# PSC 504: Instrumental Variables

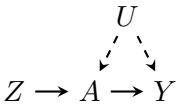
Matthew Blackwell

3/28/2013

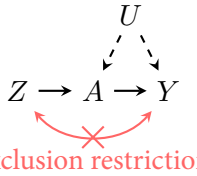
## Instrumental Variables and Structural Equation Modeling

### Setup

- The basic idea behind instrumental variables is that we have a treatment with unmeasured confounding, but that we have another variable, called the instrument, that affects the treatment, but not the outcome. With DAGs, it looks something like this:




- This DAG implies that the instrument, $Z$, is actually randomly assigned.




- This exclussion restriction means that there can be no common causses of the instrument and the outcome and no direct or indirect effect of the instrument on the outcome that does not go through the treatment.
- Another assumption implicit in this setup is that the instrument has a "first-stage" effect. That is, the instrument actually causes changes in the treatment.

### IV with constant effects

- Let's write down a causal model for $Y_i$ with constant effects and an unmeasured confounder, $U_i$. Here we assume that $E[A_i \eta_i] = 0$, so if we measured $U_i$, then we would be able to estimate $\tau$.

$$Y_i = \alpha + \tau A_i + U_i' \gamma + \eta_i$$

- If we have an instrument, $Z_i$, that satisfies the exclusions restriction, then we know that $\text{cov}(U_i'\gamma + \eta_i, Z_i) = 0$, because it must be independent of $U_i$ and it has no correlation with $\eta_i$ because neither does the treatment. With this in hand, we can formulate an expression for the average treatment effect here:

$$\tau = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(A_i, Z_i)} = \frac{\text{Cov}(Y_i, Z_i)/V[Z_i]}{\text{Cov}(A_i, Z_i)/V[Z_i]}$$

- Here, we can see that the average treatment effect is the population regression coefficient of $Y_i$ on $Z_i$ (called the "reduced form") divided by the population regression coefficient of $A_i$ on $Z_i$ (called the "first stage").
- With a binary instrument, there is a simply estimator based on this formulation called the Wald estimator. It is easy to show that:

$$\tau = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(A_i, Z_i)} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[A_i|Z_i = 1] - E[A_i|Z_i = 0]}$$

- Intuitively, these are the effects of $Z_i$ on $Y_i$ divided by the effect of $Z_i$ on $A_i$)

### Two-Stage Least Squares (2SLS)

- Now, let's write a model for the treatment and the instrument:

$$\begin{aligned} Y_i &= X_i'\beta + \tau A_i + \varepsilon_i \\ A_i &= X_i'\alpha + \gamma Z_i + \nu_i \end{aligned}$$

- We can plug the treatment equation into the outcome equation:

$$\begin{aligned} Y_i &= X_i'\beta + \tau[X_i'\alpha + \gamma Z_i + \nu_i] + \varepsilon_i \\ &= X_i'\beta + \tau[X_i'\alpha + \gamma Z_i] + [\tau\nu_i + \varepsilon_i] \\ &= X_i'\beta + \tau[X_i'\alpha + \gamma Z_i] + \varepsilon_i^* \end{aligned}$$

- Here we assume that $E[Z_i\nu_i] = 0$, $E[X_i'\nu_i] = 0$, and $E[X_i'\varepsilon_i] = 0$so that the first-stage parameters are identified. There are two things to note here. First is that the value in the brackets in the last line is the population fitted value of the treatment. Second, note that since $Z_i$ and $X_i$ are uncorrelated with $\nu_i$ and $\varepsilon_i$, then this fitted value is also independent of $\varepsilon_i^*$.
- Thus, the population regression coefficient of a $Y_i$ on $[X_i'\alpha + \gamma Z_i]$ is the average treatment effect, $\tau$.
- In practice, we estimate the first stage from a sample and calculate fitted values:

$$\hat{A}_i = X_i'\hat{\alpha} + \hat{\gamma} Z_i.$$

- Here, $\hat{\alpha}$ and $\hat{\gamma}$ are estimates from OLS. Then, we estimate a regression of $Y_i$ on $X_i$ and $\hat{A}_i$. We plug this into our equation for $Y_i$ and note that the error for $A_i$ is now a residual:

$$Y_i = X_i'\beta + \tau\hat{A}_i + [\varepsilon_i + \tau(A_i - \hat{A}_i)]$$

- This is valid because the fitted values are uncorrelated with $\varepsilon_i$ by the exclusion restriction and uncorrelated with the first-stage residuals by construction. Thus, this regression will consistently estimate the average treatment effect.
- Note that this isn't how we actually estimate 2SLS because the standard errors are all wrong. This is because the computer wants to calculate the standard errors based on $\varepsilon_i^*$, but what we really want is the standard errors based on $\varepsilon_i$.

### Overidentification

- Of course, there's nothing stopping us from including more instruments in the first stage equation. But they must be instruments: if we include a variable in the first stage, but not the second and it does affect $Y_i$, then we could get biased estimates.
- With more instruments than causal parameters of interest, we say that the model is **overidentified**, whereas with one instrument and one causal parameters, we say it is **just-identified**. With more than one instrument and constant effects, we can test for the plausibility of the exclussion restriction(s) using an overidentification test.
- If we reject the null hypothesis in these overidentification tests, then it means that the exclusion restrcitions for our instruments are probably incorrect. Note that it won't tell us which of them are incorrect, just that at least one is.
- These overidentification tests depend heavily on the constant effects assumption, which is why I'm not going into detail about this. Once we move away from constant effects, we no longer can generally pool multiple instruments together in this way.

## Instrumental Variables and Potential Outcomes

### Setup

- The basic idea behind instrumental variable approaches is that we do not have ignorability for $A_i$, but we do have a variable, $Z_i$, that affects $A_i$, but only affects the outcome through $A_i$.
- Note that we allow the instrument, $Z_i$ to have an effect on $A_i$, so the treatment must have potential outcomes, $A_i(1)$ and $A_i(0)$, with the usual consistency assumption: $A_i = Z_i A_i(1) + (1 - Z_i)A_i(0)$.
- Of course, now the outcome can depend on both the treatment and the instrument, so we have $Y_i(a, z)$ is the outcome if unit $i$ had received treatment $A_i = a$ and instrument value $Z_i = z$.
- The effect of the treatment given the value of the instrument is $Y_i(1, Z_i) - Y_i(0, Z_i)$.

### Key assumptions

- Randomization: the first assumption is that the instrument is randomized, which is a strong assumption that we can weaken to conditional ignorability later. In general, though, it's often difficult to know why we would believe conditional ignorability for the instrument but not for the treatment. Thus, the most plausible instruments are those that are truly randomized.

$$[\{Y_i(a,z), \forall a, z\}, A_i(1), A_i(0)] \perp\!\!\!\perp Z_i$$

- Exclusion restriction: here we put it in a more concrete, explicitly causal form. The instrument has no effect on the outcome, once we fix the value of the treatment. In some sense, the instrument would be completely useless if we had simply randomly assigned the treatment. It has no interesting value to the outcome separate from its effect on the treatment.

$$Y_i(a,1) = Y_i(a,0) \quad \text{for } a = 0,1$$

- Given this exclusion restriction, we know that the potential outcomes for each treatment status only depend on the treatment, not the instrument:

$$Y_i(1) \equiv Y_i(1,1) = Y_i(1,0)$$
$$Y_i(0) \equiv Y_i(0,1) = Y_i(0,0)$$

- Rewriting the usual consistency assumption gives us a linear model with heterogeneous effects (we have seen this before in randomized experiments):

$$\begin{aligned} Y_i &= Y_i(0) + (Y_i(1) - Y_i(0))A_i \\ &= \alpha_0 + \tau_i A_i + \eta_i \end{aligned}$$

- Here, we have $\alpha_0 = E[Y_i(0)]$ and $\tau_i = Y_i(1) - Y_i(0)$.

- First stage: the next assumption is a little mundane, but turns out to be very important. The instrument must have an effect on the treatment. Otherwise, what would we be doing? The instrument wouldn't affect anything. This might seem harmless, but it can wreak havoc on the efficiency of our causal estimates.

$$E[A_i(1) - A_i(0)] \neq 0$$

- Monotonicity. Lastly, we need to make another assumption about the relationship between the instrument and the treatment. Namely, that the presence of the instrument never dissuades someone from taking the treatment (or, alternatively, that the presence of the instrument never encourages someone from taking the treatment).

$$A_i(1) - A_i(0) \geq 0 \quad \text{or} \quad A_i(1) - A_i(0) \leq 0$$

- This is sometimes called "no defiers". It turns out that with a binary treatment and a binary instrument, we can group units into four categories:

| Name | $A_i(1)$ | $A_i(0)$ |
|---|---|---|
| Always Takers | 1 | 1 |
| Never Takers | 0 | 0 |
| Compliers | 1 | 0 |
| Defiers | 0 | 1 |

- The monotonicity assumption remove the possibility of there being defiers in the population. This gives us a lot of information. It means that anyone with $A_i = 1$ when $Z_i = 0$ must be an always-taker and anyone with $A_i = 0$ when $Z_i = 1$ must be a never-taker. We'll see how it factors into into identification of effects.

- It turns out that under these assumptions, we can show that the Wald estimator is equal what we call Local average treatment effect (LATE) or the complier average treatment effect (CATE). This is average effect among the compliers: those that take the treatment when encouraged to do so. That is, the LATE theorem (Angrist and Pischke 1994), states that:

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[A_i|Z_i = 1] - E[A_i|Z_i = 0]} = E[Y_i(1) - Y_i(0)|A_i(1) > A_i(0)]$$

- Under the exclusion restriction, we know that $E[Y_i|Z_i = 1] = E[Y_i(0) + (Y_i(1) - Y_i(0))A_i|Z_i = 1]$, which is then equal to $E[Y_i(0) + (Y_i(1) - Y_i(0))A_i(1)]$ by randomization. The same applies to when $Z_i = 0$, so we have $E[Y_i|Z_i = 0] = E[Y_i(0) + (Y_i(1) - Y_i(0))A_i(0)]$. Thus, we know that

$$\begin{aligned}
E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] =& E[(Y_i(1) - Y_i(0))(A_i(1) - A_i(0))] \\
=& E[(Y_i(1) - Y_i(0))(A_i(1) - A_i(0))|A_i(1) > A_i(0)] \Pr[A_i(1) > A_i(0)] \\
&+ E[(Y_i(1) - Y_i(0))(A_i(1) - A_i(0))|A_i(1) < A_i(0)] \Pr[A_i(1) < A_i(0)] \\
=& E[Y_i(1) - Y_i(0)|A_i(1) > A_i(0)] \Pr[A_i(1) > A_i(0)]
\end{aligned}$$

- The third equality comes from monotonicity: with this assumption, $A_i(1) < A_i(0)$ never occurs. We can use the same argument for the denominator:

$$E[A_i|Z_i = 1] - E[A_i|Z_i = 0] = E[A_i(1) - A_i(0)] = \Pr[A_i(1) > A_i(0)]$$

- Dividing these two expressions through gives the LATE.

### Is the LATE useful?

- A couple of things to note about what we have shown. Basically, once we allow for heterogeneous effects, all we can estimate with IV is the effect of treatment among compliers. Note that this is a unknown subset of the data. Among treated units with $Z_i = 1$, we cannot distinguish them from the always-takers and similarly for the control units with $Z_i = 0$.

- For instance, we can show that, generally the ATT and the LATE differ:

$$\begin{aligned}
E[Y_i(1) - Y_i(0)|A_i = 1] =& E[Y_i(1) - Y_i(0)|A_i = 1, A_i(0) = 1] \Pr[A_i(0) = 1|A_i = 1] \\
&+ E[Y_i(1) - Y_i(0)|A_i(1) > A_i(0), A_i = 1] \Pr[A_i(1) > A_i(0)|A_i = 1]
\end{aligned}$$

- Without further assumptions, this estimand is not equal to overall treatment effect or the treatment effect on the treated. Furthermore, since the complier group depends on the instrument, an IV estimate with one instrument will generally estimate a different quantity than an IV estimate of the **same effect** with a different instrument.

- But it's also true that this is the **only** causal effect of $A_i$ on $Y_i$ that we can identify given the above assumptions. This leads to the title of Imbens's paper: better LATE than nothing.

- In general, the best interpretation of the LATE estimate is that it might have weaker external validity. It's unclear if we were to intervene and actually randomly assign $A_i$, we would get a similar result because the LATE might be very different than the ATE. This is the thrust of the paper by Deaton on the syllabus in his skepticism.
- We can derive bounds for the average treatment effect in this setting, but those bounds tend to be quite wide. In general, though, it is good to calculate such bounds to give a sense of what is happening in the data.

### Randomized trials with one-sided compliance

- It turns out that with additional assumptions, we can get the LATE to be equal to a parameter of interest: the ATT. In general, this is most plausible in a specific setting: randomized control trials with one-sided compliance.
- Note that we can think of a randomized experiment with issues of compliance as the type of situation that is ideal for IV. We have a randomized instrument (the treatment assignment) and we have a non-randomized treatment affected by the instrument (the treatment actually taken). Here, we get randomization by design,
- In this siutation, it might be plausible to make an additional assumption of **one-sided compliance**. This means that compliance problems can only come from one direction. That is, we might have $\Pr[A_i = 1|Z_i = 0] = 0$ because no one that was randomly assigned to control $(Z_i = 0)$ has access to the treatment. Maybe this is because only those treated actually get pills or only they are invited to the job training location.
- With this assumption, we know that there are no "always-takers" and since there are no defiers, then we know that anyone in the treated $(Z_i = 1)$ group that takes the treatment $(A_i = 1)$ is a complier. Thus, we know that:

$$\begin{aligned}
E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] &= E[Y_i(0) + (Y_i(1) - Y_i(0))A_i|Z_i = 1] - E[Y_i(0)|Z_i = 0] \\
&= E[Y_i(0)|Z_i = 1] + E[(Y_i(1) - Y_i(0))A_i|Z_i = 1] - E[Y_i(0)|Z_i = 0] \\
&= E[Y_i(0)] + E[(Y_i(1) - Y_i(0))A_i|Z_i = 1] - E[Y_i(0)] \\
&= E[(Y_i(1) - Y_i(0))A_i|Z_i = 1] \\
&= E[Y_i(1) - Y_i(0)|A_i = 1, Z_i = 1] \Pr[A_i = 1|Z_i = 1] \\
&= E[Y_i(1) - Y_i(0)|A_i = 1] \Pr[A_i = 1|Z_i = 1]
\end{aligned}$$

- The first equality comes from the exclusion restriction (we used this before) and from the the fact that under one-sided compliance, $Z_i = 0$ implies that $A_i = 0$. The second equality comes from the linearity of expectations. The third comes from the randomization of the instrument. The fourth is just algebra. The fifth comes from the fact that the treatment is binary. The last comes from the fact that $A_i = 1$ implies that $Z_i = 1$ because only those that were randomized to take treatment can take treatment. Thus, it's clear that we have:

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{\Pr[A_i = 1|Z_i = 1]} = E[Y_i(1) - Y_i(0)|A_i = 1]$$

- Thus, under the additional assumption of one-sided compliance, we can estimate the ATT. Again, the reason is quite simple: before we showed that the ATT is a combination of the LATE and the effect for the always-takers. If we remove the possibility of the always takers, then anyone who actually takes the treatment is a complier. Not only this, but we can identify that group and learn about their characteristics.
- It's also easy to see that if we switch the direction of one-sided compliance, then we can esimate the average treatment effect for the controls.

### Size, characteristics of the compliers

- While we cannot identify who is a complier and who is not a complier in general, we can estimate the size of the complier group:

$$\Pr[A_i(1) > A_i(0)] = E[A_i(1) - A_i(0)] = E[A_i|Z_i = 1] - E[A_i|Z_i = 0]$$

- Angrist and Pischke describe ways to calculate the difference between the compliers and overall population in terms of binary covariates. Abadie (2003) shows how to calculate the mean of any covariate in the complier group.

### Multiple instruments

- Again, since each instrument implies a different complier group, each instrument estimates a causal effect for a different subset of the population. Thus, if we had two instrument, then there would be two different LATEs, $\rho_1$ and $\rho_2$ for instruments $Z_{1i}$ and $Z_{2i}$. We might try to use 2SLS to estimate an overall effect with these instruments with following first stage:

$$\hat{A}_i = \pi_1 Z_{1i} + \pi_2 Z_{2i}.$$

- In Angrist and Pischke, they show that the 2SLS estimator using these two instruments is a weighted sum of the two component LATEs:

$$\rho_{2SLS} = \psi\rho_1 + (1 - \psi)\rho_2,$$

where the weights are:

$$\psi = \frac{\pi_1 \text{Cov}(A_i, Z_{1i})}{\pi_1 \text{Cov}(A_i, Z_{1i}) + \pi_2 \text{Cov}(A_i, Z_{2i})}$$

- Thus, the 2SLS estimate is a weighted average of causal effects for each instrument, where the weights are related to the strenght of prediction for each of the first stage effects of the instruments.

### Covariates and heterogeneous effects

- It might be the case that the above assumptions only hold conditional on some covariates, $X_i$. That is, instead of randomization, we might have conditional ignorability:

$$[\{Y_i(a,z), \forall a, z\}, A_i(1), A_i(0)] \perp\!\!\!\perp Z_i | X_i$$

- We would also have exclusion conditional on the covariates:

$$\Pr[Y_i(a,0) = Y_i(a,1) | X_i] = 1 \quad \text{for } a = 1, 0$$

- Under these assumptions, Angrist and Pischke show that if you fully saturate the first stage and the second stage in the covariates, then 2SLS estimates a weighted average of the covariates-specific LATEs (very similar to regression).

- Abadie (2003) shows how to estimate the overall LATE using a weighting approach based on a "propensity score" for the instrument.