ECON 3040 - Instrumental Variables (IV) / Two-stage least-squares (2SLS)

Ryan T. Godwin

University of Manitoba

Instrumental Variables

For least-squares to work well, we need to make a very important assumption about the error term ϵ.

The error term ϵ must be independent from the x variables, or else least-squares is biased and inconsistent.

For example, in the simple model:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

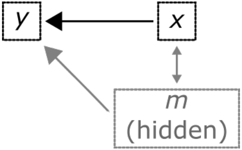
if x is correlated with ϵ then the least-squares estimator for β_1 will be wrong (biased and inconsistent)!

Missing, lurking, or confounding variables

The error term contains missing variables, that determine y. So, those missing variables need to be uncorrelated with the x variables for LS to work. This is often unreasonable!

A *lurking*, or *confounding* variable is one that threatens our ability to correctly estimate the effect that an x variable has on a y variable. Confounding variables are a major issue in analyses of *causal inference*, and are of tremendous import in many areas, not just economics.

Figure: A missing m variable that is correlated with x and that determines y will make estimation of the effect of x on y difficult (or impossible).



The situation depicted in the above Figure, where m is correlated with both x and y, implies that the effect of x on y cannot be estimated correctly by LS. That is, the estimated β_1 (b_1) is *wrong* in the population model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

The reason that b_1 gives the *wrong* answer for the true effect of x on y is that:

- A change in m is associated with a change in both x and y.
- When we "see" x changing, we know m is also changing.
- Attributing changes in y due to changes in x alone becomes impossible, since we don't know how much of the change in y came from m.

The solution to the problem is to include the m variable in the model! If we can't actually observe m (but we can imagine that it is there) then we must use clever strategies and more advanced methods to attempt to estimate the effect of x on y. One of those possible methods is Instrumental Variables (IV) estimation, the focus of this chapter.

House price again

Let's return to the house price data:

```r
house<- read.csv("https://rtgodwin.com/data/houseprice.csv")
bad.mod <- lm(Price ~ Fireplaces, data=house)
summary(bad.mod)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   171824       3234   53.13   <2e-16 ***
Fireplaces     66699       3947   16.90   <2e-16 ***
```

This model is suffering from omitted variable bias. The estimated effect of an additional fireplace on house price is wrong (biased and inconsistent). $67,000 is likely not the true effect. This is because there is a missing variable `Living.Area` (the size of the house in square feet), that is correlated with fireplaces and that also determines price. Notice that the missing variable is *inside* the error term (as are all other variables that determine y), but that this missing variable is correlated with x. This means that ϵ and x are correlated, and that least-squares will be biased and inconsistent.

Once we include the missing variable `Living.Area`, the problem is solved:

```
better.mod <- lm(Price ~ Fireplaces + Living.Area, data=
    house)
summary(better.mod)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14730.146   5007.563   2.942  0.00331 **
Fireplaces   8962.440   3389.656   2.644  0.00827 **
Living.Area   109.313      3.041  35.951  < 2e-16 ***
```

But what if we can't include the missing variable, because we don't observe it? All hope is not lost. If we can find an *instrument*, then we can still get a consistent estimator for the β.

Endogeneity

- When an x variable is correlated with the error term, that variable is sometimes said to be **endogenous**.
- Simultaneous causality (or just “simultaneity”) is another way that we can have endogeneity. We will soon see that this is the case with demand and supply.

Instrumental variable (IV)

A variable, z, qualifies as an instrument if it satisfies two conditions.

An instrumental variable, z, must be:

1. Correlated with the endogenous variable x.
   - This is sometimes called the “relevance” of an IV.
   - This condition can be tested.
2. Uncorrelated with the error term, or equivalently, uncorrelated with the missing variable m.
   - This is sometimes called the “exclusion” restriction.
   - This restriction cannot be easily tested.

If we can find a valid instrument, then we can use it to extract the “good” or “clean” variation from x. With endogeneity, changes in x are associated with changes in ϵ. But, changes in x **due to** z are not associated with the error term, because z is not correlated with ϵ.

IV estimation / Two-stage least-squares (2SLS)

Instrumental variables estimation, also called two-stage least-squares, is a statistical method for estimating β_1 in the equation:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

when x is endogenous (correlated to ϵ), but when we have a valid instrument z. The IV estimation gives us a consistent estimator for β_1, whereas LS gives us an inconsistent estimator and should not be used.

1st stage of 2SLS

In the first stage, we estimate an auxiliary regression to *extract* variation from x which is independent from ϵ. The 1st stage regression model is:

$$x = \alpha_0 + \alpha_1 z + u \tag{1}$$

After estimating this model by least-squares, we have the estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$. We then use this model to get the least-square predictions for x:

$$\hat{x} = \hat{\alpha}_0 + \hat{\alpha}_1 z \tag{2}$$

The LS predicted values $\hat{x}$ from equation 2 are independent from the error term! That is, $\hat{x}$ contains changes in x that are due to z only, and since z is uncorrelated with ϵ, so is $\hat{x}$ uncorrelated with ϵ.

2nd stage of 2SLS

In the second stage, we estimate the population model by LS, but instead of using x, we replace it with $\hat{x}$ from the 1st stage. Although x is endogenous, $\hat{x}$ is not! Estimating the following equation by LS gives us the IV estimator:

$$y = \beta_0 + \beta_1 \hat{x} + \epsilon$$

Direct formula for the IV/2SLS estimator

For the model $y = \beta_0 + \beta_1 x + \epsilon$, recall that the formulas for the LS estimators are:

$$b_1 = \frac{\sum \left[(y - \bar{y}) (x - \bar{x}) \right]}{\sum (x - \bar{x})^2}$$
$$b_0 = \bar{y} - b_1 \bar{x}$$

Applying these formulas to the 1st stage regression in equation 1, the formulas look like:

$$\hat{\alpha}_1 = \frac{\sum \left[(x - \bar{x}) (z - \bar{z}) \right]}{\sum (z - \bar{z})^2}$$
$$\hat{\alpha}_0 = \bar{x} - \hat{\alpha}_1 \bar{z}$$

The LS predicted values from the 1st stage are:

$$\hat{x} = \hat{\alpha}_0 + \hat{\alpha}_1 z = \bar{x} - \frac{\sum \left[(x - \bar{x}) (z - \bar{z}) \right]}{\sum (z - \bar{z})^2} \bar{z} + \frac{\sum \left[(x - \bar{x}) (z - \bar{z}) \right]}{\sum (z - \bar{z})^2} z \quad (3)$$

and the LS slope estimator for the model in the 2nd stage is:

$$b_1 = \frac{\sum \left[(y - \bar{y}) \left(\hat{x} - \bar{\hat{x}} \right) \right]}{\sum \left(\hat{x} - \bar{\hat{x}} \right)^2} \tag{4}$$

Plugging the predicted values (equation 3) into the 2nd stage estimator in equation 4) yields the formula for the IV estimator:

$$\hat{\beta}_{IV} = \frac{\sum \left[(y - \bar{y}) (z - \bar{z}) \right]}{\sum \left[(x - \bar{x}) (z - \bar{z}) \right]} \tag{5}$$

Example of a missing variable: Distance from college

Let's look at data from Card (1993).[1]

- Data contains *wage*, *years of education*, and demographic variables.
- Goal: estimate the returns to education in terms of *wage*.
- Problem: ability (intelligence) may be correlated with (cause) both wage and education.
- Since ability is unobservable (a missing variable), it is contained in the error term.
- The education variable is then correlated with the error term (endogenous).
- LS estimation of the returns to education may be inconsistent.

The population model that we want to estimate is:

$$wage = \beta_0 + \beta_1 education + \beta_2 urban + \beta_3 gender + \beta_4 ethnicity + \beta_5 unemp + \epsilon \tag{6}$$

[1] Card, D. (1993). *Using geographic variation in college proximity to estimate the return to schooling* (No. w4483). National Bureau of Economic Research.

We are primarily interested in β_1 (the returns to education). The other variables are included as controls, in order to avoid omitted variable bias. The difficulty with estimating equation 7 is that education is *endogenous*. From the Card (1993) paper:

"One of the most important "facts" about the labor market is that better-educated workers earn higher wages. Hundreds of studies in virtually every country show earnings gains of 5-15 percent (or more) per additional year of schooling. Despite this evidence, most analysts are reluctant to interpret the earnings gap between more and less educated workers as a reliable estimate of the economic return to schooling. Education levels are not randomly assigned across the population; rather, individuals make their own schooling choices. Depending on how these choices are made, measured earnings differences between workers with different levels of schooling may over-state or under-state the "true" return to education."

LS is the wrong method

First, let's try LS. It is the wrong method to use because it is inconsistent when there is endogeneity. Load the data, and estimate the model:

```r
college <- read.csv("https://rtgodwin.com/data/collegedist.
    csv")
ls <- lm(wage ~ education + urban + gender + ethnicity +
    unemp, data = college)
summary(ls)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.000192   0.156928  50.980   <2e-16 ***
education         0.005369   0.010362   0.518   0.6044
urbanyes          0.070117   0.044727   1.568   0.1170
gendermale        0.085242   0.037069   2.300   0.0215 *
ethnicityhispanic 0.012048   0.062385   0.193   0.8469
ethnicityother    0.556056   0.052167  10.659   <2e-16 ***
unemp             0.133101   0.006711  19.834   <2e-16 ***
```

Notice that the returns to education are estimated to be very small (an additional year of education leads to an increase in wage of half of a cent per hour). No point in going to school! But we know that LS is wrong (inconsistent) if *education* is correlated with the error term.

2SLS using distance from college as an IV

Now let's try using *distance from college* (while attending high school) as an instrument for education. The argument for the validity of this instrument is that:

- *distance from college* is correlated with *education*, since the closer a student is, the cheaper it is to get an education
- *distance from college* is uncorrelated with the missing variables that simultaneously determine education and wage

1st stage

To use this *distance from college* variable in two-stage least-squares, we first regress *education* (the problem endogenous variable) on *distance from college* and all the controls. Then we save the LS predicted values from this regression:

```
first.stage <- lm(education ~ urban + gender + ethnicity +
    unemp + distance,
                  data = college)
education.hat <- first.stage$fitted.values
```

2nd stage

Now, we estimate the original population model in equation 7 using LS, but we replace the *education* variable with 1st stage predicted values $\widehat{education}$. That is, we estimate the equation:

$$wage = \beta_0 + \beta_1 \widehat{education} + \beta_2 urban + \beta_3 gender + \beta_4 ethnicity + \beta_5 unemp + \epsilon \tag{7}$$

The R code is:

```r
iv <- lm(wage ~ education.hat + urban + gender + ethnicity +
     unemp, data = college)
summary(iv)
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.657025   1.358890  -0.484  0.62876
education.hat      0.647099   0.100592   6.433 1.38e-10 ***
urbanyes           0.046144   0.044691   1.033  0.30188
gendermale         0.070753   0.036978   1.913  0.05576 .
ethnicityhispanic -0.124051   0.065641  -1.890  0.05884 .
ethnicityother     0.227240   0.072984   3.114  0.00186 **
unemp              0.139163   0.006748  20.622  < 2e-16 ***
---
Signif.codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.263 on 4732 degrees of freedom
Multiple R-squared:  0.1175,  Adjusted R-squared:  0.1163
F-statistic:   105 on 6 and 4732 DF,  p-value: < 2.2e-16
```

The return to education is now positive and significant!

Under LS the estimated returns to education are 0.005, but under IV they are 0.647.

Using the direct formula: `ivreg()`

We can use a direct formula like in equation 5 to get the IV estimates (instead of using the two-stage approach). Install and load the `ivreg` package:

```r
install.packages("ivreg")
library(ivreg)
```

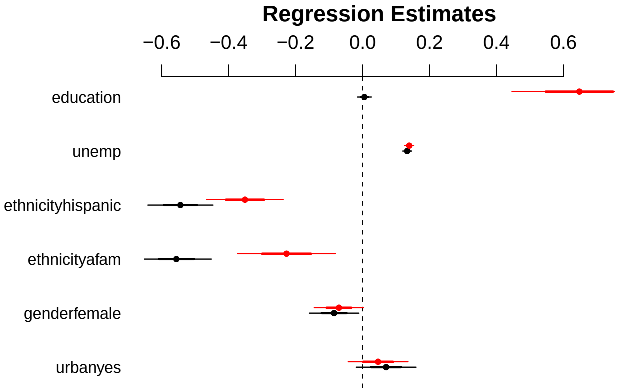
To use the `ivreg()` function, we need to specify the population model that we want to estimate, and then the list of instruments that we will use. The population model and list of instruments are separated by `|`:

```r
iv <- ivreg(wage ~
           education + urban + gender + ethnicity + unemp |
            distance + urban + gender + ethnicity + unemp,
            data = college)
summary(iv)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.65702    1.83641  -0.358   0.7205
education          0.64710    0.13594   4.760 1.99e-06 ***
urbanyes           0.04614    0.06039   0.764   0.4449
gendermale         0.07075    0.04997   1.416   0.1569
ethnicityhispanic -0.12405    0.08871  -1.398   0.1621
ethnicityother     0.22724    0.09863   2.304   0.0213 *
unemp              0.13916    0.00912  15.259  < 2e-16 ***
```

We get the same results as from using the two-stage method!

Figure: Results of LS and IV (in red) regression using Card (1993) data. Dependent variable is *wage*; *distance from college* is an instrument for *education*. Horizontal lines are 95% and 99% confidence intervals. Notice that the returns to education are insignificant under LS, but significant under IV.



Estimating demand with IV

We have tried to estimate a demand curve several times in this course. We have been doing it wrong! This is because the *price* variable that we have been using as a regressor (on the RHS of the model) is endogenous! The price and quantity values that we observe in our data set are actually due to the *intersection* of demand and supply. The price and quantity values that we observe are due to *two* equations, demand and supply:

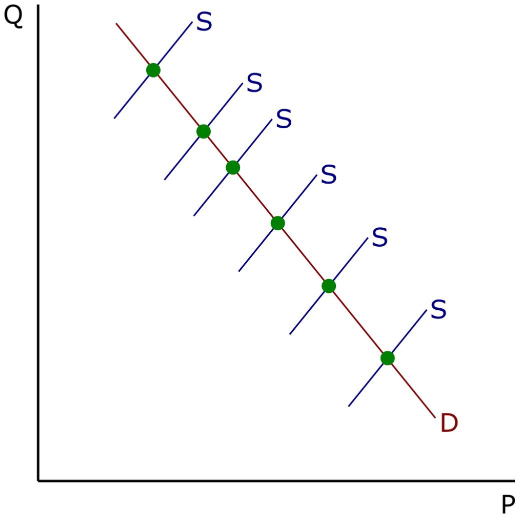
$$\begin{array}{lllllllll} q & = & \alpha_0 & + & \alpha_1 p & + & \alpha_2 s & + & \varepsilon \qquad \text{(supply)} \\ q & = & \beta_0 & + & \beta_1 p & + & \beta_2 d & + & \epsilon \qquad \text{(demand)} \end{array} \tag{8}$$

where:

- q is *both* quantity demanded and supplied
- p is price
- d are "demand-shifters" (such as income, prices of complements and substitutes, etc.)
- s are "supply-shifters" (such as prices of inputs, weather, etc.)
- α_1 should be $(+)$ and β_1 should be $(-)$

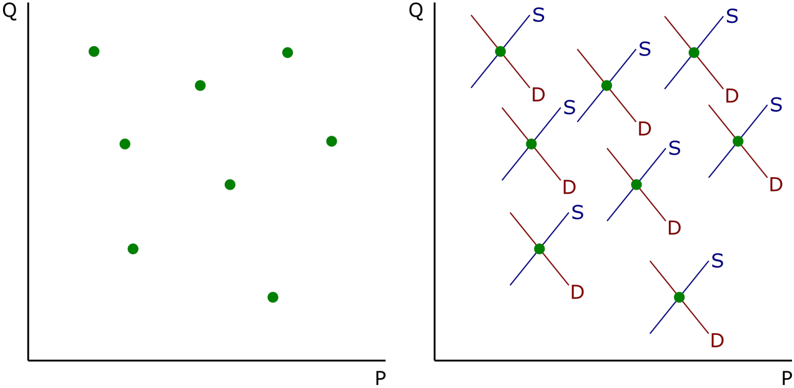
The relationship between q and p is *both* positive and negative (depending on whether we look at the supply or demand curve)! How can we fit a line through price and quantity data, and call it a demand curve? We could be estimating the supply curve, or (most likely) a combination of the two. If we want to estimate the slope of the demand curve, then we need to hold it's position constant. That is, the variation in price would have to come only from shifts in supply, so that we are tracing out points along a demand curve.

Figure: In order to estimate the slope of the demand curve, variation in quantity and price must come from shifts in supply.



The problem is, the demand curve is shifting along with the supply curve! The data that we observe is the result of demand and supply intersecting.

Figure: Price and quantity data is the result of the intersection of shifting demand and supply curves. We cannot attribute changes in quantity due to changes in price as coming just from the demand curve. Quantity and price are *endogenous* variables.



Fulton fish market data

Graddy (1995) produces data on the Fulton fish market, and Angrist, Graddy, and Imbens (2000) estimate the demand curve in this market using instrumental variables. The version of the data that we use is from Wooldridge (2020). Download the data:

```r
fish <- read.csv("https://rtgodwin.com/data/fish.csv")
```

Table: Description of **some** of the variables in the Graddy (1995) Fulton fish market data. We only use a few variables for this example. In parentheses the variables are labeled as either demand-shifters d or supply-shifters s.

totqty (q)	quantity of fish sold that day
avgprc (p)	price of fish that day
mon (d)	dummy variable equal to 1 if it's Monday
tues (d)	
wed (d)	
thurs (d)	
wave2 (s)	average max last 2 days wave height
wave3 (s)	average max wave heights of 3 and 4 day lagged heights

The variables in the data set are shown in Table 1. Demand may change depending on the day: the dummy variables are the demand-shifters. Supply is affected by the weather: if the sea is rough it is harder to fish. Using the variables `wave2` and `wave3` as instruments for price, we can use variations in price that are due to changes in supply only, in order to estimate the slope of the demand curve. Graddy's own description of the 2SLS approach:

"...first a regression is run with log price as the dependent variable and the storminess of the weather as the explanatory variable. This regression seeks to measure the variation in price that is attributable to stormy weather. The coefficients from this regression are then used to predict log price on each day, and these predicted values for price are inserted back into the regression."

To estimate the demand equation:

$$\log(totqty) = \beta_0 + \beta_1 avgprc + \beta_2 mon + \beta_3 tues + \beta_4 wed + \beta_5 thurs + \epsilon$$

using 2SLS/IV, we can use the R code:

```r
install.packages("ivreg")
library(ivreg)
iv.fish <- ivreg(log(totqty) ~
              log(avgprc) + mon + tues + wed + thurs |
            wave2 + wave3 + mon + tues + wed + thurs,
                 data = fish)
summary(iv.fish)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.16410    0.18171  44.930  < 2e-16 ***
log(avgprc) -0.81582    0.32744  -2.492  0.01453 *  
mon         -0.30744    0.22921  -1.341  0.18317    
tues        -0.68473    0.22599  -3.030  0.00318 ** 
wed         -0.52061    0.22357  -2.329  0.02209 *  
thurs        0.09476    0.22521   0.421  0.67492    
```

Since the variables are in logs, we have estimated the *elasticity* of the demand curve: when price increases by 1%, the quantity demanded is estimated to decrease by 0.81582%.

Let's compare this to the LS estimates (as we would have done in previous chapters):

```
ls.fish <- lm(log(totqty) ~ log(avgprc) + mon + tues + wed +
     thurs, data = fish)
summary(ls.fish)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.24432    0.16281  50.637  < 2e-16 ***
log(avgprc) -0.52466    0.17611  -2.979  0.00371 **
mon         -0.31093    0.22582  -1.377  0.17193
tues        -0.68279    0.22267  -3.066  0.00285 **
wed         -0.53389    0.21994  -2.427  0.01717 *
thurs        0.06723    0.22042   0.305  0.76107
```

The LS estimate for the elasticity is much lower (0.52466%). The LS estimator is inconsistent because price is an endogenous variable!