusted R^2	0.956

Note: *p<0.1; **p<0.05; ***p<0.01. Standard Error in Parentheses.

fully described in the following set of equations:

$$\begin{aligned} Y_t &= CO_t + I_t + G_t + NX_t \\ &= \beta_0 + \beta_1 YD_t + \beta_2 CO_{t-1} + \beta_3 + \beta_4 Y_t + \beta_5 r_{t-1} + G_t + NX_t + \epsilon' \\ &= \beta_0 + \beta_1 Y_t - \beta_1 T_t + \beta_2 CO_{t-1} + \beta_3 + \beta_4 Y_t + \beta_5 r_{t-1} + G_t + NX_t + \epsilon' \qquad (17) \\ (1 - \beta_1 - \beta_4) Y_t &= \beta_0 - \beta_1 T_t + \beta_2 CO_{t-1} + \beta_3 + \beta_5 r_{t-1} + G_t + NX_t + \epsilon' \\ \therefore Y_t &= \pi_0 + \pi_1 T_t + \pi_2 CO_{t-1} + \pi_5 r_{t-1} + \pi_6 G_t + \pi_7 NX_t + \epsilon'. \end{aligned}$$

As a result, we can use OLS to estimate Y_t and obtain the following equation:

$$\widehat{Y_t} = \pi_0 + \pi_1 T_t + \pi_2 CO_{t-1} + \pi_5 r_{t-1} + \pi_6 G_t + \pi_7 NX_t \qquad (18)$$

In Section 3, we noted that researchers may come up with an instrumental variables to resolve the endogeneity problem. In our example, however, we already have a set of equations that describe the theoretical relationship between variables. In a sense, economic theory *provided* us with the relevant instrumental variables. Nevertheless, the researcher may choose to consider additional instrumental variables to reduce endogeneity within her model.

Moving on to the second stage of 2SLS, we substitute the endogenous variable, Y_t, with the estimated reduced-form equation given in Equation 18. We then obtain the following equation as a revised version of our structural equation - namely, the revised investment function:

$$I_t = \beta_3 + \beta_4 \widehat{Y}_t + \beta_5 r_{t-1} + \epsilon_{2t}. \tag{19}$$

This time, we use the *fitted.values()* function in R to extract the estimated values, $\widehat{Y}_t$. We plug the $\widehat{Y}_t$ values into the original investment equation (Equation 15) and generate the revised investment equation (Equation 19). Table 3 presents the 2SLS estimation of the investment function, and the following equation indicates the linear model with the estimated coefficients by 2SLS:

$$\widehat{I}_t = -261.480 + 0.192\widehat{Y}_t - 9.550 r_{t-1}. \tag{20}$$

Table 3: Regression Results Using 2SLS "By Hand"

	Dependent variable:
	I_t
$\widehat{Y}_t$	0.192^{***}
	(0.012)
r_{t-1}	-9.550
	(11.491)
Constant	-261.480
	(184.038)
Observations	32
R^2	0.956
Adjusted R^2	0.953
Note:	*p<0.1; **p<0.05; ***p<0.01. Standard Error in Parentheses.

As indicated in Table 3, we have performed the 2SLS procedure "by hand". Recall from Section 4.3 that the standard errors in 2SLS are different from merely executing OLS twice. To note the difference, we will now utilize a function in R dedicated to correcting the standard error of the coefficient estimates.

5.1.3 2SLS Estimation with Computer

The *ivreg* function built in the *AER* package can be used to estimate an instrumental variable regression using R. To implement the *ivreg* function, we first define the following variables to be our instrumental variables for the investment function: T_t, CO_{t-1}, r_{t-1}, G_t, and NX_t. Table 4 displays an estimation of the investment function with 2SLS, not by hand but with computer.

Table 4: Regression Results Using 2SLS with Computer

	Dependent variable:
	I_t
Y_t	0.192^{***}
	(0.012)
r_{t-1}	-9.550
	(11.199)
Constant	-261.480
	(179.367)
Observations	32
R^2	0.958
Adjusted R^2	0.956
Note:	*p<0.1; **p<0.05; ***p<0.01. Standard Error in Parentheses.

So far, we have used three different methods to estimate the investment function. The results are summarized in Table 5. We clearly notice that the standard error of the constant term and the coefficient of r_{t-1} are reduced, compared to the results from 2SLS by hand. Thus, we confirm that the *ivreg* function has corrected the standard errors of coefficient estimates. Furthermore, we also observe that both 2SLS methods (by hand and with computer) show higher standard error compared to OLS estimates. This observation corresponds to the property of 2SLS mentioned in Section 4.4, which is that 2SLS estimates have increased variances and standard error.

Now, we directly answer our first research question. Our 2SLS estimation results indicate that a one-unit change in GDP will increase investment by 0.192 units, holding interest rates constant. In the meantime, a one-unit increase in the previous year's interest rate will decrease investment by 9.528 units, holding GDP constant. The coefficients make sense because investment will generally

Table 5: Comparison of Regression Results: OLS, 2SLS by hand, and 2SLS with computer

	Dependent variable: Investment		
		I_t	
	(OLS)	(2SLS by hand)	(2SLS by 'ivreg')
Y_t	0.193^{***}		0.192^{***}
	(0.012)		(0.012)
$\widehat{Y_t}$		0.192^{***}	
		(0.012)	
r_{t-1}	-9.258	-9.550	-9.550
	(11.193)	(11.491)	(11.199)
Constant	-267.166	-261.480	-261.480
	(179.222)	(184.038)	(179.367)
Observations	32	32	32
R^2	0.958	0.956	0.958
Adjusted R^2	0.956	0.953	0.956
Note:	*p<0.1; **p<0.05; ***p<0.01. Standard Error in Parentheses.		

increase as a country becomes wealthier. Also, investors will have a weaker incentive to take risks in their investment if securing money in the bank creates a high return. In addition, we note that the results for GDP are statistically significant at the 99% level, whereas interest rate does not show to be a statistically significant predictor.

Notice the difference in the coefficient estimates of the GDP variable, Y_t. The difference between the OLS and 2SLS estimations is only 0.001. However, this is a considerable amount of money if we consider the unit of observation. Holding all other variables constant, the OLS estimation of investment will always be 0.2 billion dollars greater than that of 2SLS for each unit of investment. Given these results, would it be reasonable to claim that 2SLS is a better estimation procedure than OLS? This leads us to consider our second research question.

5.2 Comparison of Simultaneity Bias

Does 2SLS successfully reduce simultaneity bias? To answer this question, we first need to identify the bias that OLS and 2SLS each produces in the estimated coefficients of the investment function.

By definition, $Bias = E(\widehat{\beta}) - \beta$, and we can use this to calculate the bias for each of the methods:

$$Bias_{OLS} = E(\widehat{\beta}_{OLS}) - \beta \text{ and } Bias_{2SLS} = E(\widehat{\beta}_{2SLS}) - \beta.$$

Comparison of bias requires the calculation of the expected value of coefficient estimates. Only then can we compare the two expected values and decide which method results in a greater bias. Following Wooldridge (2016), we consider the formula for the OLS coefficient estimate of a multiple regression equation:

$$\widehat{\beta}_1 = \frac{(\sum_{n=1}^{N} yx_1)(\sum_{n=1}^{N} {x_2}^2) - (\sum_{n=1}^{N} yx_2)(\sum_{n=1}^{N} x_1x_2)}{(\sum_{n=1}^{N} {x_1}^2)(\sum_{n=1}^{N} {x_2}^2) - (\sum_{n=1}^{N} x_1x_2)^2} \tag{21}$$

and the formula for the 2SLS coefficient estimate:

$$\widehat{\beta}_1 = \frac{\sum_{n=1}^{N}(z_n - \overline{z})(y_n - \overline{y})}{\sum_{n=1}^{N}(z_n - \overline{z})(x_n - \overline{x})}, \tag{22}$$

where z is an instrumental variable.

A slight difficulty arises in actually calculating the bias because we do not know the 'true' value of β. However, we are only interested in showing that the bias in the OLS estimate is greater than the 2SLS estimate. In other words, we want to show:

$$E(\widehat{\beta}_{2SLS}) - \beta < E(\widehat{\beta}_{OLS}) - \beta.$$

The inequality shows that knowledge of the true value of β is not necessary. In order to answer our second research question, all we need to show is that the following holds true:

$$E(\widehat{\beta}_{2SLS}) < E(\widehat{\beta}_{OLS}).$$

5.2.1 Methodology

We choose computer simulation as our method of analysis. Using Equation 21, we will obtain the OLS estimate, $\widehat{\beta}_1$, for Y_t, which is the endogenous variable in the investment function. To be consistent with the above formula, we consider I_t as our y, Y_t as x_1, and r_{t-1} as x_2. Then, we generate

100 samples of each x_1, x_2, and y, respectively; substitute the corresponding values into Equation 21; and thus obtain a sampling distribution for the 100 values of $\widehat{\beta}_1$. In a similar manner, we will also investigate the 2SLS coefficient estimate. Equation 22 requires an instrumental variable, z, and we have already established $r_{t-1}$to be an instrumental variable for Y_t. Therefore, Y_t will be our x in the formula, I_t our y, and r_{t-1}our z. Then, we generate 100 samples of x, y, and z, respectively, and follow the same procedure described above. By substituting the three values into Equation 22, we obtain a sampling distribution of our $\widehat{\beta}_1$s.

5.2.2 Results and Discussions

In our computational results, we find that $E(\widehat{\beta}_{OLS}) = 0.06463382$ and $E(\widehat{\beta}_{2SLS}) = 0.05448795$. This result is indeed consistent with the property of 2SLS estimation that we discussed in Section 4.4. Also, we have computationally shown that $E(\widehat{\beta}_{OLS}) > E(\widehat{\beta}_{2SLS})$ and thus answered our second research question of whether 2SLS successfully reduces simultaneity bias.

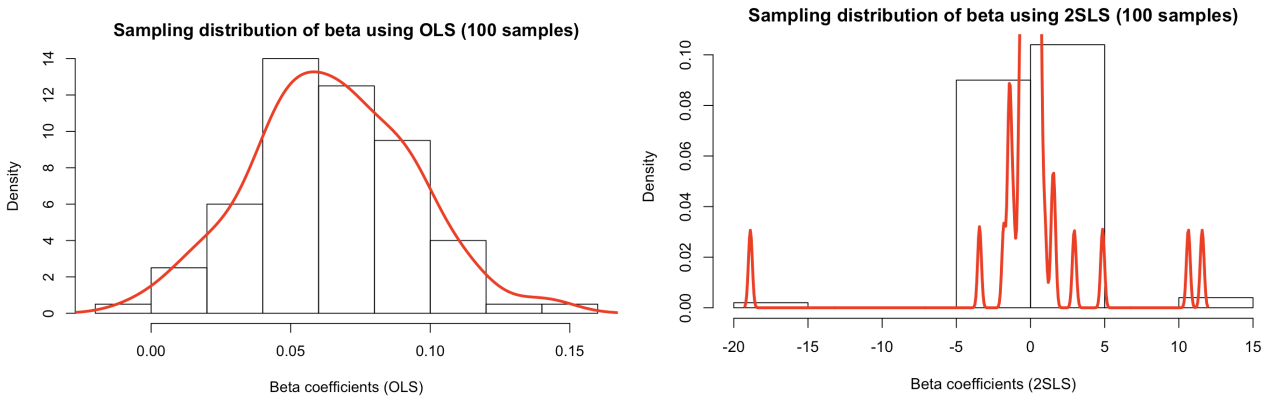

Figure 1: Sampling distributions of $\widehat{\beta}_1$ using OLS and 2SLS

Figure 1 provides a visual overview of the variance and simultaneity bias in the estimation procedures. The sampling distribution of $\widehat{\beta}_{OLS}$ is smoother than that of $\widehat{\beta}_{2SLS}$ with less variation at the tails. This verifies the shortcoming of 2SLS, which is that the coefficient estimates have higher standard errors and thus increased variances. This may be offset by the advantage of 2SLS, which is that simultaneity bias is reduced.

6 Conclusion

In this paper, we have introduced instrumental variable regression as a solution to the endogeneity problem. We have carefully examined the two stages of the 2SLS estimation procedure and highlighted its properties in comparison to OLS estimation. With a theoretical understanding of the 2SLS method, we estimated an instrumental variable regression against a real-world data set. In this exercise, we estimated annual investment in the United States as a function of GDP and interest rates (of the previous year). To resolve the endogeneity problem, we expressed GDP in terms of four instrumental variables, which were net export, consumption (of the previous year), government spending, and tax revenue.

In our empirical analysis, we aimed to answer two research questions. First, we identified the impact of each explanatory variable on annual investment by using OLS, 2SLS "by hand", and 2SLS with computer. We presented interpretations of the coefficient estimates as well as a comparison between the three methods. Second, we showed via simulation that 2SLS reduces simultaneity bias. We used the formulae for the beta coefficients in OLS and in 2SLS, generated a sampling distribution of the coefficient estimates, and then compared the expected value of the two.

There may not be a 'perfect' tool capable of completely eliminating endogeneity or simultaneity bias. However, the researcher must diligently aim to devise or utilize methods that enhance the validity of her statistical analysis. We have shown that instrumental variable and 2SLS estimation can be an effective method when analyzing simultaneous models.

References

[1] Chair of the Council of Economic Advisers. (2009). *The Economic Report of the President, 2009*. Washington D.C.

[2] Studenmund, A.H. (2017). *Using Econometrics: A Practical Guide (7th ed.)*. Boston, MA: Pearson.

[3] Wooldrige, J. M. (2016). *Introductory Econometrics: A Modern Approach (6th ed.)*. Boston, MA: Cengage Learning.

7 Appendix A: R Codes

```
setwd("~/Desktop")
library(readxl)
library(stargazer)
library(AER)
stat <- read_excel("stat.xlsx") # Import Data set
```

```r
#-------Compare 2SLS with OLS --------------

stat$r_t[1] <- NA
for (i in 2:33){
  stat$r_t[i] <- stat$r[i-1] # r_{t-1}
}

stat$CO_t[1] <- NA
for (i in 2:33){
  stat$CO_t[i] <- stat$CO[i-1] # CO_{t-1}
}

invfunc <- lm(I~Y+r_t, data = stat) # (a) Investment function with OLS
summary(invfunc)

stat$nx <- stat$Y - stat$CO - stat$I - stat$G
stat$t <- stat$Y - stat$YD

y <- lm(Y~t+CO_t+r_t+G+nx, data = stat) # (b) Reduced Y with OLS
summary(y)

y.hat1 <- fitted(y)
y.hat <- fitted.values(y) # (c)
```

```r
y.hat1 <- fitted(y)
y.hat <- fitted.values(y) # (c)


stat$y.hat[1] <- NA
for (i in 2:33){
  stat$y.hat[i] <- y.hat[i-1]
}

invfunc2 <- lm(I~y.hat+r_t, data = stat)
summary(invfunc2)

invfunc3 <- ivreg(I~Y+r_t|t+CO_t+r_t+G+nx, data = stat)
summary(invfunc3)

stargazer(invfunc, invfunc2, invfunc3)
```

```r
#---------BIAS - OLS vs SLS---------------------------#

b_sls <- c()
for (i in 1:100){
rs_x <- sample(stat$Y[!is.na(stat$Y)]); rs_x
rs_y <- sample(stat$I[!is.na(stat$I)]); rs_y
rs_z <- sample(stat$r_t[!is.na(stat$r_t)]); rs_z

num <- sum((rs_z - mean(rs_z, na.rm = TRUE)) * (rs_y - mean(rs_y, na.rm = TRUE)))
denom <- sum((rs_z - mean(rs_z, na.rm = TRUE)) * (rs_x - mean(rs_x, na.rm = TRUE)))

beta <- num/denom
b_sls <- c(b_sls, beta)
}

b_sls
```

```r
b_ols <- c()
for (i in 1:100){
rs_x1 <- sample(stat$Y[!is.na(stat$Y)]); rs_x1
rs_x2 <- sample(stat$r_t[!is.na(stat$r_t)]); rs_x2
rs_y <- sample(stat$I[!is.na(stat$I)]); rs_y


num <- sum(rs_x2^2)*sum(rs_x1*rs_y) - (sum(rs_x1*rs_x2))*(sum(rs_x2*rs_y))
denom <- sum(rs_x1^2)*sum(rs_x2^2) - (sum(rs_x1*rs_x2))^2

beta_ols <- num/denom
b_ols <- c(b_ols, beta_ols)
}

#b_ols

#hist(b_ols)

c(mean(b_sls),mean(b_ols))
```

```r
hist(b_sls,xlab="Beta coefficients (2SLS)",
     main="Sampling distribution of beta using 2SLS (100 samples)",
     prob = TRUE)
lines(density(b_sls), col = "red", lwd = "3")
```

```r
hist(b_ols, xlab="Beta coefficients (OLS)",
     main="Sampling distribution of beta using OLS (100 samples)",
     prob = TRUE )
lines(density(b_ols), col = "red", lwd = "3")
```