Chapter 20: Omitted Variables and the Instrumental Variable Estimation Procedure

Chapter 20 Outline

- **Revisit Omitted Explanatory Variable Bias**
  - Review of Our Previous Explanation of Omitted Explanatory Variable Bias
  - Omitted Explanatory Variable Bias and the Explanatory Variable/Error Term Independence Premise
- **The Ordinary Least Squares Estimation Procedure, Omitted Explanatory Variable Bias, and Consistency**
- **Instrumental Variable Estimation Procedure: A Two Regression Estimation Procedure**
  - Mechanics
  - The "Good" Instrument Conditions
- **Omitted Explanatory Variables Example: 2008 Presidential Election**
  - Might the Ordinary Least Squares (OLS) Estimation Procedure Suffer from a Serious Econometric Problem?
- **Instrument Variable (IV) Application: 2008 Presidential Election**
  - The Mechanics
  - "Good" Instrument Conditions Revisited
- **Justifying the Instrumental Variable (IV) Estimation Procedure**

Prep Question

1. Consider two regression models:

$$y_t = \beta_{Const} + \beta_{x1} x1_t + \beta_{x2} x2_t + e_t$$

and

$$y_t = \beta_{Const} + \beta_{x1} x1_t + \varepsilon_t$$

a. Express the second model's error term, ε_t, in terms as a function of the first model's terms.

Assume that

- the coefficient β_{x2} is positive

and

- the explanatory variables, $x1_t$ and $x2_t$, are positively correlated.

b. Will the explanatory variable, $x1_t$, and the second model's error term, ε_t, be correlated? If so, how?

c. Focus on the second model. Suppose that the ordinary least squares (OLS) estimation procedure were used to estimate the parameters of

the second model. Would the ordinary least squares estimation (OLS) estimation procedure for the value of β_{x1} be biased? If so, how?

Omitted Explanatory Variables Example: 2008 Presidential Election
2008 Presidential Election Data: Cross section data of election, population, and economic statistics from the 50 states and the District of Columbia in 2008.

$AdvDeg_t$	Percent adults who have advanced degrees in state *t*
$Coll_t$	Percent adults who graduated from college in state *t*
HS_t	Percent adults who graduated from high school in state *t*
$PopDen_t$	Population density of state *t* (persons per square mile)
$RealGdpGrowth_t$	GDP growth rate for state *t* in 2008 (percent)
$UnemTrend_t$	Change in the unemployment rate for state *t* in 2008 (percent)
$VoteDemPartyTwo_t$	Percent of the vote received in 2008 received by the Democratic party in state *t* based on the two major parties (percent)

[Link to MIT-PresElectionByState-2008.wf1 goes here.]

2. Consider the following model explaining the vote received by the Democratic Party in the 2008 Presidential election:

$$\begin{aligned} VoteDemPartyTwo_t &= \beta_{Const} + \beta_{PopDen}PopDen_t + \beta_{Lib}Liberal_t + e_t \\ &= \beta_{Const} + \beta_{PopDen}PopDen_t + (\beta_{Lib}Liberal_t + e_t) \\ &= \beta_{Const} + \beta_{PopDen}PopDen_t + \varepsilon_t \end{aligned}$$

The variable $Liberal_t$ reflects the "liberalness" of the electorate in state *t*. If the electorate is by nature liberal in state *t*, the $Liberal_t$ would be high; on the other hand, if the electorate is conservative the $Liberal_t$ would be low.

a. Express the second model's error term, ε_t, in terms as a function of the first model's terms.

Assume that

- the coefficient β_{Lib} is positive

and

- the explanatory variables $PopDen_t$ and $Liberal_t$ are positively correlated

b. Will the explanatory variable $PopDen_t$ and the second model's error term, ε_t, be correlated? If so, how?

c. Focus on the second model. Suppose that the ordinary least squares (OLS) estimation procedure were used to estimate the parameters of the second model. Would the ordinary least squares estimation (OLS) estimation procedure for the value of β_{Lib} be biased? If so, how?

3. What does the correlation coefficient of $PopDen_t$ and $Coll_t$ equal?

Revisit Omitted Explanatory Variable Bias

We shall briefly review our previous discussion of omitted explanatory variables that appears in Chapter 14. Then, we shall show that omitted explanatory variable phenomenon can also be analyzed in terms of explanatory variable/error term correlation.

Review of Our Previous Explanation of Omitted Explanatory Variable Bias

In Chapter 14, we argued that omitting an explanatory variable from a regression will bias the ordinary least squares (OLS) estimation procedure for the coefficient value whenever two conditions are met. Bias results if the omitted variable:

- influences the dependent variable;
- is correlated with an included variable.

When these two conditions are met, the ordinary least squares (OLS) procedure to estimate the coefficient of the included explanatory variable captures two effects:

- **Direct Effect:** The effect that the included explanatory variable actually has on the dependent variable.
- **Proxy Effect:** The effect that the omitted explanatory variable has on the dependent variable because the included variable is acting as a proxy for the omitted variable.

Recall the goal of multiple regression analysis:

- **Goal of Multiple Regression Analysis:** Multiple regression analysis attempts to sort out the individual effect that each explanatory variable has on the dependent variable.

Consequently, we want the coefficient estimate of the included variable to capture only the direct effect and not the proxy effect. Unfortunately, this does not occur when the omitted variance influences the dependent variable and when it is also correlated with an included variable.

To illustrate this, we considered a model with two explanatory variables, *x1* and *x2*:

Model: $y_t = \beta_{Const} + \beta_{x1} x1_t + \beta_{x2} x2_t + e_t$

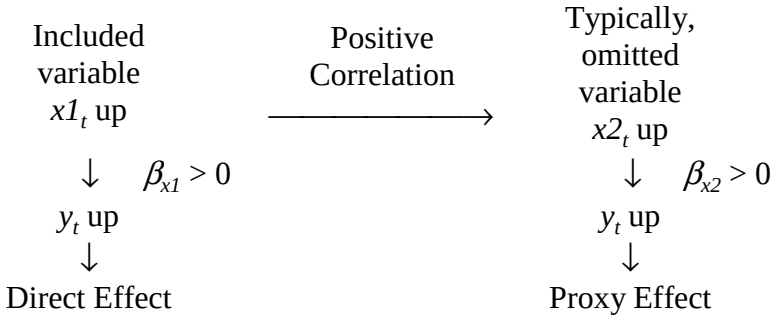
For purposes of illustration, assume that the coefficients are positive and that the explanatory variables are positively correlated:

- $\beta_{x1} > 0$ and $\beta_{x2} > 0$.
- $x1_t$ and $x2_t$ are positively correlated.

What happens when we omit the explanatory variable $x2_t$ from the regression? The two conditions necessary for the omitted variable bias are satisfied:

- Since β_{x2} is positive, the omitted variable influences the dependent variable.
- Since $x1_t$ and $x2_t$ are positively correlated, the omitted variable is correlated with an included variable.

An increase in $x1_t$ directly affects y_t, causing y_t to increase. But the story does not end here. Since the two explanatory variables are positively correlated, an increase in $x1_t$ is typically accompanied by an increase in $x2_t$ which in turn leads to an additional increase in y_t:



When the explanatory variable $x2_t$ is omitted from a regression, the ordinary least squares (OLS) estimation procedure for the value of $x1_t$'s coefficient, β_{x1}, is biased upward because it reflects not only the impact of $x1_t$ itself (direct effect) but also the impact of $x2_t$ (proxy effect).

Omitted Explanatory Variable Bias and the Explanatory Variable/Error Term Independence Premise

We can also use what we just learned about correlation between the explanatory variable and error term to explain why bias occurs. When we omit the explanatory variable $x2_t$ from the regression, the error term of the new equation, ε_t, includes not only the original error term, e_t, but also the "omitted variable term," $\beta_{x2}\, x2_t$:

$$\begin{aligned} y_t &= \beta_{Const} + \beta_{x1}x1_t + \beta_{x2}\, x2_t + e_t \\ &= \beta_{Const} + \beta_{x1}x1_t + (\beta_{x2}x2_t + e_t) \\ &= \beta_{Const} + \beta_{x1}x1_t + \varepsilon_t \qquad \text{where } \varepsilon_t = \beta_{x2}x2_t + e_t \end{aligned}$$

Recall that

- $\beta_{x2} > 0$.
- $x1_t$ and $x2_t$ are positively correlated.

The new error term, ε_t, includes the "omitted variable term," $\beta_{x2}x2_t$. Therefore, the included explanatory variable, $x1_t$, and the new error term, ε_t, are positively correlated:

- Since $x1_t$ and $x2_t$ are positively correlated, when $x1_t$ increases, $x2_t$ typically increases also.
- Since β_{x2} is positive when $x1_t$ increases, $\beta_{x2}x2_t$ and the new error term, ε_t, increases also. the new error term will typically increase.

$x1_t$ and $x2_t$ positively correlated

Included variable $x1_t$ up ⟶ Typically, omitted variable $x2_t$ up

$x1_t$ up ↘ ↙ ε_t up $\beta_{x2} > 0$

$\varepsilon_t = \beta_{x2}x2_t + e_t$

$x1_t$ and ε_t positively correlated

What did we learn about the consequence of correlation between the explanatory variable and error term? When the explanatory variable and error term are positively correlated, the ordinary least squares (OLS) estimation procedure for the value of $x1_t$'s coefficient is biased upward.

$x1_t$ and ε_t positively correlated

↓

OLS estimation procedure for the value of *x1*'s coefficient is biased upward

When $x2_t$ is omitted, $x1_t$ becomes a "problem" explanatory variable because it is correlated with the new error term. Our two analyzes arrive at the same conclusion.

The Ordinary Least Squares Estimation Procedure, Omitted Explanatory Variable Bias, and Consistency

When an omitted explanatory variable causes the ordinary least squares (OLS) estimation procedure to be biased, might the procedure still be consistent? We shall use a simulation to address this question.

Econometrics Lab 20.1: Ordinary Least Squares, Omitted Variables, and Consistency

[Link to MIT-Lab 20.1 goes here.]

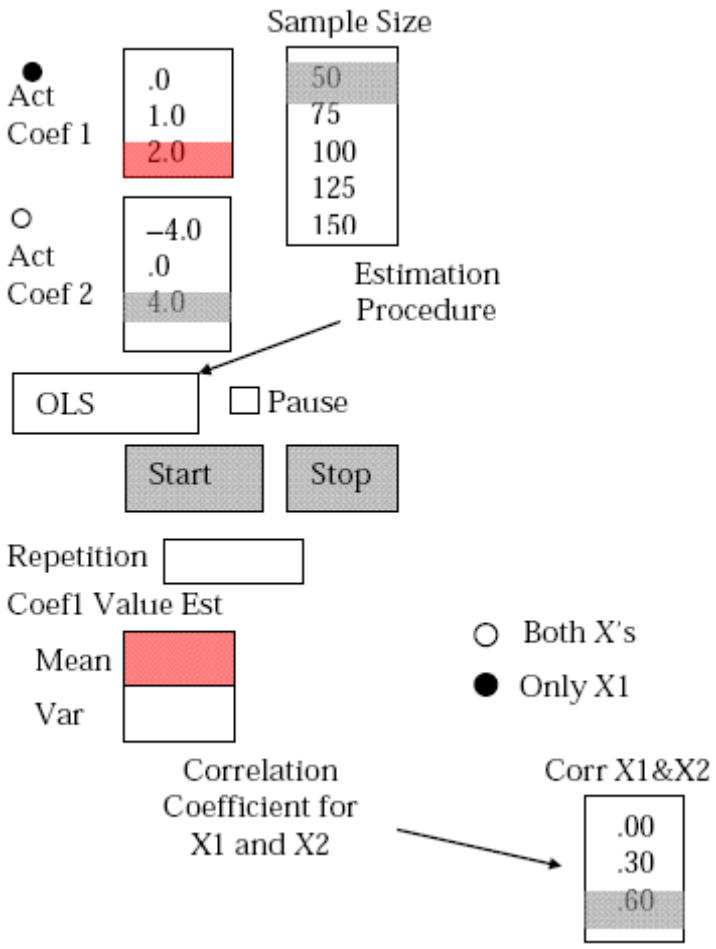


Figure 20.1: *OLS Omitted Variable Simulation*

By default the actual value of *x1* coefficient, Coef1, equals 2.0 and the actual value of the *x2* coefficient, Coef2, equals 4.0. The correlation coefficient for the explanatory variables *x1* and *x2* equals .60; they are positively correlated; the explanatory variables are positively correlated. Furthermore, the Only X1 option is selected; the explanatory variable *x2* is omitted. The included explanatory variable, *x1*, will be positively correlated with the error term. *x1* becomes a "problem" explanatory variable.

Initially, the sample size equals 50. Click Start and then after many, many repetitions click Stop. The mean of the coefficient estimates for the explanatory variable *x1* equals 4.4. Our logic is confirmed; upward bias results. Nevertheless, to determine if the ordinary least squares (OLS) estimation procedure might consistent we increase the sample size from 50 to 100 and once more from 100 to 150. As Table 20.1 reports, the mean of the coefficient estimates remains at 4.4.

Estimation Procedure	**Corr X1&X2**	**Actual Coef2**	**Sample Size**	**Actual Coef**	**Mean of Coef1 Ests**	**Magnitude of Bias**	**Variance of Coef1 Ests**
OLS	.60	4.0	50	2.0	≈4.4	≈2.4	≈6.6
OLS	.60	4.0	100	2.0	≈4.4	≈2.4	≈3.2
OLS	.60	4.0	150	2.0	≈4.4	≈2.4	≈2.1

Table 20.1: *Ordinary Least Squares: Bias and Consistency*

Unfortunately, the ordinary least squares (OLS) estimation procedure proves not only biased, but also not consistent whenever an explanatory variable is omitted that

- affects the dependent variable

and

- is correlated with an included variable.

What can we do?

Instrumental Variable Estimation Procedure: a Two Regression Estimation Procedure

The instrumental variable (IV) estimation procedure can deal with situations when the explanatory variable and the error term are correlated:

Original Model:

$y_t = \beta_{Const} + \beta_x x_t + \varepsilon_t$ where y_t = Dependent variable

x_t = Explanatory variable

ε_t = Error term

$t = 1, 2, \ldots, T$ T = Sample size

When x_t and ε_t are correlated

↓

x_t is the "problem" explanatory variable

Figure 20.2: *"Problem" Explanatory Variable*

When an explanatory variable, x_t, is correlated with the error term, ε_t, we refer to the explanatory variable as the "problem" explanatory variable. The correlation of the explanatory variable and the error term creates the bias problem for the ordinary least squares (OLS) estimation procedure. The instrumental variable estimation procedure can mitigate, but not completely remedy these cases. Let us briefly review the procedure and motivate it.

Mechanics

- **Choose a "Good" Instrument:** A "good" instrument, z_t, must have two properties:
  - Correlated with the "problem" explanatory variable, x_t.
  - Uncorrelated with the error term, ε_t.
- **Instrumental Variables (IV) Regression 1:** Use the instrument, z_t, to provide an "estimate" of the problem explanatory variable, x_t.
  - Dependent variable: "Problem" explanatory variable, x_t.
  - Explanatory variable: Instrument, z_t.
  - Estimate of the "problem" explanatory variable: $Estx_t = a_{Const} + a_z z_t$ where a_{Const} and a_z are the estimates of the constant and coefficient in this regression, IV Regression 1.

- **Instrumental Variables (IV) Regression 2:** In the original model, replace the "problem" explanatory variable, x_t, with its surrogate, $Estx_t$, the estimate of the "problem" explanatory variable provided by the instrument, z_t, from IV Regression 1.
    - Dependent variable: Original dependent variable, y_t.
    - Explanatory variable: Estimate of the "problem" explanatory variable based on the results from IV Regression 1, $Estx_t$.

The "Good" Instrument Conditions

Let us again provide the intuition behind why a "good" instrument, z_t, must satisfy the two conditions:

- **Instrument/"Problem" Explanatory Variable Correlation:** The instrument, z_t, must be correlated with the "problem" explanatory variable, x_t. To understand why, focus on IV Regression 1. We are using the instrument to create a surrogate for the "problem" explanatory variable in IV Regression 1:

  $$Estx_t = a_{Const} + a_z z_t$$

  The estimate, $Estx_t$, will be a good surrogate only if it is a good predictor of the "problem" explanatory variable, x_t. This will occur only if the instrument, z_t, is correlated with the "problem" explanatory variable, x_t.
- **Instrument/Error Term Independence:** The instrument, z_t, must be independent of the error term, ε_t. Focus on IV Regression 2. We begin with the original model and then replace the "problem" explanatory, x_t, variable with its surrogate, $Estx_t$:

  $$y_t = \beta_{Const} + \beta_x x_t + \varepsilon_t$$

  $$\downarrow$$

  Replace "problem" with surrogate

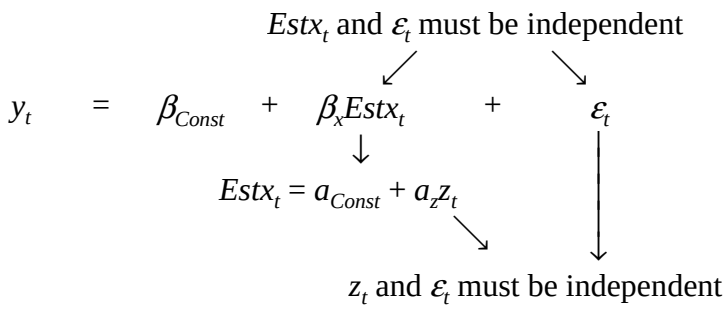
  $$= \beta_{Const} + \beta_x Estx_t + \varepsilon_t$$

  where $Estx_t = a_{Const} + a_z z_t$ from IV Regression 1

  To avoid violating the explanatory variable/error term independence premise in IV Regression 2, the surrogate for the "problem" explanatory variable, $Estx_t$, must be independent of the error term, ε_t. The surrogate, $Estx_t$, is derived from the instrument, z_t, in IV Regression 1:

  $$Estx_t = a_{Const} + a_z z_t$$

  Consequently, to avoid violating the explanatory variable/error term independence premise the instrument, z_t, and the error term, ε_t, must be independent.



Omitted Explanatory Variables Example: 2008 Presidential Election
2008 Presidential Election Data: Cross section data of election, population, and economic statistics from the 50 states and the District of Columbia in 2008.

$AdvDeg_t$	Percent of adults who have advanced degrees in state *t*
$Coll_t$	Percent of adults who graduated from college in state *t*
HS_t	Percent adults who graduated from high school in state *t*
$PopDen_t$	Population density of state *t* (persons per square mile)
$RealGdpGrowth_t$	GDP growth rate for state *t* in 2008 (percent)
$UnemTrend_t$	Change in the unemployment rate for state *t* in 2008 (percent)
$VoteDemPartyTwo_t$	Percent of the vote received in 2008 by the Democratic party in state *t* based on the two major parties (percent)

[Link to MIT-PresElectionByState-2008.wf1 goes here.]

Now, we introduce a model to explain the Democratic vote:

Model: $VoteDemPartyTwo_t = \beta_{Const} + \beta_{PopDen} PopDen_t + \beta_{Lib} Liberal_t + e_t$

The variable $Liberal_t$ reflects the "liberalness" of the electorate in state *t*. If the electorate is by nature liberal in state *t*, the $Liberal_t$ would be high; on the other hand, if the electorate is conservative the $Liberal_t$ would be low. The theories described below suggest that coefficients both $Liberal_t$ and $PopDen_t$would be positive:

- **Population Density Theory:** States with high population densities have large urban areas which are more likely to vote for the Democratic candidate, Obama; hence, $\beta_{PopDen} > 0$.
- **"Liberalness" Theory:** Since the Democratic party is more liberal than the Republican party, a high "liberalness" value would increase the vote of the Democratic candidate, Obama; hence, $\beta_{Lib} > 0$.

Unfortunately, we do not have any data to quantify the "liberalness" of a state; according, *Liberal* must be omitted from the regression.

Ordinary Least Squares (OLS)

Dependent Variable: *VoteDemPartyTwo*

Explanatory Variable(s):	Estimate	SE	*t*-Statistic	Prob
EstPopDen	0.004915	0.000957	5.137338	0.0000
Const	50.35621	1.334141	37.74431	0.0000

Number of Observations 51

Estimated Equation: *VoteDemPartyTwo* = 50.3 + .005*EstPopDen*

Interpretation of Estimates:

$b_{EstPopDen}$ = .005: A 1 person increase in a state's population density increases the state's Democratic vote by .005 percentage points; that is, a 100 person increase in a state's population density increases the state's Democratic vote by .5 percentage points.

Critical Result: The *EstPopDen* coefficient estimate equals .005. The positive sign of the coefficient estimate suggests that increases in a state with a higher population density will have a greater Democratic vote. This evidence supports the theory.

Table 20.2: *Democratic Vote OLS Regression Results*

Might the Ordinary Least Squares (OLS) Estimation Procedure Suffer from a Serious Econometric Problem?

Since the "liberalness" variable must be omitted from the regression, the explanatory variable/error term premise would be violated if the included variable, $PopDen_t$, is correlated with the new error term, ε_t.

$$\begin{aligned} VoteDemPartyTwo_t &= \beta_{Const} + \beta_{PopDen}PopDen_t + \beta_{Lib}Liberal_t + e_t \\ &= \beta_{Const} + \beta_{PopDen}PopDen_t + (\beta_{Lib}Liberal_t + e_t) \\ &= \beta_{Const} + \beta_{PopDen}PopDen_t + \varepsilon_t \end{aligned}$$

where $\varepsilon_t = \beta_{Lib}Liberal_t + e_t$

We have good reason to believe that they will be correlated because we would expect $PopDen_t$ and $Liberal_t$ to be correlated. States that tend to elect liberal representatives and senators then to have high population densities. That is, we suspect that $PopDen_t$ and $Liberal_t$ are positively correlated:

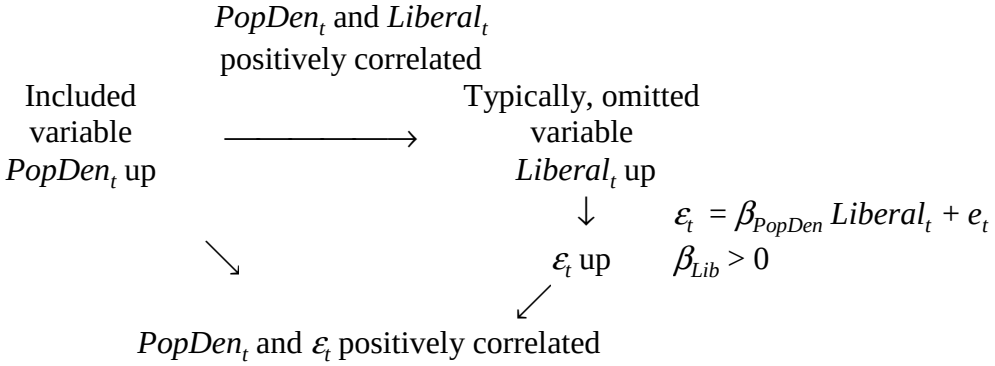


Figure 20.3: *PopDen – A "Problem" Explanatory Variable*

Consequently, the included explanatory variable, $PopDen_t$, and the error term, ε_t, will be positively correlated. The ordinary least squares (OLS) estimation procedure for the value of the coefficient will be biased upwards

To summarize, when the explanatory variable *Liberal* is omitted, as it must be, *PopDen* becomes a "problem" explanatory variable become it is correlated with the error term, ε_t. Now we shall apply the instrumental variable (IV) estimation procedure to understand how the instrument variable estimation procedure can address the omitted variable problem.

Instrument Variable (IV) Application: 2008 Presidential Election

The Mechanics

Choose an Instrument: In this example, we shall use the percent of high school graduates, $Coll_t$, as our instrument. In doing so, we believe that it satisfies the two "good" instrument conditions; that is, we believe that the percent of high school graduates, $Coll_t$, is

- positively correlated with the "problem" explanatory variable, $PopDen_t$

and

- uncorrelated with the error term, $\varepsilon_t = \beta_{Lib} Liberal_t + e_t$. Consequently, we believe that the instrument, $Coll_t$, is uncorrelated with the omitted variable, $Liberal_t$.

Instrumental Variables (IV) Regression 1

- **Dependent variable:** "Problem" explanatory variable, *PopDen*.
- **Explanatory variable:** Instrument, *Coll*.

Ordinary Least Squares (OLS)

Dependent Variable: *PopDen*

Explanatory Variable(s):	Estimate	SE	*t*-Statistic	Prob
Coll	149.3375	28.21513	5.292816	0.0000
Const	–3676.281	781.0829	–4.706647	0.0000
Number of Observations	51			

Estimated Equation: *EstPopDen* = –3,676.3 + 149.3*Coll*

Table 20.3: *Democratic Vote IV Regression 1 Results*

Instrumental Variables (IV) Regression 2:

- **Dependent variable:** Original dependent variable, *VoteDemPartyTwo*.
- **Explanatory variable:** Estimate of the "problem" explanatory variable based on the results from IV Regression 1, *EstPopDen*.

Ordinary Least Squares (OLS)

Dependent Variable: *VoteDemPartyTwo*

Explanatory Variable(s):	Estimate	SE	*t*-Statistic	Prob
EstPopDen	0.009955	0.001360	7.317152	0.0000
Const	48.46247	1.214643	39.89852	0.0000

Number of Observations 51

Estimated Equation: *VoteDemPartyTwo* = 48.5 + .001*EstPopDen*

Interpretation of Estimates:

$b_{EstPopDen}$ = .001: A 1 person increase in a state's population density increases the state's Democratic vote by .001 percentage points; that is, a 100 person increase in a state's population density increases the state's Democratic vote by .1 percentage points.

Critical Result: The *EstPopDen* coefficient estimate equals .001. The positive sign of the coefficient estimate suggests that increases in a state with a higher population density will have a greater Democratic vote. This evidence supports the theory.

Table 20.4: *Democratic Vote IV Regression 2 Results*

Good Instrument Conditions Revisited

IV Regression 1 allows us to assess the first "good" instrument condition.

- **Instrument/"Problem" Explanatory Variable Correlation:** The instrument, $Coll_t$, must be correlated with the "problem" explanatory variable, $PopDen_t$. We are using the instrument to create a surrogate for the "problem" explanatory variable in IV Regression 1:

  EstPopDen = –3,676.3 + 149.3*Coll*

  The estimate, $EstPopDen_t$, will be a "good" surrogate only if the instrument, $Coll_t$, is correlated with the "problem" explanatory variable, $PopDen_t$; that is, only if the estimate is a good predictor of the "problem" explanatory variable.

Correlation Matrix

	Coll	*PopDen*
Coll	1.000000	0.603118
PopDen	0.603118	1.000000

Table 20.5: *Correlation Matrix Coll and PopDen*

Table 20.5 reports that the correlation coefficient for $Coll_t$ and $PopDen_t$ equals .51. Furthermore, the IV Regression 1 results appearing in Table 20.4 suggest that the instrument, $Coll_t$, will be a good predictor of the "problem" explanatory variable, $PopDen_t$. Clearly, the coefficient estimate is significant at the 1 percent level. So, it is reasonable to judge that the instrument meets the first condition.

Next, focus on the second "good" instrument condition:

- **Instrument/Error Term Independence:** The instrument, $Coll_t$, and the error term, ε_t, must be independent. Otherwise, the explanatory variable/error term independence premise would be violated in IV Regression 2.

Recall the model that IV Regression 2 estimates:

$$VoteDemPartyTwo_t = \beta_{Const} + \beta_{PopDen} EstPopDen_t + \varepsilon_t$$

Question: Are $EstPopDen_t$ and ε_t independent?

$$EstPopDen_t = -3{,}676.3 + 149.3 Coll_t \qquad \varepsilon_t = \beta_{Lib} Liberal_t + e_t$$

Answer: Only if $Coll_t$ and $Liberal_t$ are independent.

The explanatory variable/error term independence premise will be satisfied only if the surrogate, $EstPopDen_t$, and the error term, ε_t, are independent. $EstPopDen_t$ is a linear function of $Coll_t$ and ε_t is a linear function of $PopDen_t$:

- $EstPopDen_t = -3{,}676.3 + 149.3 Coll_t$

and

- $\varepsilon_t = \beta_{Lib} Liberal_t + e_t$

Hence, the explanatory variable/error term independence premise will be satisfied only if the instrument, $Coll_t$, and the omitted variable, $Liberal_t$, are independent. If they are correlated, then we have gone "from the frying pan into the fire." It was the violation of this premise that created the problem in the first place. Unless we were to believe that liberals are better educated than conservatives or vice versa, it is not unreasonable to believe that education and political leanings are independent. Many liberals are highly educated and many conservatives are highly educated. Unfortunately, there is no way to confirm this empirically with our data. This can be the "Achilles heel" of the instrumental variable (IV) estimation procedure. When we choose an instrument, it must be uncorrelated

with the omitted variable. Since there is no way to assess this empirically, we are implicitly assuming that the second good instrument condition is satisfied when we use the instrumental variables estimation procedure to address the omitted explanatory variables problem.

Justifying the Instrumental Variable (IV) Approach: A Simulation

We claim that while the instrumental variable (IV) estimation procedure for the coefficient value is still biased when an omitted explanatory variable problem exists, it will be consistent when we use a "good" instrument.

Econometrics Lab 20.2: Instrumental Variables - Omitted Variables: Good Instrument

While this claim can be justified rigorous, we shall avoid the complicated mathematics by using a simulation.

[Link to MIT-Lab 20.2 goes here.]

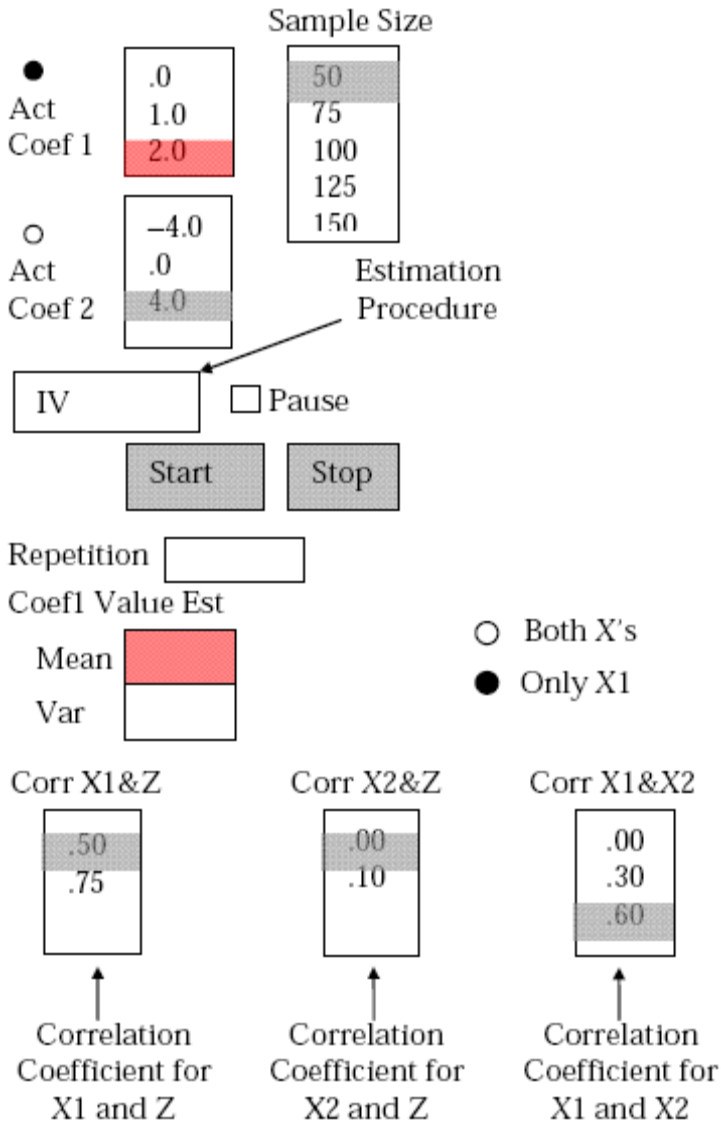


Figure 20.4: *IV Omitted Variable Simulation*

The model upon which this simulation is based on the following model:

$$y_t = \beta_{Const} + \beta_{x1}x1_t + \beta_{x2}x2_t + e_t$$

The Only X1 button is selected; hence, only the first explanatory variable will be included in the analysis. The model becomes:

$$\begin{aligned} y_t &= \beta_{Const} + \beta_{x1}x1_t + \beta_{x2}x2_t + e_t \\ &= \beta_{Const} + \beta_{x1}x1_t + (\beta_{x2}x2_t + e_t) \\ &= \beta_{Const} + \beta_{x1}x1_t + \varepsilon_t \qquad \text{where } \varepsilon_t = \beta_{x2}x2_t + e_t \end{aligned}$$

The second explanatory variable, *x2*, is omitted.

As before the values for the actual coefficients of the two explanatory variables are 2.0 and 4.0. The explanatory variable, *x2*, is omitted and only the first explanative variable, *x1*, is included. The correlation coefficient of the two explanatory variables equals .60. The included and omitted variables are positively correlated.

Since the instrumental variable (IV) estimation procedure, IV, is specified, two new lists appear. These two lists concern the instrumental variable, *z*. Recall that to be a good instrument two conditions must be met:

- included "problem" explanatory variable must be correlated so that the instrument acts as a good surrogate for the "problem" explanatory variable.
- error term must be independent so that we do not violate the explanatory variable/error term independence premise.

The Corr X1&Z list specifies the correlation coefficient of the included explanatory variable, *x1*, and the instrument, *z*. This correlation indicates how good a surrogate the instrument will be. An increase in correlation means that the instrument should become a better surrogate. By default this correlation coefficient equals .50. The Corr X2&Z list specifies the correlation coefficient of the omitted explanatory variable, *x2*, and the instrument, *z*. Recall how the omitted variable, *x2*, and the error term, ε_t, are related:

$$y_t = \beta_{Const} + \beta_{x1} x1_t + \varepsilon_t \quad \text{where } \varepsilon_t = \beta_{x2}\, x2_t + e_t$$

By default, .00 is selected from the Corr X2&Z list. Hence, the instrument, *z*, and the error term, ε_t, are independent. The second condition required for a "good" instrument is also satisfied. Initially, the sample size equals 50. Then, we increase from 50 to 100 and subsequently from 100 to 150. Table 20.6 reports results from this simulation:

Estimation Procedure	Correlation Coefficients X1&Z	X2&Z	X1&X2	Sample Size	Actual Coef1	Mean of Coef1 Ests	Magnitude of Bias	Variance of Coef1 Ests
IV	.50	.00	.60	50	2.0	≈1.82	≈.18	≈32.8
IV	.50	.00	.60	100	2.0	≈1.89	≈.11	≈13.7
IV	.50	.00	.60	150	2.0	≈1.95	≈.05	≈9.2

Table 20.6: *IV Estimation Procedure – Biased but Consistent*

Both bad news and good news emerge:

- **Bad News:** The instrumental variable estimation is biased. The mean of the estimates for the coefficient of the first explanatory variable, *x1*, does not equal the actual value we specified, 2.
- **Good News:** As we increase the sample size, the mean of the coefficient estimates gets closer to the actual value and the variance of the coefficient estimates becomes smaller. This illustrates the fact that the instrumental variable (IV) estimation procedure is consistent.

Next, let us see what happens when we improve the instrument by making it more correlated with the included "problem" explanatory variable. We do this by increasing the correlation coefficient of the included explanatory variable, *x1*, and the instrument, *z*, from .50 to .75 when the sample size equals 150:

Estimation Procedure	Correlation Coefficients X1&Z	X2&Z	X1&X2	Sample Size	Actual Coef1	Mean of Coef1 Ests	Magnitude of Bias	Variance of Coef1 Ests
IV	.50	.00	.60	150	2.0	≈1.95	≈.05	≈9.2
IV	.75	.00	.60	150	2.0	≈1.97	≈.03	≈3.9

Table 20.7: *IV Estimation Procedure – Improved Instrument*

The magnitude of the bias decreases and the variance of the coefficient estimates also decreases. We now have a better instrument.

Econometrics Lab 20.3: Instrumental Variables - Omitted Variables: Bad Instrument

[Link to MIT-Lab 20.3 goes here.]

Last, let us use the lab to illustrate the important role that the independence of the error term, ε_t, and the instrument, *z*, plays:

$$\varepsilon_t = \beta_{x2} x2_t + e_t$$

By specifying .10 from the Corr X2&Z list the error term, ε_t, and the instrument, *z*, are no longer independent:

Estimation Procedure	Correlation Coefficients X1&Z	X2&Z	X1&X2	Sample Size	Actual Coef1	Mean of Coef1 Ests	Magnitude of Bias	Variance of Coef1 Ests
IV	.50	.10	.60	50	2.0	≈2.67	≈.67	≈31.1
IV	.50	.10	.60	100	2.0	≈2.70	≈.70	≈13.8
IV	.50	.10	.60	150	2.0	≈2.73	≈.73	≈8.7

Table 20.8: *IV Estimation Procedure – Instrument Correlated with Omitted Variable*

As we increase the sample size from 50 to 100 to 150, the magnitude of the bias does not decrease. The instrumental variable (IV) estimation procedure is no longer consistent. This illustrates the “Achilles heel” of the instrument variable (IV) estimation procedure.