Instrumental Variables

Introduction

- So far, in our quest to capture causal effects, we have tried to use controlling variables in our regression.

- The idea was that if we can control for all potential confounders (omitted variables), then the only channel through which X can affect Y must be a causal channel.

- Remember our definition of a causal effect was the effect of X on Y while holding **everything** else constant. And when we say 'everything', we mean **everything**!

- Now, is it realistic that we can hold everything else constant?? Not really!

Introduction

- So, is regression a waste of time and is all hope of finding causal effects lost? No!

- First, regression is incredibly useful for more than just finding causal effects. It is very widely used for prediction tasks of all sorts.

- And, even when we are interested in causality, regression can definitely give us a much better idea of a causal effect than a simple correlation. Every relevant variable you add to a regression marks an improvement over a simple correlation.

- However, in the next 8 or so lectures, we will learn several more advanced techniques for identifying causal effects. The first is **Instrumental Variables**.

Instrumental Variables

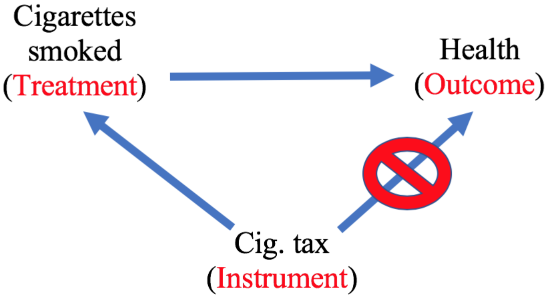
- Suppose we want to know the causal effect of smoking on your health. In it's simplest form, an instrumental variable is a variable which affects the outcome (your health) only through the regressor of interest (the treatment) (in this example, smoking).
- At first glance this seems fairly easy. However, good instrumental variables (IVs) are notoriously hard to find.
- For me, it has always made sense that IVs are so hard to come up with. We gain such a considerable amount that they should be hard to find!
- Any ideas of a potential instrument for our health and smoking example?

Instrumental Variables

- A possible IV could be the tax rate on cigarettes.
- This has actually been used in practice many times in both the economics and medical literature. It is particularly popular in looking at the causal effect of pregnant women's smoking habits on the birth weight of their children.
- For the IV to be valid, we need to think if there is any possible channel through which cigarette tax can affect your health other than through the amount you smoke.
- We will discuss many more examples of IV as we go on. But first we should understand how it works and how to use this new tool...

Instrumental Variables Explained

- The following picture highlights what we need from our IV and also helps us understand why it works.



Instrumental Variables Explained

- If the above graphic is correct, we can see the effect of cigarette tax on health (instrument on the outcome) is:

$$\begin{aligned}\text{Effect of tax on health} \quad &= \quad \text{Effect of tax on cigs} \\ &\quad\quad \times\text{Effect of cigs on health}\end{aligned}$$

- The thing we're interested in, the causal effect of cigarette smoking on health (the top arrow in the picture), is actually in the above equation!
- All we need to do is rearrange:

$$\text{Effect of cigs on health} \quad = \quad \frac{\text{Effect of tax on health}}{\text{Effect of tax on cigs}}$$

- We can calculate both of the effects on the right-hand-side of this equation, so we can get our causal effect!! No need to control for everything under the sun, we 'just' need a suitable IV!

IV estimator

- The equation above allows us to write a closed-form estimator. We need to estimate each of the two effects, which we can do using regression. Denote the outcome, health, as Y. Denote the treatment, cigarettes smoked, as X. Finally, denote the IV, cigarette tax, as Z.

$$\begin{aligned}\text{Effect of tax on health} &= (Z'Z)^{-1}Z'Y \\ \text{Effect of tax on cigs} &= (Z'Z)^{-1}Z'X\end{aligned}$$

- We can plug these into the previous equation to obtain

$$\begin{aligned}\text{Effect of cigs on health} &= \left[(Z'Z)^{-1}Z'X\right]^{-1}(Z'Z)^{-1}Z'Y \\ &= \left(Z'X\right)^{-1}(Z'Z)(Z'Z)^{-1}Z'Y \\ &= \left(Z'X\right)^{-1}Z'Y\end{aligned}$$

- **Notice**: this is not simply a regression of Z on Y; we are not replacing X with Z. The estimator uses information from all three variables.

IV estimator

- This estimator can also be motivated from a method of moments perspective.

- Recall that for the MM estimator in the bivariate case, we used

$$
\begin{aligned}
E[u] &= 0 \\
E[ux] &= 0
\end{aligned}
$$

- We set it so that these two expectations held in the population. Our $\hat{\beta}_0$ and $\hat{\beta}_1$ were defined by

$$
\begin{aligned}
\frac{1}{n}\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\frac{1}{n}\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i &= 0
\end{aligned}
$$

IV Estimator

- However, now we do not believe $E[ux] = 0$. We think that there may be some confounding variable in u which is correlated with x; we haven't controlled for everything. This was what was hurting our causal interpretation.

- Note: when $E[ux] \neq 0$ we say we have an **endogeneity** problem, or that x is **endogenous**. As opposed to **exogenous.**

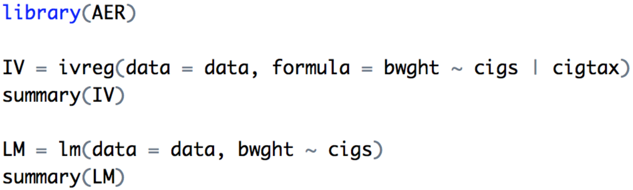
- But, we now have a new moment condition, $E[uZ] = 0$.

- Check for yourselves that the IV estimator is defined by

$$\begin{aligned} \frac{1}{n}\sum(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) &= 0 \\ \frac{1}{n}\sum(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)z_i &= 0. \end{aligned}$$

Example: Birthweight and cigarettes (bwght)

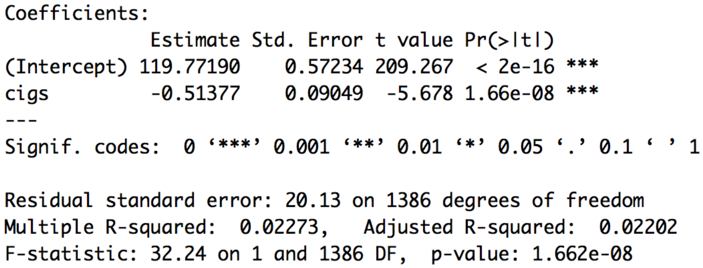
- In this example, we look at the causal effect of cigarettes on birthweight.



```
library(AER)

IV = ivreg(data = data, formula = bwght ~ cigs | cigtax)
summary(IV)

LM = lm(data = data, bwght ~ cigs)
summary(LM)
```

Example: Birthweight and cigarettes - OLS

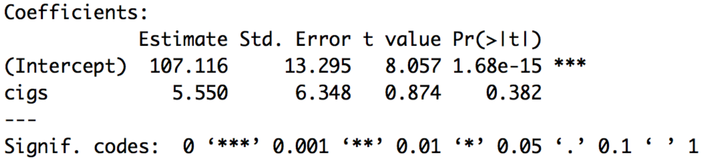
- The results for OLS are given first:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 119.77190    0.57234 209.267  < 2e-16 ***
cigs         -0.51377    0.09049  -5.678 1.66e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.13 on 1386 degrees of freedom
Multiple R-squared:  0.02273,   Adjusted R-squared:  0.02202
F-statistic: 32.24 on 1 and 1386 DF,  p-value: 1.662e-08
```

Example: Birthweight and cigarettes - OLS

- Now the IV results:



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  107.116     13.295   8.057 1.68e-15 ***
cigs           5.550      6.348   0.874    0.382
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Surprising? What's going on?

Properties of IV

- To work out what's going on in the previous example, we will first take a detour to look at some of the properties of the IV estimator.
- First, in general $\tilde{\beta}_{IV}$ is biased. This is because

$$E\left[\tilde{\beta}_{IV}\middle|X,Z\right]=\beta+(Z'X)^{-1}Z'E\left[u\middle|X,Z\right]$$

  and, in general, $E\left[u\middle|X,Z\right]\neq 0$ even if $E[u|Z]=0$.

- However, we can appeal to large sample properties. In particular, consistency, which is defined as

$$\text{plim }\hat{\theta}=\theta,$$

  i.e. the probability limit of an estimator is equal to the thing it is estimating.

plim Operator

- The plim operator is far nicer to deal with than the expectation operator. We can write

  $$\begin{aligned} \text{plim } g(X) &= g(\text{plim } X) \\ \text{plim } (XY) &= (\text{plim } X)(\text{plim } Y) \\ \text{plim } \frac{1}{n}\sum_i X_i &= E[X] \end{aligned}$$

  Notice that the first two properties generally don't hold for the expectation.The first result is from the continous mapping theorem, the second from Slutsky's theorem, and the last is a (weak) law of large numbers (LLN).

- Note, a sometimes useful sufficient condition (not necessary!) for consistency is that the bias and the variance converge to zero in the limit.

Consistency of IV

- We show the consistency of IV in a similar way to the unbiasedness of OLS. First plug in the true model and simplify, then take the plim

  $$\begin{aligned} \text{plim } \tilde{\beta}_{IV} &= \text{plim } \left(\beta + (Z'X)^{-1}Z'u\right) \\ &= \beta + (\text{plim } [Z'X])^{-1}\text{plim } Z'u \\ &= \beta + \left(\text{plim } \left[\frac{1}{n}Z'X\right]\right)^{-1}\text{plim } \frac{1}{n}Z'u \\ &= \beta + (E[Z'X])^{-1}E\left[Z'u\right] \\ &= \beta \end{aligned}$$

  where we have used that the plim of a constant is just that constant, the cont. map. theorem and Slutsky's theorem in the second equality, and the LLN in the penultimate equality.
- Notice two key conditions we need for the consistency of IV: (1) $E[Z'X]$ is invertible and (2) $E\left[Z'u\right] = 0$.

Consistency of IV

- The first condition requires that $E[Z'X]$ is full rank, which is satisfied if we have the same number of instruments, Z, as regressors, X, and that Z and X are **correlated**. This is commonly known as **relevance** of the instruments.

- The second condition requires that Z is exogenous, i.e. it is uncorrelated with the error term. This is known as the **validity** of the instruments.

- The correlation of Z with X and the uncorrelatedness of Z with u is what leads to our intuitive idea of what constitutes a suitable instrument: that Z can only affect Y through X.

- Note that it would be straightforward to test the relevance condition by regressing X on Z (or the other way around). However, how could we test the validity? We'll discuss this later.

Variance of IV

- It can be shown pretty easily that the conditional asymptotic variance of our IV estimator is

$$AVar\left(\tilde{\beta}_{IV}\middle|X,Z\right)=\sigma_u^2\left(Z'X\right)^{-1}Z'Z\left(X'Z\right)^{-1}.$$

- Notice that the correlation between Z and X appears 'on the bottom'. So the larger the correlation (in absolute value) the smaller the variance and the more accurate our estimator is.

- As the correlation between X and Z approaches 0, the variance approaches infinity. This is a problem known as **weak instruments**.

- As a rule-of-thumb (Stock and Yogo, 2005), we want the t-stat from a regression of X on Z to be greater than 3.2.

Weak Instruments

- When the correlation between X and Z is very weak, our IV estimator suffers badly.
- In particular, when we have weak instruments:
  - IV can become severely biased in small samples.
  - Inconsistency from a small violation of the validity condition gets magnified.
- Go back to our example on cigarettes and birthweight.... could weak instruments be the cause of the surprising result? Why do we think the instrument is weak?

Some Examples of Instruments

- Unfortunately, I don't have access to many interesting datasets that include reasonable instruments (and I want to keep some for the exercises). But it's fun to see what people have used in academic papers:

Angrist and Kreuger (1991)

Shows individuals' season of birth is related to their educational attainment because of the combined effects of school start age policy and compulsory school attendance laws. Individuals born in the beginning of the year start school at a slightly older age, and therefore are eligible to drop out of school after completing fewer years of schooling than individuals born near the end of the year. Our estimates suggest that as many as 25% of potential dropouts remain in school because of compulsory schooling laws. **We estimate the impact of compulsory schooling on earnings by using quarter of birth as an instrumental variable for education in an earnings equation.** This provides a valid identification strategy because date of birth is unlikely to be correlated with omitted earnings determinants. The IV estimate of the rate of return to education is remarkably close to the OLS estimate, suggesting that there is little ability bias in conventional estimates of the return to education. The results also imply that individuals who are compelled to attend school longer than they desire by compulsory schooling laws reap a substantial return for their extra schooling.

Angrist and Kreuger (1992)

Between 1970 and 1973 priority for military service was randomly assigned to draft-age men in a series of lotteries. Many men who were at risk of being drafted managed to avoid military service by enrolling in school and obtaining an educational deferment. **This paper uses the draft lottery as a natural experiment to estimate return to education**. The results suggest that an extra year of schooling acquired in response to the lottery is associated with 6.6 percent higher weekly earnings. This figure is about 10 percent higher than the OLS estimate of the return to education in this sample, which suggests there is omitted-variable bias in conventional estimates of the return to education.

Hotz et al. (2005)

In this paper, we exploit a "natural experiment" associated with human reproduction to identify the **effect of teen childbearing on subsequent educational attainment, family structure, labor market outcomes and financial self-sufficiency**. In particular, we exploit the fact that a substantial fraction of women who become pregnant **experience a miscarriage** and thus do not have a birth. If miscarriages were purely random then women who had a miscarriage as a teen would constitute an ideal control group with which to contrast teenage mothers. We devise an IV for the consequences of teen mothers not delaying their childbearing. Our major finding is that many of the negative consequences of not delaying childbearing until adulthood are much smaller than has been estimated in previous studies. While we do find adverse consequences of teenage childbearing immediately following a teen mother's first birth, these negative consequences appear short-lived.

Cesarini et al. (2017)

We study the effect of wealth on labor supply using the randomized assignment of monetary prizes in a sample of Swedish lottery players. We find winning a lottery prize modestly reduces labor earnings, with the reduction being immediate, persistent, and similar by age, education, and sex. A calibrated dynamic model of individual labor supply implies an average lifetime marginal propensity to earn out of unearned income of -0.11, and labor-supply elasticities in the lower range of previously reported estimates. The earnings response is stronger for winners than their spouses, which is inconsistent with unitary household labor supply models.

- (We'll discuss more examples later, but now we're moving on to extend our IV estimator)

Including Other Variables

- We said that we needed $(Z'X)$ to be square, i.e. X and Z are the same dimension.
- First, suppose we have one endogenous regressor and one instrument for it. However, we also want to add in some other controlling variables (which are exogenous - remember, exogenous variables are the good guys!).
- First, why might we want to do this? Consider the following example:
- We want to determine how an extra year spent in prison affects your earnings later in life. However, if we regress earnings on previous time in prison, there is likely to be a whole bunch of endogeneity.
- People who spend more time in prison are likely to be very different to people who spend less time in prison, these differences could be numerous and difficult to measure.

Including Other Variables

- So, we use an instrument. A popular instrument in this context is to use the harshness/leniency of the judge in the court trial. This works because judges are randomly assigned to cases (validity) and their leniency affects the amount of prison time they face (relevance).

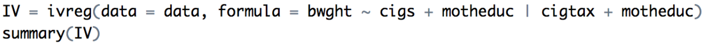
- However, in some states, judges are not assigned completely randomly. Some judges are more likely to preside over minor cases and others are more likely to preside over major cases. So, judges are assigned randomly after conditioning on the severity of the crime.

- Hence, for our instrument to satisfy the validity condition we must include the severity of the crime as a regressor.

- But now X includes two variables and Z only has one variable...

Including Other Variables

- The solution is not to think of Z as the instrument, but instead as the set of **exogenous** regressors.
- So in our justice example. X contains prison time and severity of the crime, and Z contains the leniency of the judge and the severity of the crime.
- In R, we can do this using the code:



```
IV = ivreg(data = data, formula = bwght ~ cigs + motheduc | cigtax + motheduc)
summary(IV)
```

More Than One Instrument

- But what about the other situation, when Z has a higher dimension than X.
- Why would we encounter this situation? It may be that you're super smart and can come up with several different potential instruments for your endogenous regressor. Or it may be that your instrument is measured using several variables.
- For example, in the season of birth example, it would require 3 dummy variables in order to capture the season of birth. That means we technically have 3 instruments.
- The solution is called **Two Stage Least Squares (2SLS).**

Two Stage Least Squares (2SLS)

- Assuming that each of the potential instruments are valid, we could choose to use the instrument which is the most relevant.
- However, this would be throwing away useful information (we never want to throw away information in statistics!).
- So, perhaps we can combine the instruments into one ’super instrument’. It turns out that there is a simple method to combine them in an optimal way.
- Suppose we have Z_1 and Z_2 which are both valid instruments for X. To create our optimal instrument we run the following regression

$$X = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \epsilon.$$

Two Stage Least Squares (2SLS)

- Denote the fitted value from this regression as $\hat{X}$. It turns out that $\hat{X}$ is the optimal instrument!

- Think about why this makes sense... $\hat{X}$ represents all of the information from Z_1 and Z_2 that explains X. This ensures I get the instrument that is as highly correlated with X as possible.

- The above regression represents the **first-stage** of 2SLS (rule-of-thumb: F-stat > 10). The **second-stage** is using $\hat{X}$ as the instrument for X.

- Thus, our 2SLS estimator is given by

$$\tilde{\beta}_{2SLS} = \left(\hat{X}'X\right)^{-1}\hat{X}'Y.$$

Two Stage Least Squares (2SLS)

- For practical purposes, it's useful to note that because $\hat{X} = Z\hat{\beta}_{OLS} = Z\left(Z'Z\right)^{-1}Z'X$, we can write our estimator as

$$\tilde{\beta}_{2SLS} = \left(\hat{X}'\hat{X}\right)^{-1}\hat{X}'Y.$$

- As a slightly more complicated example, suppose our model is

$$Y = \gamma_0 + \gamma_1 X + \gamma_2 W + u,$$

  where X is endogenous, W exogenous, and (Z_1, Z_2) are instruments for X.

- The first-stage would regress X on Z_1, Z_2, **and** W. So we regress the endogenous variable on **all exogenous** variables.
- One final point: we cannot use W as an instrument for X as it is already being used as an instrument for itself. We require our instruments to be **excluded** from the model. This is why you will sometimes see the validity condition being called the **exclusion restriction**.

2SLS with 2 or more Endogenous Regressors

- Suppose we have the following model

  $$Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 Y_3 + \beta_4 X + u,$$

  where Y_2 and Y_3 are endogenous variables (it is typical to denote endogenous variables as Y), X is an exogenous control variable, and we have access to Z_1, and Z_2 as instruments for Y_2, and Z_3 as an instrument for Y_3.

- Note that to estimate this model we need at least 2 excluded exogenous variables (we have 3 here, so we're fine).

- To estimate this model by 2SLS we proceed as follows:
  1. Regress Y_2 on all exog. variables (X, Z_1, Z_2, Z_3), obtain $\hat{Y}_2$.
  2. Regress Y_3 on all exog. variables (X, Z_1, Z_2, Z_3), obtain $\hat{Y}_3$.
  3. Regress X on all exog. variables (X, Z_1, Z_2, Z_3), obtain $\hat{X} = X$.
  4. Regress Y_1 on all fitted values $(\hat{Y}_2, \hat{Y}_3, X)$.

Some more Examples: Lundborg et al. (2017)

This paper introduces a new **IV strategy based on IVF** (in vitro fertilization) induced fertility variation among childless women to **estimate the causal effect of having children on their career**. For this purpose, we use administrative data on IVF treated women in Denmark. Because observed chances of IVF success do not depend on labor market histories, IVF treatment success provides a plausible instrument for childbearing. Our IV estimates indicate that fertility effects on earnings are: (a) negative, large and long lasting; (b) driven by fertility effects on hourly earnings and not so much on labor supply; and (c) much stronger at the extensive margin than at the intensive margin.

Fetzer et al. (2018)

This paper answers the question whether extreme power rationing can induce changes in human fertility and thus, generate “mini baby booms”. We study a period of extensive power rationing in Colombia that lasted for most of 1992 and see whether this has **increased births** in the subsequent year, exploiting variation from a newly constructed measure of the the extent of power rationing. We find that **power rationing** increased the probability that a mother had a baby by 4 percent and establish that this effect is permanent as mothers who had a black out baby were not able to adjust their total long-run fertility. Exploiting this variation, we show that women who had a black-out baby find themselves in worse socio-economic conditions more than a decade later, highlighting potential **social costs of unplanned motherhood.**

Draca et al. (2011)

In this paper we study the **causal impact of police on crime** by looking at what happened to crime and police **before and after the terror attacks** that hit central London in July 2005. The attacks resulted in a large redeployment of police officers to central London as compared to outer London – in fact, police deployment in central London increased by over 30 percent in the six weeks following the July 7 bombings, before sharply falling back to pre-attack levels. During this time crime fell significantly in central relative to outer London. Study of the timing of the crime reductions and their magnitude, the types of crime which were more likely to be affected and a series of robustness tests looking at possible biases all make us confident that our research approach identifies a causal impact of police on crime. The instrumental variable approach we use uncovers an elasticity of crime with respect to police of approximately -0.3, so that a 10 percent increase in police activity reduces crime by around 3 percent.

Zimmerman (2003)

I implement a quasi-experimental empirical strategy aimed at measuring **peer effects in academic outcomes**. In particular, I use data on individual students' SAT scores, and the SAT scores of their roommates. I argue that 1st-year roommates are assigned randomly with respect to academic ability. This allows me to measure differences in grades of high-, medium-, or low-SAT students living with high-, medium-, or low-SAT roommates. With random assignment these estimates would provide compelling estimates of the **effect of roommates' academic characteristic's on an individual's grades**. The results suggest that peer effects are almost always linked more strongly with verbal SAT scores than with math SAT scores. The effects are not large, but are statistically significant in many models.

Bronars and Grogger (1994)

We estimate the short-run and life-cycle effects of unplanned children on unwed mothers by **comparing unmarried women who first gave birth to twins with unwed mothers who bore singletons**. We find large short-term effects of unplanned births on labor-force participation, poverty, and welfare recipiency among unwed mothers, but not among married mothers. Although most of the adverse economic effects of unplanned motherhood dissipate over time for whites, there are larger and more persistent negative effects on black unwed mothers.

Angrist and Evans (1998)

Research on the **labor-supply consequences of childbearing** is complicated by the endogeneity of fertility. This study uses parental preferences for a mixed sibling-sex composition to construct instrumental variables (IV) estimates of the effect of childbearing on labor supply. IV estimates for women are significant but smaller than OLS estimates. The IV estimates are also smaller for more educated women and show no impact of family size on husbands' labor supply. A comparison of estimates using sibling-sex composition and twins instruments implies that the impact of a third child disappears when the child reaches age 13.

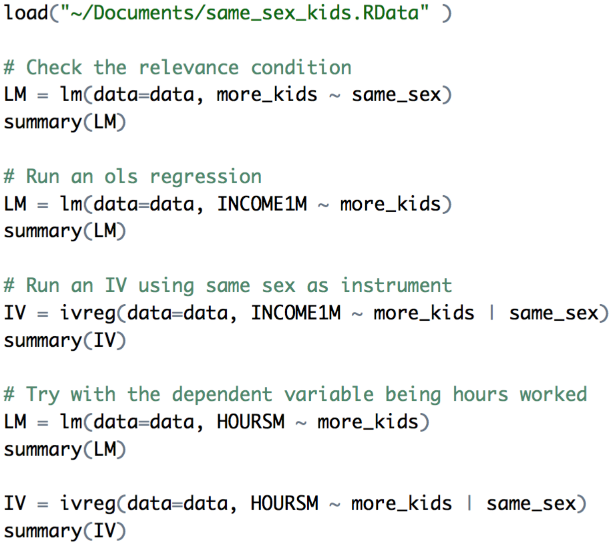
Miguel, Satyanath and Sergenti (2004)

Estimating the impact of economic conditions on the likelihood of civil conflict is difficult because of endogeneity and omitted variable bias. We use rainfall variation as an instrumental variable for economic growth in 41 African countries during 1981–99. Growth is strongly negatively related to civil conflict: a negative growth shock of five percentage points increases the likelihood of conflict by one-half the following year

Hanandita and Tampubolon (2014)

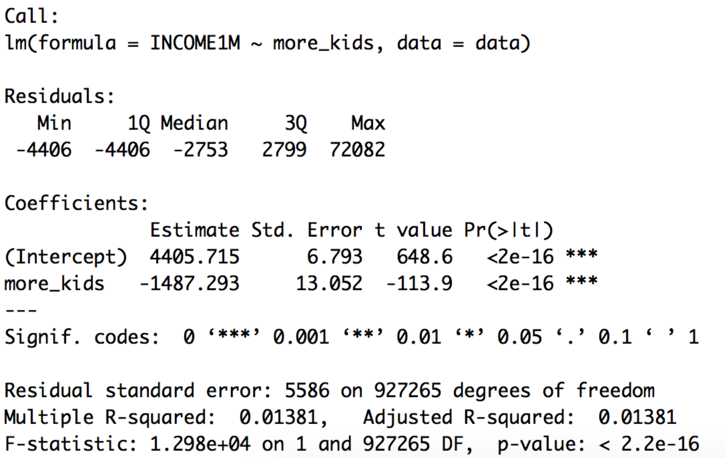
That poverty and mental health are negatively associated in developing countries is well known among epidemiologists. Whether the relationship is causal or associational, however, remains an open question. This paper aims to estimate the causal effect of poverty on mental health by exploiting a natural experiment induced by weather variability across 440 districts in Indonesia (N = 577,548). Precipitation anomaly in two climatological seasons is used as an instrument for poverty status, which is measured using per capita household consumption expenditure. Results of an instrumental variable estimation suggest that poverty causes poor mental health: halving one's consumption expenditure raises the probability of suffering mental illness by 0.06 point; in terms of elasticity, a 1% decrease in consumption brings about 0.62% more symptoms of common mental disorders. This poverty effect is approximately five times stronger than that obtained prior to instrumenting.

Example: Fertility and Labour (same_sex_kids.RData)



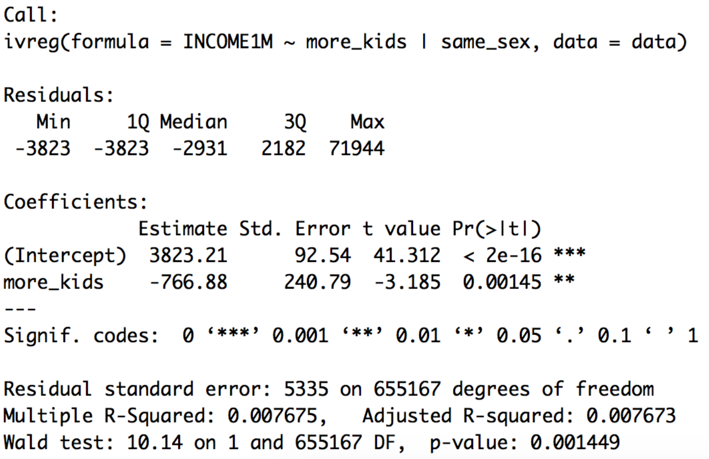
```
load("~/Documents/same_sex_kids.RData" )

# Check the relevance condition
LM = lm(data=data, more_kids ~ same_sex)
summary(LM)

# Run an ols regression
LM = lm(data=data, INCOME1M ~ more_kids)
summary(LM)

# Run an IV using same sex as instrument
IV = ivreg(data=data, INCOME1M ~ more_kids | same_sex)
summary(IV)

# Try with the dependent variable being hours worked
LM = lm(data=data, HOURSM ~ more_kids)
summary(LM)

IV = ivreg(data=data, HOURSM ~ more_kids | same_sex)
summary(IV)
```

Example: Fertility and Labour



```
Call:
lm(formula = INCOME1M ~ more_kids, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
 -4406  -4406  -2753   2799  72082 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4405.715      6.793   648.6   <2e-16 ***
more_kids   -1487.293     13.052  -113.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5586 on 927265 degrees of freedom
Multiple R-squared:  0.01381,	Adjusted R-squared:  0.01381 
F-statistic: 1.298e+04 on 1 and 927265 DF,  p-value: < 2.2e-16
```

Example: Fertility and Labour (same_sex_kids.RData)



```
Call:
ivreg(formula = INCOME1M ~ more_kids | same_sex, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -3823  -3823  -2931   2182  71944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3823.21      92.54  41.312  < 2e-16 ***
more_kids    -766.88     240.79  -3.185  0.00145 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5335 on 655167 degrees of freedom
Multiple R-Squared: 0.007675,   Adjusted R-squared: 0.007673
Wald test: 10.14 on 1 and 655167 DF,  p-value: 0.001449
```

Comments

- Ignore the R^2 in an IV regression. It isn't valid and doesn't have the usual interpretation.
- All of the same tests can be used with no issues. But, of course, the correct variance estimate should be used. Recall, the variance of IV is not the same as the variance of OLS.
- On this point, if you are carrying out 2SLS manually, the standard errors you obtain will be wrong because you are not accounting for the uncertainty in the first stage. The second stage treats the $\hat{X}$ as if they have not been previously estimated.

Application of IV: Measurement Error Models

- Here, we briefly look at 'another' useful application of IV.
- Measurement error occurs when a regressor we are interested in is measured with error. The model is written as

$$Y = \beta_0 + \beta_1 X + u$$

  but we don't observe X, we observe W, given by

$$W = X + \epsilon.$$

- A simple example: we want to look at how intelligence affects your wage. However, we only have access to your score on an IQ test.

Application of IV: Measurement Error Models

- In the 'classical measurement error model', X and ϵ are independent. This is unlikely to be a realistic assumption in many cases.

- However, we can make this assumption more palatable by supposing

$$\begin{aligned} W &= X \times \epsilon \\ \ln W &= \ln X + \ln \epsilon. \end{aligned}$$

- Here, we can recover an additive structure (which is very convenient for the mathematical analysis) by taking the natural logarithm, but now we allow the variance of ϵ to depend on X; a more reasonable assumption.

Application of IV: Measurement Error Models

- It turns out that if we have a repeated measurement of X

$$\begin{aligned} W_1 &= X + \epsilon_1 \\ W_2 &= X + \epsilon_2, \end{aligned}$$

  where ϵ_1 and ϵ_2 are independent. Then we can use W_2 as an instrument for W_1 (or the other way around).

- We will look at this model in more detail in Problem Set 6.

Comparing IV and OLS

- Suppose you have a suitable IV. Should you always use IV in this case?
- It's true that IV is always consistent, whether we have endogeneity or not. And it is also true that OLS is consistent only when we have exogeneity. However, recall the variance of the IV estimator $\sigma_u^2 \left(Z'X\right)^{-1} Z'Z \left(X'Z\right)^{-1}$, and the OLS estimator $\sigma_u^2 \left(X'X\right)^{-1}$.
- It can be shown that the variance of the IV estimator is larger than the variance of the OLS estimator. Only when $Z = X$ are they equal, in which case, the IV collapses to the OLS anyway.
- So the answer is: no! If you have exogeneity, OLS is far more precise.
- However, this idea of comparing OLS to IV gives me an idea (more specifically, it gave Durbin, Hausman, and Wu an idea in the '70s).

Testing for Endogeneity

- Suppose we want to test whether y_2 in the below model is endogenous or not

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x + u.$$

- This is just asking whether y_2 and u are correlated. Of course, we don't have access to u by definition. How to proceed?
- We use the fact that under exogeneity, both OLS and IV are consistent, but, under endogeneity, OLS is inconsistent but IV is still consistent.
- Essentially, if OLS and IV are similar, we may as well just use OLS because endogeneity doesn't seem to be a problem.
- Of course, in order to implement this test, we need a suitable instrument!

Testing for Endogeneity

- An equivalent and easier way to test endogeneity is to directly look at the covariance between u and y_2.

- The 'reduced form' for y_2 is given by

  $$y_2 = \gamma_0 + \gamma_1 z + \gamma_2 x + v,$$

  where z is the instrument for y_2. (The reduced form for y_2 is just a regression on **all** of the exogenous variables)

- Since z and x are exogenous by definition, the only way that y_2 can be correlated with u is if v is correlated with u.

- So, we look to test whether $Cov(v, u) = 0$. To do this with a regression, we would run

  $$u = \rho v + \epsilon$$

  and test if $\rho = 0$, a simple t-test of significance.

Testing for Endogeneity

- If we plug in for u we get

$$y_1 \quad = \quad \beta_0 + \beta_1 y_2 + \beta_2 x + \rho v + \epsilon$$

  and still, we're testing if $\rho = 0$ and we estimate this model by OLS.

- Of course, we don't have v, but we could just use $\hat{v}$ from the reduced form regression.

- So, the full procedure is:

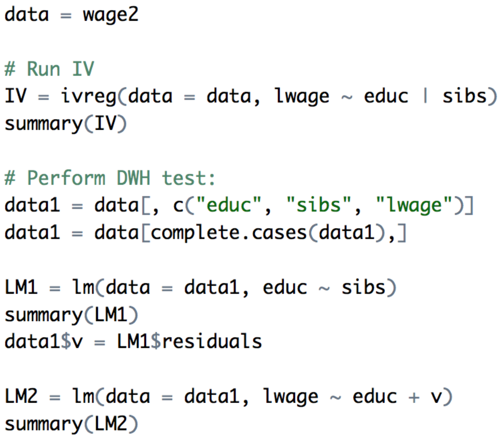
1. Estimate the reduced form equation for y_2

$$y_2 = \gamma_0 + \gamma_1 z + \gamma_2 x + v.$$

   Save the residuals $\hat{v}$.

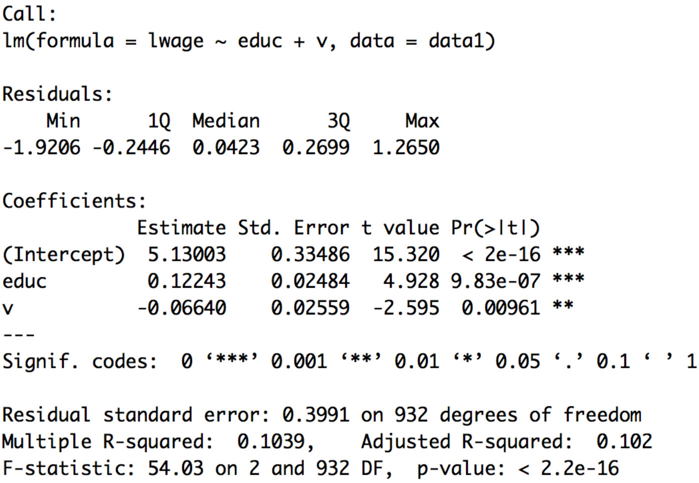
2. Add $\hat{v}$ to the original regression. If the coefficient on $\hat{v}$ is significant then we have an endogeneity problem.

Example: Testing for Endogeneity (wage2)



```
data = wage2

# Run IV
IV = ivreg(data = data, lwage ~ educ | sibs)
summary(IV)

# Perform DWH test:
data1 = data[, c("educ", "sibs", "lwage")]
data1 = data[complete.cases(data1),]

LM1 = lm(data = data1, educ ~ sibs)
summary(LM1)
data1$v = LM1$residuals

LM2 = lm(data = data1, lwage ~ educ + v)
summary(LM2)
```

Example: Testing for Endogeneity



```
Call:
lm(formula = lwage ~ educ + v, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9206 -0.2446  0.0423  0.2699  1.2650 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.13003    0.33486  15.320  < 2e-16 ***
educ         0.12243    0.02484   4.928 9.83e-07 ***
v           -0.06640    0.02559  -2.595  0.00961 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3991 on 932 degrees of freedom
Multiple R-squared:  0.1039,	Adjusted R-squared:  0.102 
F-statistic: 54.03 on 2 and 932 DF,  p-value: < 2.2e-16
```

- So we have a problem of endogeneity because v (actually $\hat{v}$) is significant.

Testing the Validity of the Instrument

- Another useful test in the context of IV is to **test the validity** of the instrument, i.e. if $Cov(z, u) = 0$. Recall that earlier in these lectures we said that we could test the relevance easily, but testing validity is more involved.

- This is sometimes called a **test of overidentification**. This terminology is from classical econometrics. A model is identified if we have enough information to estimate it. It is overidentified if we have more than enough information... eg. if we have two instruments for one endogenous regressor.

- Suppose we have one endogenous regressor and two instruments for it. Hausman (again!) suggested estimating β using just the first IV, estimate it again using the other IV, then compare the two. It turns out, there's a slightly simpler approach but we'll first discuss how to interpret the test.

Testing the Validity of the Instrument

- If the two estimates are different, then we have to conclude that one, or both instruments are invalid. However, there is no way to know which is causing the problem. Unless you are very sure about the validity of one, then you are effectively testing the validity of the other.

- There is a bit of an issue when testing the validity of instruments. It could be that the instruments give similar answers but are both in fact not valid. This could especially be true if the two instruments are similar (perhaps eg. a mum's education and a dad's education as instruments for their child's education).

- The test we are going to use is based on the fact that if all of our instruments are valid, then the error term from our 'structural equation' (the original regression of interest) is uncorrelated with the instruments.

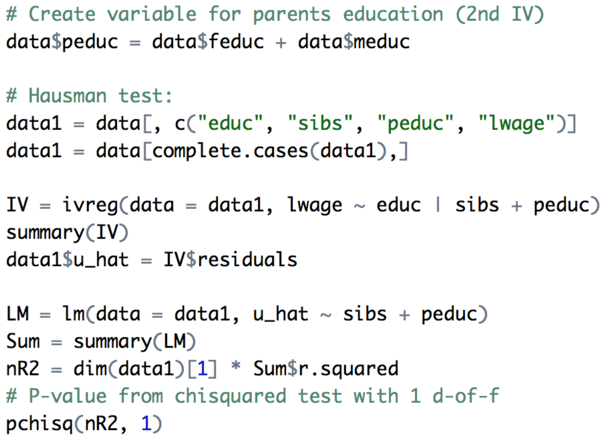
Testing the Validity of the Instrument

- The process works as follows:

1. Estimate the structural equation by IV (2SLS). Save the residuals $\hat{u}$.
2. Regress $\hat{u}$ on **all exogenous** variables. Calculate nR^2.

- Under H_0, we have $nR^2 \sim \chi_q^2$ where q is equal to the number of instruments minus the number of endogenous variables.
- If nR^2 is large, it implies that the instruments and u are highly correlated, and hence the instruments are not valid.

Example: Testing the Validity of the Instrument (wage2)



```r
# Create variable for parents education (2nd IV)
data$peduc = data$feduc + data$meduc

# Hausman test:
data1 = data[, c("educ", "sibs", "peduc", "lwage")]
data1 = data[complete.cases(data1),]

IV = ivreg(data = data1, lwage ~ educ | sibs + peduc)
summary(IV)
data1$u_hat = IV$residuals

LM = lm(data = data1, u_hat ~ sibs + peduc)
Sum = summary(LM)
nR2 = dim(data1)[1] * Sum$r.squared
# P-value from chisquared test with 1 d-of-f
pchisq(nR2, 1)
```

- The p-value comes out at 0.38, so we do not reject the null and conclude that the instruments may indeed be valid.

Summary

- We have seen how to use IV to solve our omitted bias problem and obtain a causal effect
- We have looked at the properties of the IV estimator, and seen how to use plim to show consistency.
- We've seen many examples of IVs that have been used in the academic literature.
- We've seen how to handle more than one instrument (2SLS) and how to deal with more than one endogenous variable.
- Finally, we've seen how to test for endogeneity and how to test the validity of our instruments.