Instrumental Variables and Two Stage Least Squares

- Why Use Instrumental Variables?

- Instrumental Variables (IV) estimation is used when your model has endogenous \( x \)'s

- That is, whenever \( \text{Cov}(x, u) \neq 0 \)

- IV can be used to address the problem of omitted variable bias

- Omitted variable bias is particularly difficult to address when the omitted variable is not observable.
The leading example in labor economics is ability. It is very difficult to find convincing proxies for ability.

Consider the wage equation that tries to determine the benefits of education in terms of future earnings:

\[ \text{earn} = \beta_0 + \beta_1 \text{educ} + u \]

where 'earn' is the wage rate and 'educ' is the number of years of schooling. The error term \( u \) contains omitted factors such as sociodemographics, location, as well as unobservable factors, most importantly, ability. It is reasonable to assume that ability is correlated with 'educ'. Therefore, in this example, we have to work under the assumption that \( \text{Cov}(x, u) \neq 0 \). This means that conventional OLS regression can not be used to estimate the parameter of interest, \( \beta_1 \) (measuring the effect on earnings of one unit of additional schooling).
• IV can also be used to address the problem of simultaneity, i.e. the problem that two variables are determined simultaneously.

• Consider the problem of estimating a demand function:

\[ q^d = \beta_0 + \beta_1 p^d + u \]

• Firms (especially monopolists) are very interested to know the parameter \( \beta_1 \). It measures the effect on quantity demanded of raising the price by $1. Thus, with the knowledge of \( \beta_1 \), the firm can find the optimal price in order to maximize profit.
• Estimating the demand equation is difficult. The reason is that observed price and quantity are the result of market equilibrium which equates demand and supply. In other words, there is a second equation, the supply equation

\[ q^s = \gamma_0 + \gamma_1 p^s + v \]

and the condition that for observed price and quantity it must hold that

\[ q^s = q^d. \]

• The solution to the problem of estimating demand is to find variables that only affect the supply of the product. These variables shift the supply curve without affecting the demand curve.
• And Example: The Fulton Market for Fish (Graddy, 1996; Angrist, Graddy and Imbens, 2000).

• The Fulton Fish market is a wholesale fish market with the customers owning retail fish stores or restaurants. The study focuses on the sale of Whiting which is a cheap fish in large supply. This fish is ideal for the analysis because there is no good substitute for it (ie. other fish available at large quantities and comparable prices).

• Data available for the study is in the form of total price and quantity transacted by one particular dealer on a particular day, observed over roughly 100 days of trading. There is no set price at the market and each dealer is free to charge a different price for each customer.
The data are thus in a form where we only observe the transacted quantity and price which are the result of a market equilibrium (ie. if there is more fish available, dealers will be willing to offer discounts to their customers in order to sell the fish). Demand may equally depend on seasonal variation (weekdays) and other factors affecting consumer preferences (other than price).

In order to estimate the demand function we need variation in the observed quantities and prices that is not due to consumer preference but only is due to variation in supply.
The variable used in this study is 'weather at sea'. Weather conditions at sea are measured in terms of wave heights and wind strength. At the same time one needs to control for weather conditions on shore and seasonal effects, such as the day of the week. Weather at sea affects the supply of fish, and as much as it is not related to severe weather on shore, does not affect demand for fish.

'Weather at sea' is a so called instrumental variable.
What Is an Instrumental Variable?

- In order for a variable, $z$, to serve as a valid instrument for $x$, the following must be true:

1. The instrument must be exogenous, that is,

   $$ \text{Cov}(z, u) = 0 $$

2. The instrument must be correlated with the endogenous variable $x$, that is,

   $$ \text{Cov}(z, x) \neq 0 $$
Are these conditions likely satisfied in the fish example? We need to convince ourselves, that in the demand equation

\[ q^d = \beta_0 + \beta_1 p^d + u \]

the variable 'Weather at sea' is not correlated with the error term \( u \) but is correlated with the price for fish \( p^d \). This is likely to be the case if weather at sea does not affect consumers preferences for fish and if it does affect the amount of fish caught on a particular day.
• We have to use common sense and economic theory to decide if it makes sense to assume $Cov(z, u) = 0$.

• Finding a good instrument for a particular application requires ingenuity and knowledge of the institutional details of the particular market or empirical problem.

• The second condition for a valid instrument is easier to check:

• We can test if $Cov(z, x) \neq 0$ by testing

$$H_0 : \beta_1 = 0$$

in the regression

$$x = \beta_0 + \beta_1 z + v$$

• Sometimes we refer to this regression as the first-stage regression
The IV Estimator, one X and one Z

- For

\[ y = \beta_0 + \beta_1 x + u, \]

and given our assumptions

\[
\begin{align*}
\text{Cov}(z, y) &= \text{Cov}(z, \beta_0 + \beta_1 x + u) \\
&= \text{Cov}(z, \beta_0) + \text{Cov}(z, \beta_1 x) + \text{Cov}(z, u) \\
&= \beta_1 \text{Cov}(z, x) + \text{Cov}(z, u) \\
&= \beta_1 \text{Cov}(z, x)
\end{align*}
\]

so

\[ \beta_1 = \frac{\text{Cov}(z, y)}{\text{Cov}(z, x)} \]
We see from this derivation that we need both conditions on $z$ to be able to identify $\beta_1$. In particular, if $\text{Cov}(z, x) = 0$ then $\beta_1$ is not well defined. In other words, in this case $\text{Cov}(z, y)$ does not contain useful information about $\beta_1$. 
Then the IV estimator for $\beta_1$ is obtained by replacing the population parameters $\text{Cov}(z, y)$ and $\text{Cov}(z, x)$ with sample averages.

We obtain

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})}$$

where

$$\bar{z} = n^{-1} \sum_{i=1}^{n} z_i$$

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

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$$
Two Stage Least Squares (TSLS)

- As it sounds, TSLS has two stages – two regressions:

- (1) First isolates the part of \( x \) that is uncorrelated with \( u \): regress \( x \) on \( z \) using OLS

\[
x_i = \pi_0 + \pi_1 z_i + u_i
\]

- Because \( z_i \) is uncorrelated with \( u_i \), \( \pi_0 + \pi_1 z_i \) is uncorrelated with \( u_i \). We don’t know \( \pi_0 \) or \( \pi_1 \) but we have estimated them, so...

- Compute the predicted values of \( x_i \), \( \hat{x}_i \), where

\[
\hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i,
\]

\( i = 1, \ldots, n \).
• (2) Replace \( X_i \) by \( \hat{X}_i \) in the regression of interest: regress \( Y \) on \( \hat{X}_i \) using OLS:

\[
Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i
\]

• Because \( \hat{X}_i \) is uncorrelated with \( u_i \) (if \( n \) is large), the first least squares assumption holds (if \( n \) is large)

• Thus \( \beta_1 \) can be estimated by OLS using regression (2)

• This argument relies on large samples (so \( \pi_0 \) and \( \pi_1 \) are well estimated using regression (1))
• This the resulting estimator is called the Two Stage Least Squares (TSLs) estimator, $\hat{\beta}_{1,TSLs}$. 
Two Stage Least Squares, ctd.

Suppose you have a valid instrument, \( Z_i \).

Stage 1: Regress \( X_i \) on \( Z_i \), obtain the predicted values \( \hat{X}_i \).

Stage 2: Regress \( Y_i \) on \( \hat{X}_i \); the coefficient on \( \hat{X}_i \) is the TSLS estimator, \( \hat{\beta}_{1,\text{TSL}} \).

\( \hat{\beta}_{1,\text{TSL}} \) is a consistent estimator of \( \beta_1 \).
We note that the first stage equation is

\[ x_i = \pi_0 + \pi_1 z_i + v_i \]

and the OLS estimator for \( \pi_1 \) is given by

\[ \hat{\pi}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^{n} (z_i - \bar{z})^2}. \]

We now form the prediction

\[ \hat{x}_i = \bar{x} - \hat{\pi}_1 \bar{z} + \hat{\pi}_1 z_i \]

or after rearranging,

\[ \hat{x}_i - \bar{x} = \hat{\pi}_1 (z_i - \bar{z}). \]

We now regress \( y_i \) on \( \hat{x}_i - \bar{x} \) and a constant such that

\[ \hat{\beta}_{1,TLS} = \frac{\sum_{i=1}^{n} (\hat{x}_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (\hat{x}_i - \bar{x})^2} \]
After substituting for \( \hat{x}_i - \bar{x} \) we obtain

\[
\hat{\beta}_{1,TSLS} = \frac{\sum_{i=1}^{n} \hat{\pi}_1 (z_i - \bar{z}) (y_i - \bar{y})}{\sum_{i=1}^{n} \hat{\pi}_1^2 (z_i - \bar{z})^2} = \frac{\sum_{i=1}^{n} (z_i - \bar{z}) (y_i - \bar{y})}{\hat{\pi}_1 \sum_{i=1}^{n} (z_i - \bar{z})^2} = \frac{\sum_{i=1}^{n} (z_i - \bar{z}) (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x}) (z_i - \bar{z})} = \beta_{IV}
\]

Thus, we have shown that the two stage procedure is algebraically identical to the IV estimator.
We return to the wage equation example

\[ \text{earn} = \beta_0 + \beta_1 \text{educ} + u. \]

One way to estimate the parameter \( \beta_1 \) is to find an instrumental variable \( z \) such that

\[ \text{Cov}(z, u) = 0 \]

and

\[ \text{Cov}(z, \text{educ}) \neq 0. \]

In the empirical literature, much effort is devoted to finding convincing instrumental variables.

The instrument is used to find variation in 'educ' that is not correlated with ability.
• We discuss three examples of instrumental variables:

– Mzor (1987) is using PSID (Panel Study of Income Dynamics, http://psidonline.isr.umich.edu/) data for the year 1975: In estimating the return to education for married women, a proposed instrument is number of years of schooling of the father of a married woman. The idea is that children of better educated parents tend to be better educated themselves, independent of individual ability. A first stage regression of 'educ' on 'fatheduc' gives a statistically significant coefficient and an $R^2 = .173$ for a sample of size 428. Since 'educ' and unobserved ability are positively correlated we expect the OLS estimate to be upward biased (overestimating the return to schooling). The instrumental variables estimates using 'fatheduc' as an instrument reduce the estimated returns to schooling by about 50% (see Example 15.1 in the textbook).
– Blackburn and D. Neumark (1992) use the NLS young man cohort data (NLS: National Longitudinal Survey, published by the U.S. Department of Labor, Bureau of Labor Statistics http://www.bls.gov/nls/home.htm): In estimating the return to education for men, they use number of siblings as an instrument for 'educ'. The first stage regression finds a significant negative effect of siblings on 'educ'. If it can be assumed that number of siblings is unrelated to unobserved ability then number of siblings is a valid instrument. Comparing OLS and IV estimates for the return to schooling we see that the return to schooling increases for the IV estimator using the number of siblings as an instrument. This casts doubt on the validity of the instrument. It could be that number of siblings is itself correlated with unobserved ability. For example, children with more siblings could receive less parental attention which could lead to lower ability.

– Card (1995) is using the NLS young man cohort data to analyze the return to education. The instrument used is proximity to a two or four
year college. Card finds that 'necr4' is a significant factor in explaining 'educ' when tested in a first stage regression. If location is uncorrelated with unobserved ability then nearc4 is a valid instrument.
Literature:


