LECTURE 7. INSTRUMENTAL VARIABLES: ASSUMPTIONS AND APPLICATIONS

Dr. Tom O'Grady

Last time, we learned how to do instrumental variables estimation, both for experiments with non-compliance and natural experiments. We discussed instrumental variables as a randomised *encouragement* to take a treatment, when the treatment itself is not randomly assigned. Today, we're going to delve more deeply into the assumptions underlying instrumental variables estimation. These are particularly important, not to mention controversial, when it comes to natural experiments. In this lecture, we will begin by defining the assumptions, and how we can (and can't) test for them. Then we'll apply them in practice to four different studies of varying quality.

1. Assumptions for Valid Estimation

There are four key assumptions needed for internal validity. For instrumental variables estimation to give causal estimates of the Local Average Treatment Effect in a natural experiment (or in a genuine experiment, to give us the Complier Average Causal Effect under two-sided non-compliance):

1. **Randomisation:** The instrument must be genuinely (as-if) randomly assigned [note that Angrist and Pischke call this 'ignorability']
2. **Relevance:** The instrument must be effective for at least some units, i.e. it must cause at least some people to actually take (more of) the treatment. It should also not be "too weak", which we'll define properly below.
3. **Exclusion Restriction:** The instrument must have no effect on the outcome except through its impact on treatment uptake.
4. **Monotonicity:** No units can be defiers

In addition, we must assume non-interference, as in standard experiments, although this requirement tends to be relatively unimportant in an instrumental variables setting. We won't be discussing it here, but Gerber and Green cover it. We also won't be heavily discussing the monotonicity assumption, since it is usually met very trivially.

1.1. Randomisation

Instrumental variables estimation relies on the fact that we have a randomised *encouragement* to take a treatment, even though treatment uptake itself is not randomly assigned because those who choose to comply with the encouragement may have very different potential outcomes than those who do not choose to comply. In the same way that a randomised experiment allows us to estimate the average treatment effect without bias, a randomised encouragement allows us to estimate both the proportion of compliers/first-stage effect and the intent-to-treat effect/reduced form without bias.

Why is this? Let's start with the case of a binary instrument. For the intent-to-treat effect, the logic is identical to a randomised experiment. We cannot recover the causal effect of being assigned to treatment if the potential outcomes of the encouraged and unencouraged are on average different. We require balance. When it comes to the proportion of compliers, recall the formula for it under two-sided noncompliance:

$$\begin{aligned} Prop.\ of\ Compliers &= \frac{1}{m}\sum_{i=1}^{m}(D_i \mid Z_i = 1) - \frac{1}{N-m}\sum_{i=m+1}^{N}(D_i \mid Z_i = 0)] \\ &= Prop.\ of\ Compliers\ and\ Always\ Takers - Prop.\ of\ Always\ Takers \end{aligned}$$

In order to get a causal estimate of the proportion of compliers, we require that there be an equal number of always-takers in the encouraged ($Z = 1$) and unencouraged ($Z = 0$) groups. This is ensured by randomisation. Just as there will, on average, be balance between the two groups in terms of personal characteristics like age and gender (if our units are people), so too will there be balance in terms of compliance types. With a continuous instrument, we require a causal estimate of the first stage effect:

$$\hat{D}_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i$$

If values of Z are not randomly assigned, we will obtain a biased estimate of the fitted values $\hat{D}_i$. Suppose, for instance, that units with higher potential outcomes are more likely to receive higher values of Z. Then, we cannot be certain that our second-stage estimates will give us the causal effect of changes in the treatment induced by the instrument, or merely selection bias resulting from different types of units having higher values of Z. The variation in the treatment that we are using in the second stage might no longer be exogenous.

Therefore a randomly assigned instrument is essential. Of course, many instruments in experiments are not actually randomly assigned by the analyst, as demonstrated in today's papers. In those cases, **we require that the instrument is "as if" randomly-assigned, i.e. it is allocated in such a way that those who receive it (or more of it) have the same average potential outcomes as those who do not receive it (or receive less of it).** How likely is this in practice? The assumption of randomisation is, of course, most plausible in cases with a genuinely randomly assigned instrument, like the Hajj lottery example. But even then, randomisation failures are possible, such as people with higher potential outcomes bribing officials to fix the visa lottery. In other cases without actual random assignment, the analyst needs to carefully justify why the particular natural experiment approximates random assignment. As in randomised experiments, balance tests can be used to test for random assignment of the instrument, although imbalance on unobserved characteristics of course cannot be tested. Strong knowledge on the part of researchers about the particular case being studied can also help a lot. We need to understand carefully the circumstances under which the instrument was "assigned" in order to be sure that it approximates random assignment. The Hajj researchers needed to delve into the mechanism by which Pakistan assignes visas, for example, to be certain that the lottery was not open to abuses.

1.2. Relevance

For instrumental variables estimation to work at all, the instrument must be *relevant* in the sense that it causes at least some people to take (more of) the treatment. In an extreme case where the instrument has no impact on the treatment at all, no estimates can be computed. This can easily be tested using the proportion of compliers or first-stage effect. The proportion of compliers must be greater than zero, or the coefficient on Z in the first-stage must be non-zero.

A more subtle but much more common problem is of weak instruments, where the instrument has a positive but very small effect on the treatment. This is quite a common occurrence in natural experiments where compliance with the instrument is entirely voluntary. It is not unknown for instruments to have a proportion of compliers that is less than 10%, for instance. The problem is that this allows us to estimate a Local Average Treatment Effect, but it is likely to be heavily biased, particularly in small samples. We won't go into the technical details of why. Importantly though, the resulting bias can sometimes be almost as bad as the selection bias that would result from simple regression estimates. Thus weak instruments are a very serious problem. The proportion of compliers or the magnitude of the first stage effect provide an indication of the 'strength' of an instrument. In general, instruments that cause only a small number of units to take the treatment, or that explain only a tiny proportion of the variation in the treatment, should be avoided.

A more precise test of the strength of an instrument is possible using an F test. An F test compares two different regression models, one with more independent variables than the other, and determines how much explanatory power is added by the model with more independent variables. The model with more independent variables is known as the "unrestricted model." The F test can be used for any type of regression, but here we examine the first stage, with the restricted model excluding the instrument and the unrestricted model including it. The idea is that if the instrument is "strong", then the unrestricted model will do a much better job of explaining the treatment than the restricted model. The *F statistic* can be calculated by the following procedure, where D is the treatment and X is a set of covariates, if they are included:

1. Fit the Unrestricted Model (UR) with the instrument:

   $$D = \beta_0 + \gamma_1 X + \beta_1 Z + \epsilon_1$$

2. Fit the Restricted Model (R) without the instrument:

   $$D = \beta_0 + \gamma_1 X + \epsilon_2$$

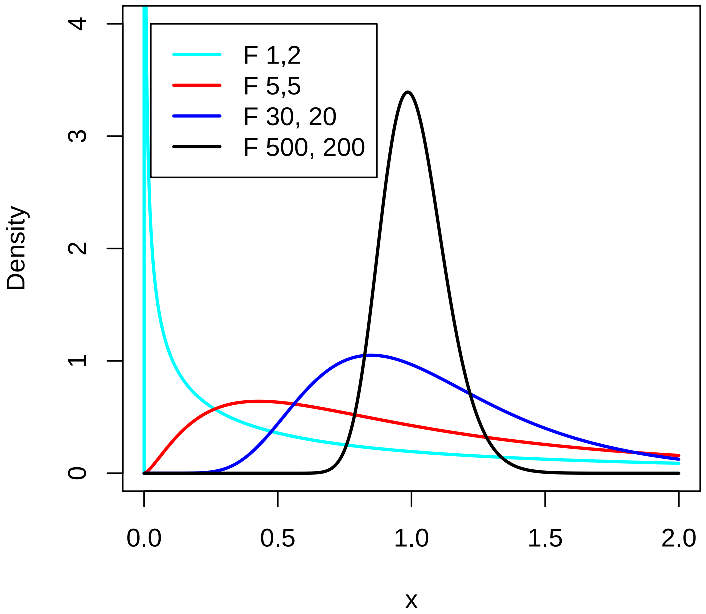
3. From the two results, compute the F Statistic:

   $$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

   where SSR=sum of squared residuals, q=number of instruments, k=number of independent variables in the unrestricted model, and n= number of observations.

Formally, both halves of this fraction have chi-squared distributions because the sum of a sequence of squared random variables follows this distribution. The F distribution arises from

Figure 1: **The F Distribution**



the ratio of two chi-squared distributions. It is described by $F(q, n-k-1)$. and with small n and q it looks like a right-skewed normal distribution.

The F statistic that we compute is just like a t statistic. Under the null hypothesis that the instrument adds no explanatory power, a high F value allows us to reject the null hypothesis and conclude that the instrument does explain the treatment. How big does the F statistic need to be? There is some controversy about this, but a rule of thumb that is often used is that an F statistic must be **greater than 10** for an instrument to be sufficiently strong. Below that, we fall into the "danger zone" where the results from two-stage least squares are likely to be very biased indeed. Conducting a test with an F statistic is also known as a **Wald test**, and we use a function with that name to carry out the test in R:

```
library(lmtest)
waldtest(model1,model2)
```

where `model1` and `model2` are the names of estimated regression models with and without the instrument. If the estimated F statistic falls below 10, the instrument is probably not relevant, in the sense that it does not have a strong enough effect on the treatment for instrumental variables estimation to provide unbiased estimates.

1.3. Exclusion Restriction

The exclusion restriction is not usually a serious problem in instrumental variables estimation of randomised experiments, but becomes much more serious once we move to studying natural experiments. The reason for it is that ultimately, we are usually not interested in the effect of the random encouragement to take the treatment, but only the effect of the treatment itself. Medical researchers who want to know whether vitamin A "works" only care about the impact of vitamin A on mortality, not the impact of randomly encouraging people to take vitamin A on mortality. The encouragement is only interesting to researchers because of its impact on the treatment. If the encouragement Z has some independent effect on Y that is not due to changes in D, then we cannot be sure that any treatment effect amongst the compliers is really due to D or to some other feature of being encouraged (Z).

One way to think about it is that factors that coincide with the encouragement beyond having an impact on treatment uptake should have no impact on the outcome. For example, in the Hajj lottery, it must be the case that merely winning or losing the lottery does not cause people to change their views on religious and social tolerance. Instead, the only way that the lottery should affect views can be through the fact that it causes some people to go on the Hajj, or not, which then has an impact on people's opinions. In many cases without a genuinely randomised instrument, the exclusion restriction is by far the most controversial assumption. It is rare to find naturally-occurring instruments that affect the outcome of interest only via the treatment. Importantly, it is also impossible to formally test for violations of it. Instead, we need to use theoretical and practical knowledge of the case being studied in order to justify the exclusion restriction in practice.

2. External Validity

In randomised experiments, we must worry about whether the experimental population that is being studied is typical of the population at large. But at least researchers have some leeway to choose who they experiment on. With natural experiments, that goes out of the window. First, we can only study places or groups where a natural experiment happened to occur, which may or may not be interesting to researchers. And second, even within that particular place or group, only some of the units will actually be affected by the instrument (the compliers). The compliers may be an atypical subgroup of an already atypical case.

It is always important to think about both the features of the particular case being studied and the likely characteristics of compliers. Beyond the exclusion restriction, external validity is the other area of instrumental variables estimation that tends to be controversial. Many economists, in particular, remain unconvinced that Local Average Treatment Effects for small subgroups are interesting. Others would reply that it is better to have clear causal estimates for some people than estimates for larger populations that are subject to selection bias.

3. Examples from Recent Studies

Now we'll look at four recent papers to help solidify the concepts of randomisation, relevance, the exclusion restriction and external validity. In each case, we can ask how convincingly the authors deal with these issues.

3.1. Clingingsmith, Khwaja and Kremer (2009)

This is an example of a particularly convincing instrumental variables study, in part because the instrument is genuinely randomly assigned (in this case via a lottery). Therefore in this study we have the following setup:

1. **Outcome of Interest:** Tolerance of other religions, women's rights, etc.
2. **Treatment:** Going on the Hajj pilgrimage
3. **Instrument:** The visa lottery

The assumptions, and tests of them, are as follows:

1. **Randomisation:** The visa lottery wins and losses must be genuinely randomly assigned. For example, people must not be able to manipulate the results of the lottery through bribery, corruption, etc. The authors use balance tables to show no evidence of randomisation failures, and state that the lottery was operated rigorously and was not subject to manipulation
2. **Relevance:** The lottery instrument was very highly relevant, with a proportion of compliers of 85.4%, as shown in Problem Set 4.
3. **Exclusion Restriction:** The lottery must only affect the outcome through the impact it has on people going on the pilgrimage, or not. This seems very plausible. It is possible, but not particularly likely, that losing or winning the lottery could directly alter someone's political/religious beliefs. The fact that the survey was taken 8-11 months after the lottery bolsters this (it seems highly implausible that the lottery could alter beliefs semi-permanently, even if it could cause short-term anger/happiness)
4. **External Validity:** It is possible that Pakistanis who enter the visa lottery are atypical of Muslims more generally. Perhaps they are more open to change and new experiences, for example. That would make the treatment effects larger than might be the case in other countries.

3.2. Madestam *et al* (2013)

These authors ask whether political protests work, in the sense of increasing participation in political movements, donations to those movements, and even votes by politicians. They look at the first big protests by the Tea Party movement in the USA, which occurred across the country on 9th April 2009. Did these protests have subsequent effects? This is a difficult question to answer with conventional regression due to selection bias. Protestors are likely to target areas that have higher potential outcomes – where their protests are more likely to work. Instead, the authors use rainfall on the day of the protests as an instrument that randomly encourages smaller protests in places with high rainfall.

1. **Outcomes of Interest:** Subsequent online and offline tea party participation, donations, subsequent media coverage in area, survey beliefs, votes by congress member

2. **Treatment:** Size of initial protest

3. **Instrument:** Rainfall on day of initial protest

The assumptions, and tests of them, are as follows:

1. **Randomisation:** Rainfall on the day of the protest needs to have occurred at random. Of course, certain places are in general rainier than others, and rainier places may have different potential outcomes. What matters here is whether rain fell on April 9th 2009 in a way that was correlated with potential outcomes. The authors do a good job of showing that this was probably not the case. Balance tests indicate that places experiencing high rainfall were not politically different to places with low rainfall. In addition they include "probability of rain" as a covariate in their two-stage least squares regressions, which measures the general 'raininess' of an area. Within a certain probability-band, it is much more likely that rainfall was genuinely random

2. **Relevance:** The instrument is strong. The mean decrease in attendance was 51% between rainy and non-rainy places

3. **Exclusion Restriction:** Here, we require that rainfall does not in itself affect tea party participation, voting by congress-people, etc., other than through its effect on the initial protest size. This seems plausible. One day of rainfall is very unlikely to directly alter any of those political variables. However, as the authors discuss, it could alter the experience people had at the protest rather than just its size, in which case the authors would really be measuring some combination of protest size and protest experience.

4. **External Validity:** The political effects of protests seem likely to be quite similar regardless of context, at least in democratic countries. Within this US case, the compliers are districts where high rainfall leads to lower protest turnout. They seem unlikely to be atypical.

3.3. Hainmueller and Kern (2009)

They ask whether exposure to western media decreases or increases support for authoritarian regimes. A naive comparison of people who do and do not consume western media would be subject to selection bias, because those who choose to consume it will have different potential outcomes to those who do not. They use a geographical quirk of the former East Germany as an instrument. Living in the Dresden area randomly discouraged people from watching western TV, since TV reception did not reach that them due to topography.

1. **Outcome of Interest:** Support for the East German regime, from survey data

2. **Treatment:** Watching Western TV

3. **Instrument:** Living in the cut-off area (Dresden region)

The assumptions, and tests of them, are as follows:

1. **Randomisation:** Access to television must be randomly assigned with respect to potential outcomes. The authors make a fairly good argument that the 'border' between areas with and without TV access did not map to political, economic or social divides (Figure 2 - although note that Dresden does seem to have had an older and poorer populace), so it was probably "as-if" randomly-assigned. It could have been the case, though, that there was residential sorting, whereby people who disliked the East German regime tried to move across the border to areas with TV reception. They argue that this sort of residential mobility was virtually impossible in East Germany, but it is not completely implausible.

2. **Relevance:** The instrument is highly relevant: the proportion of compliers was 62%, and Table 1 illustrates the strength of the instrument more informally. Virtually everybody in the Dresden area did not watch any western TV while virtually everybody outside of it did watch at least some of it.

3. **Exclusion Restriction:** The exclusion restriction requires that the only way in which living in the Dresden area affected political beliefs was through its impact on television viewing. This assumption is much more difficult to defend than in the previous two papers. There are many ways in which the Dresden area might have been distinctive, and hence might have shaped beliefs. Figure 2 suggests it was not particularly distinctive based on the characteristics included there (but note the comment above), although we can't be certain.

4. **External Validity:** While this study seems to provide interesting findings for East Germany, we might question how well they would generalise to other authoritarian regimes, time periods, etc.

3.4. Levitt (1997)

Levitt tackles the question of whether higher numbers of police reduce crime in American cities and states. A naive comparison of high- and low-police cities makes it look like more police increases crime, since more officers are hired in response to rising crime. This is an example of a problem known as *simultaneity.* Levitt uses election years as an instrument that randomly encourages higher numbers of police. Elections occur on a fixed cycle that was randomly set many years ago, and the number of police tends to rise in election years as mayors and governors try to win favour with the electorate by reducing crime in the runup to elections.

1. **Outcome of Interest:** Crime rates

2. **Treatment:** Police numbers

3. **Instrument:** Election year in city

The assumptions, and tests of them, are as follows:

1. **Randomisation:** Election years need to be genuinely randomly assigned. That should be a reasonable assumption here. Mayors can't manipulate the timing of elections to fall in years when there happens to be low crime, for example.

2. **Relevance:** The instrument seems weak. Police numbers only increase by around 1-2% in election years. This should give us very serious concerns about the paper's validity. Even with the full set of instruments, Levitt reports an F-statistic of 3.96, which is generally considered unacceecptably low. As a result, this paper would probably not be considered publishable today.

3. **Exclusion Restriction:** This requires that the only way in which election years affect crime rates is through their impact on police numbers. This also seems very problematic. If police numbers rise in election years, surely many other things are likely to rise too, including spending on other programs like welfare, education and health that might also reduce crime. Levitt does try to control for these in his regressions, but we really can't be certain that his controls fully account for all possible channels. Essentially, this puts us back into a world of modeling assumptions, where we must assume that all possible confounders have been accounted for. That substantially reduces the value of his paper.

4. **External Validity:** Levitt is mainly interested in the US only, and his sample does cover all US states and cities. So external vaidity is relatively high in terms of the initial sample. One might question how typical the compliers are, though. Is there something unique about cities that always increase their police numbers in election years? These might be cities where the elasticity of crime to police numbers is already high, meaning that the estimated effects might be unusually large.