ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Selection On Unobservables Designs: Part 4, Instrumental Variables

To date we have studied selection on observables designs and a single selection on unobservables design: panel data with fixed effects or differences-in-differences. In the panel data models, we assume that any unobservable determinants of Y_{it} that are correlated with treatment assignments are constant over time (and thus get differenced out or absorbed by the fixed effects). This assumption often seems questionable, however – changes within individuals or cross-sectional units do not necessarily occur at random. To address this possibility, we discussed synthetic control methods that essentially combine matching with diffs-in-diffs. Note that this combination basically put us back in the selection on observables world – we essentially assume that we have enough observable characteristics to construct a synthetic control unit whose trajectory of Y_{it} would match the treated unit's trajectory of Y_{it} (absent any treatment effect).

We now turn to a true selection on unobservables design – the instrumental variables (IV) estimator. IV methods are a cornerstone of econometrics – these methods date back to the work of Tinbergen and Haavelmo in the 1930s and 1940s.[1] Our understanding of IV methods advanced significantly during the 1990s, however, with seminal work on IV in the context of treatment effect heterogeneity and IV methods in the case of a large number of weak instruments. For the purposes of these notes, I will use the phrase "IV methods" to refer generally to methods using instrumental variables, including IV, two stage least squares (2SLS), and limited information maximum likelihood (LIML).

[1] The late Saddam Hussein is reputed to have called IV "the mother of all unobservables designs."

1 Basic IV biased but consistent

Consider a model of the form

$$y_i = \beta_0 + \beta_1 d_i + \varepsilon_i \tag{1}$$

I will sometimes refer to this equation as the "structural equation." At this point we are not assuming that d_i is binary – it may have more than two points of support, or it may be continuous. The standard condition that we need for a linear regression of y_i on d_i to consistently estimate β_1 is $\text{Cov}(d_i, \varepsilon_i) = 0$. This will be true if d_i is randomly assigned, and it could be true in other situations as well. In general, however, it will not be true.

An alternative way to estimate β_1 is via instrumental variables. The goal in IV is to find some subset of the variation in d_i, call it z_i, that is uncorrelated with ε_i (i.e., as good as randomly assigned). Formally, our goal is to find an instrument z_i, not in equation (1), that satisfies the following two properties:

1. $\text{Cov}(z_i, d_i) \neq 0$ **Relevance**
2. $\text{Cov}(z_i, \varepsilon_i) = 0$ **Exclusion Restriction**

The first assumption ensures that z_i actually captures some of the variation in d_i. If it doesn't, then it will be of no use to us in estimating the effect of d_i on y_i. The second assumption ensures that z_i is uncorrelated with ε_i (obviously). This assumption is often referred to as the "exclusion restriction" because it implies that the instrument, z_i, can be excluded from equation (1). If z_i were correlated with ε_i, we would want to include it as a covariate (given that it's also correlated with d_i by our first assumption). This would violate our condition that z_i not be in equation (1).

To fix ideas, let us consider an application. Suppose that we would like to estimate the causal effect of schooling on earnings. One way to estimate this effect would be to regress

earnings on schooling and a bunch of other covariates. Is it plausible that the variation in schooling is uncorrelated with everything else that affects earnings, such as unobserved ability? Probably not, even after conditioning on covariates. We generally think that on average people who go to college are different in fundamental ways from people who only complete high school, and specifically we believe that the two groups are probably different in ways that affect earnings (e.g. perhaps the college graduates have higher innate "ability").

An alternative way to estimate the effect of schooling on earnings is to find an instrument for schooling that satisfies the criteria above. Suppose that we are lucky and we learn that many years ago the state government of California ("the greatest state in the Union," according to my jury duty briefing) decided to start giving out free scholarships to U.C. schools in order to promote higher education. However, because the government has a limited budget, they could not offer the scholarships to everyone, so in the interest of fairness they decided to hold a lottery in which every family in the state was automatically entered, and the scholarships were randomly assigned to lucky families who won the lottery. (Again, to my knowledge this has not actually happened – it's just a hypothetical example to make things more concrete.)

If we let z_i be a dummy variable that is 1 if a family wins the scholarship lottery and 0 otherwise, then z_i is a promising instrument for schooling. We know that the first condition, $\text{Cov}(z_i, d_i) \neq 0$, will be satisfied because people who win the free scholarships will be more likely to attend college than those who don't, inducing a positive correlation between z_i and d_i. We also know that the second condition for a good instrument, $\text{Cov}(z_i, \varepsilon_i) = 0$, is likely to be satisfied because the winners of the lottery were randomly chosen by the state. By definition, no characteristic, other than those characteristics directly affected by the lottery, can possibly be correlated with whether or not a family won the lottery.[2]

The IV estimator is:

both instr. and x_s — both D and x_s

$$\hat{\beta}_{IV} = (Z'D)^{-1}(Z'Y)$$

[2]I give an example in Section 2 of another treatment, besides schooling, that the lottery might be affecting. This would be a violation of the exclusion restriction.

In the general case, Z could contain not only the instrument z_i (the lottery number), but also predetermined covariates x_i (gender, race, parental education, etc.). D would then contain both the treatment of interest, d_i, and the predetermined covariates x_i. In the case in which there are no covariates, we can write $\hat{\beta}_{IV} = \text{Cov}(z_i, y_i)/\text{Cov}(z_i, d_i)$.

It is straightforward to show that $\hat{\beta}_{IV}$ is a consistent estimator of β_1 given the assumptions above.

$$\text{plim}(\hat{\beta}_{IV}) = \text{plim}[(Z'D)^{-1}(Z'Y)]$$

$$= \text{plim}[(Z'D)^{-1}(Z'D\beta + Z'\varepsilon)]$$

$$= \text{plim}[(Z'D)^{-1}(Z'D)\beta] + \text{plim}[\frac{1}{N}(Z'D)^{-1}]\text{plim}[\frac{1}{N}(Z'\varepsilon)]$$

$$= \beta$$

This formal derivation, however, gives limited intuition regarding why or how IV operates. For intuition, we will turn to alternative methods of implementing the IV estimator.

2 The Reduced Forms and 2SLS

The most popular way to implement the IV estimator is via a two stage procedure known as two stage least squares (2SLS). If we have one instrument and one variable that we want to instrument for, 2SLS and IV are the exact same thing (in this case we would say that we are "exactly identified"). IV is thus a special case of 2SLS – you can always use 2SLS in any scenario in which you can use IV, though the reverse is not true. We begin by writing out the two stages of 2SLS, and then consider what is going on:

1. *First Stage* We first estimate a regression of d_i (the variable that we want to instrument for – e.g., schooling, in our hypothetical example) on the instrument, z_i (e.g., the lottery

number), and all of the predetermined covariates, x_i. This regression looks like:

$$d_i = \gamma_1 z_i + x_i \gamma_2 + u_i$$

where z_i and γ_1 are scalars, x_i is a $1 \times K+1$ vector that includes all covariates and a 1, and u_i is a residual term. Take the predicted values of d_i (e.g., predicted schooling, $\hat{d}_i = \hat{\gamma}_1 z_i + x_i \hat{\gamma}_2$) from this regression and use them in place of the actual values of d_i in the second stage.

2. *Second Stage* In the second stage, we run the regression that we originally wanted to estimate, but instead of including the variable that we want to instrument for (d_i), we include its predicted values from the first stage ($\hat{d}_i$). In our example, instead of running earnings on schooling and other covariates, we would run earnings on predicted schooling (from the first stage) and other covariates. Thus the regression looks like:

$$y_i = \beta_0 + \beta_1 \hat{d}_i + x_i \beta_2 + \varepsilon_i$$

The estimate of β_1 from this regression will be consistent.

Note that both the first and second stages always contain the same set of covariates (you can't exclude certain covariates from the first stage and then include them in the second stage, and you can't exclude covariates from the second stage and include them in the first stage, unless you intend to use them as instruments). In matrices, define Z to be a matrix that includes the instrument (z_i) and the predetermined covariates (x_i). D is a matrix that includes the treatment (d_i) and the predetermined covariates (x_i). Then $\hat{\beta}_{2SLS} = (D' P_Z D)^{-1}(D' P_Z Y)$, where $P_Z = Z(Z'Z)^{-1}Z'$.

Now that we have introduced the first stage and the second stage, we are almost done, but before we move on to the next section I will introduce the reduced form equation. Technically the term "reduced form" refers to any regression which regresses an endogenous variable (i.e., a not-exogenous variable; in our case y_i and d_i are our two endogenous variables) on all of

the exogenous variables (z_i and x_i). So, if you consider the two regressions that we estimated above, you will see that the first stage is in fact a reduced form equation. However, in general I will use the term "reduced form" to refer specifically to the reduced form equation that regresses y_i on all of the exogenous variables (z_i and x_i). So the reduced form in our example is:

Reduced form Regression of IV:

$$y_i = \pi_1 z_i + x_i \pi_2 + v_i$$

What does the reduced form measure? The reduced form measures the causal effect of the instrument (z_i) on the outcome variable (y_i). In our example, the coefficient that we get from running the reduced form gives us an estimate of the effect on earnings of winning the scholarship lottery. Note that if z_i is a good instrument, then the causal effect should run only through the variable that is being instrumented for (d_i). In our example, that means that winning the schooling lottery should raise your income only because it encourages you to get more schooling on average, not for some other reason (e.g., because the parents of lottery winners used the money they saved on college tuition to pay for additional private tutoring for their children).

So there are three equations we want to keep in mind:

1. The first stage, which regresses the variable we're instrumenting for on the instrument(s) and the other exogenous variables. This predicts how the variable we're instrumenting for changes as our instrument changes.

2. The second stage, which regresses y_i on the predicted values from the first stage and the other exogenous variables. This gives us our IV estimate of β_1.

3. The reduced form, which regresses y_i on the instrument and the other exogenous variables. This measures how y_i changes as we change our instrument z_i. Note that we never have to run the reduced form in the 2SLS procedure, but as you will see in the next section, it is a useful concept to keep in mind.

3 IV Intuition

At this point, some might ask, "Why not just run the reduced form? Why bother with IV (2SLS) at all? After all, the reduced form gives unbiased predictions, and it's much less complex than this two stage procedure." In other words, why not simply replace the variable we want to instrument for (d_i) with the instrument (z_i)? Actually, this isn't necessarily a bad idea. As Josh Angrist says, "Many papers would do well to stop with the reduced form." The reduced form makes explicit exactly where the identification in the research design is coming from, and it does not suffer from some of the "weak instruments" issues that we will discuss later. Any time you are dealing with a single instrument, it's a good idea to estimate the reduced form and check whether it conforms to your expectations, even if you don't put it into the paper.

The answer to the question above, however, is that we usually are not interested in measuring the effect of z_i on y_i, which is what the reduced form gives us. Instead, we are interested in measuring the effect of d_i on y_i. That is what IV gives us. In our example, we are interested in measuring the effect of schooling on wages, so we run IV. If we just ran the reduced form, we would get the effect of winning the scholarship lottery on wages. While that may be of some policy interest in evaluating the scholarship program, it is not what we are looking for.

From a linear algebra perspective, 2SLS/IV estimates β by first projecting all of the data onto the subspace spanned by Z – all of the exogenous variables in the regression (i.e., the instrument and the predetermined covariates) – and then running the regression of y_i on d_i and x_i after they have been projected onto this subspace. In this sense it should be clear that we are only using the "good" variation in d_i (i.e., the variation in d_i that comes from z_i) to estimate β_1. However, in the case in which you have a single instrument (which is all we have discussed so far), there is an even cleaner interpretation.[3]

[3]If you are "overidentified," i.e. you have more instruments than you need, then this interpretation does not hold anymore, though it is still conceptually useful. In our example, we are "just identified" (one variable to instrument for, i.e. schooling, and one instrument, i.e. the scholarship lottery), so you can apply

In the case of one treatment and one instrument, the estimate of β_1 that we get from IV equals the reduced form coefficient rescaled by the first stage coefficient. That is to say:

$$\hat{\beta}_{1IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \frac{cov(z,y)/var(z)}{cov(z,d)/var(z)}$$

What this shows is that the IV estimate is very closely related to the reduced form estimate – in fact, it's exactly proportional to the reduced form estimate. Why is this a useful formulation? Well, consider what each coefficient means.

In our example, the reduced form coefficient ($\hat{\pi}_1$) measures the effect of winning the scholarship lottery on earnings. But that is not what we want; what we want is the effect of an additional year of schooling on earnings. Because the scholarship lottery only affects earnings due to its effect on increasing schooling (or so we're assuming), the reduced form coefficient represents the effect of an unknown additional amount of schooling on earnings. The problem is that we don't have the units right. If we knew that everyone who won the scholarship lottery got, on average, one more year of school than they otherwise would have, then we could interpret the reduced form coefficient as the causal effect of one more year of schooling on earnings. Why? Well, remember that because the instrument z_i is randomly assigned (i.e. winners are randomly picked), the winners and the losers are on average comparable in every way except that the winners get, on average, one more year of schooling than the losers. So any difference in earnings between the winners (i.e. those with $z_i = 1$) and the losers (i.e. those with $z_i = 0$) must be due to the extra year of schooling. Thus the coefficient on z_i in the reduced form ($\hat{\pi}_1$) is the effect of one more year of schooling on earnings.

In general, however, it is unlikely that the scholarship lottery winners get, on average, exactly one more year of schooling than the losers. So how do we rescale the reduced form coefficient so that we get the units right? The answer is that we divide through by the first stage coefficient, $\hat{\gamma}_1$. Why does this work? Consider our specific example. The first stage estimates the effect of winning the scholarship lottery on years of education. So the

the interpretation that I'm about to give.

first stage coefficient, $\hat{\gamma}_1$, tells you how much, on average, your years of schooling increase if you win the scholarship lottery. So suppose that $\hat{\gamma}_1 = 0.5$, i.e. that those who win the scholarship lottery get half a year more of education on average than those who do not win the lottery. Also suppose that $\hat{\pi}_1 = 500$, i.e., those who win the scholarship lottery earn \$500 more per year on average than those who do not win the scholarship lottery. Then we know that the people winning the lottery are earning \$500 more because they have extra schooling (this comes from the reduced form). And we know that they are on average getting an extra 0.5 years of schooling when they win the lottery. So what is the return to one additional year of schooling? It is \$500/0.5 (the change in earnings divided by the change in schooling), or \$1000. In other words, our estimate of the effect of schooling on earnings is $\hat{\beta}_{IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1}$. So the reduced form coefficient represents the causal effect of some additional amount of schooling on earnings (how much additional schooling is unknown until we see the first stage), and the first stage coefficient rescales that coefficient appropriately to reflect the amount of extra schooling that the instrument (the scholarship lottery) generates.

So far you have taken it on faith that the formula $\hat{\beta}_{IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1}$ is actually true. But it is actually simple to prove. Recall that $\hat{\beta}_{IV} = (Z'D)^{-1}(Z'Y)$. If you accept that we can apply partitioned regression to IV just like we can with OLS, then it is trivial to transform the formula for $\hat{\beta}_{IV}$ into one in which Z and D are always vectors.[4] If Z contains covariates X, simply redefine Z such that $\tilde{Z} = M_X Z_1$, where Z_1 is a column vector containing only the instrument and M_X is the orthogonal projection matrix for the covariates, $M_X = I - X(X'X)^{-1}X'$ (X is an $N \times K + 1$ matrix containing all covariates and a column of ones).[5] Thus we can always write $\hat{\beta}_{1IV}$ as

$$\hat{\beta}_{1IV} = (\tilde{Z}_1' D_1)^{-1}(\tilde{Z}_1' Y) = \text{Cov}(\tilde{z}_i, y_i)/\text{Cov}(\tilde{z}_i, d_i)$$

Now consider $\hat{\pi}_1$ and $\hat{\gamma}_1$. The former comes from a regression of y_i on $\tilde{z}_i$, so $\hat{\pi}_1 = \text{Cov}(\tilde{z}_i, y_i)/\text{Cov}(\tilde{z}_i, z_i)$. The latter comes from a regression of d_i on $\tilde{z}_i$, so $\hat{\gamma}_1 = \text{Cov}(\tilde{z}_i, d_i)/\text{Cov}(\tilde{z}_i, z_i)$.

[4] That partitioned regression works for the 2SLS procedure should be fairly obvious. It is less self-evident that partitioning must work for the IV formula as well.

[5] Also define the column vector D_1 such that D_1 contains only the treatment, d_i.

Thus

$$\hat{\pi}_1/\hat{\gamma}_1 = \frac{\text{Cov}(\tilde{z}_i, y_i)/\text{Cov}(\tilde{z}_i, z_i)}{\text{Cov}(\tilde{z}_i, d_i)/\text{Cov}(\tilde{z}_i, z_i)} = \hat{\beta}_{1IV}$$

The takeaway of all of this is that, when working with an IV estimator, the entire experiment is in the reduced form. The reduced form measures the causal impact of the instrument on the outcome – the first stage exists only to rescale that estimate and "get the units right." Thus, when applying IV, you should always consider the underlying reduced form that you are running and ascertain whether it makes sense and whether it is identifying the causal effect in the manner that you originally imagined.

4 Multiple Instruments

It's often very difficult to find one good instrument, let alone two or more good instruments. Nevertheless, in some cases a single conceptual instrument will be parameterized though multiple variables (we will see an example of this in the next section). In those cases, we say that the equation is "overidentified," in the sense that we have more instruments than we need. It's impossible to incorporate more than one instrument into the IV estimator because $\hat{\beta}_{IV} = (Z'D)^{-1}(Z'Y)$; Z and D must have the same number of columns, or else the first half of $\hat{\beta}_{IV}$ won't be conformable with the second half. One option would be to simply pick one instrument and discard the rest, but this seems undesirable from an efficiency standpoint because you're throwing away valid information for estimating β. An attractive alternative then is to use 2SLS, which can trivially accommodate more than one instrument.

In the two stage procedure, simply include all instruments in the first stage when you predict the value of d_i. For example, if you have two instruments, z_{1i} and z_{2i}, estimate the first stage as:

$$d_i = \gamma_1 z_{1i} + \gamma_2 z_{2i} + x_i\gamma_3 + u_i \implies \hat{d}_i$$

Then use $\hat{d}_i$ as the regressor in the second stage instead of d_i. In matrices, the formula remains the same: $\hat{\beta}_{2SLS} = (D'P_Z D)^{-1}(D'P_Z Y)$. Now Z contains more columns than D, but that doesn't affect the conformability of P_Z (which is an $N \times N$ matrix) with D. Under Gauss-Markov type assumptions, 2SLS efficiently combines all of the instruments to estimate β.

5 Applications

We now consider two important applications of instrumental variables. These applications are particularly helpful when studying IV in the context of heterogeneous treatment effects and the "weak instruments" issue (both of which we will cover).

5.1 Medical Trials

For a variety of reasons, medical trials are a fantastic example of an application of instrumental variables – I would argue the best, in fact. First of all, they are socially important (perhaps the most important application of IV to date). Furthermore, they are very clean in terms of experimental design, so they make a great teaching example for conveying the intuition behind what the IV estimator is doing. My personal recommendation would be to use this example whenever possible to guide you in understanding how IV operates.

The model for a medical trial is the same simple regression model that we are accustomed to: $y_i = \beta_0 + \beta_1 d_i + \varepsilon_i$. In this case, y_i represents a medical outcome, which could either be a continuous variable such as blood pressure or cholesterol level or a discrete variable such as whether or not you survive (e.g., 1 if you survive, 0 if you do not). The variable d_i is generally a dummy variable that is 1 if you receive the treatment and 0 if you do not. It could alternatively be continuous (for example, it could be the dosage in milligrams of the drug that you receive), but in this example we will assume it is binary (you either take the pill or you do not take the pill). The error term ε_i represents all other factors that affect the health outcome. Note that the regression model corresponds to the potential outcomes

$y_i = \beta_0 + \beta_1 d_i + \varepsilon_i$

Potential Outcomes Notation: $y = d \cdot y_1 + (1-d) y_0$, $y_0 = \beta_0 + \varepsilon$

$y_1 = y_0 + \beta_1$

y_i = blood pressure, d_i = blood pressure pill

Even if we have double-blind experiment, other issues:

- attrition
- non-compliance — "treated" don't take pills

↳ reg $y = \beta_0 + \beta_1 d_i + \varepsilon$ is an "intention to treat" analysis

fix: IV!

Setup
- Clinical trial w/ 50% compliance in treatment group

FS: $d_i = \gamma_0 + \gamma_1 z_i + u_i$ $\quad \gamma_1 = 0.5$

(d_i: took pill; z_i: treatment assignment)

RF: $y_i = \pi_0 + \pi_1 z_i + v_i$ (group assignment -or- increasing the P(taking pill) by 50%)

$$\hat{\beta}_{IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \frac{\hat{\pi}_1}{0.5}$$

- a naive comparison of means of control vs. complier treated creates selection bias issues
- Reduced form of IV uses full sample (doesn't throw out noncompliance treat obs) so there is no selection bias issue

? really?

model with constant treatment effects ($y = dy_1 + (1-d)y_0,\ y_0 = \beta_0 + \varepsilon,\ y_1 = y_0 + \beta_1$).

At this point I will switch to a specific example in order to make the discussion clearer. Let y_i be blood pressure, and let d_i represent a pill that is designed to treat high blood pressure, so $d_i = 1$ if individual i takes the pill and $d_i = 0$ if individual i does not take the pill. Our goal is to estimate the effect that the pill has on lowering blood pressure – our hope is that β_1 is large and negative. One way to estimate the effect is to start selling the drug to the general population and then collect some data and run a regression of blood pressure on whether or not you take the pill. However, this estimate will clearly suffer from a selection issue – people who take the pill are the ones who have high blood pressure to begin with! We will likely get a positive estimate of β_1 from this procedure, even if the true β_1 is large and negative. This may be true even after we condition on observable covariates using one of the selection on observables designs we discussed earlier. Therefore, in order to accurately estimate β_1, we design a medical trial in which we randomly assign some patients to the treatment group and assign other patients to the control group. The patients assigned to the treatment group are then given the pill and told to take it, while the patients assigned to the control group are given a placebo (or nothing at all).

Back in the old days (perhaps even older than me), people estimated the effect of the drug by simply subtracting the mean of y_i for the control group from the mean of y_i for the treatment group (in other words, regressing y_i on a variable that is 1 if you are in the treatment group and 0 if you are in the control group). This is what is known as an "intention to treat" analysis, because you are taking the difference between the group that you intend to treat and the group that you do not intend to treat. But there was the problem of "non-compliance" – some people in the treatment group would fail to take the pill and others in the control group would obtain the pill from another source, even though they were not supposed to. This non-compliance can cause a bias in the estimate of β_1, and it was not immediately clear how to fix this bias until it became obvious that what we were looking at was actually a simple IV problem.

In this case, the instrument z_i is the intention to treat, i.e. $z_i = 1$ if you are assigned

to treatment group (we intend to treat you), and $z_i = 0$ if you are assigned to the control group (we do not intend to treat you). It is easy to see that z_i satisfies the two properties of a good instrument. First of all, z_i is uncorrelated with ε_i by construction, because whether you are assigned to the treatment group or the control group is randomly determined, so $\text{Cov}(z_i, \varepsilon_i) = 0$. Second, z_i is correlated with d_i, because you are going to be more likely to take the pill if you are in the treatment group, so $\text{Cov}(z_i, d_i) \neq 0$. Therefore, z_i is a valid instrument for d_i, and the IV estimator gives us a consistent estimate of β_1, the effect of taking the pill on blood pressure.

How does this fix the non-compliance problem that we discussed before? To facilitate understanding, assume that the non-compliance problem only exists for the people in the treatment group. That is to say, assume that nobody in the control group takes the pill, but also assume that only half the people in the treatment group take the pill (i.e., half of the treatment group fails to comply and does not take the pill, while the other half takes the pill, as they were supposed to). What will the IV estimate look like?

The first stage will regress d_i on z_i, i.e. regress whether you took the pill on whether you were in the treatment group. So the first stage is:

$$d_i = \gamma_1 z_i + u_i$$

Since zero people in the control group took the pill while half the people in the treatment group took the pill, it should be intuitively clear that our estimate for γ_1 will be 0.5 (being in the treatment group raises your probability of taking the pill by 50 percentage points, so $\hat{\gamma}_1 = 0.5$).

Now recall that the IV estimate is the reduced form rescaled by the first stage. In this case, the reduced form is a regression of y_i (your blood pressure) on z_i (whether you were assigned to the treatment or control group). So the reduced form is:

$$y_i = \pi_1 z_i + v_i$$

Therefore, our IV estimate is $\hat{\beta}_{1IV} = \hat{\pi}_1/\hat{\gamma}_1 = \hat{\pi}_1/0.5$. How is this fixing the non-complier problem? Well, we know that the reduced form estimates the causal effect of the instrument on y_i, so in our case the reduced form is estimating the effect that being assigned to the treatment group has on blood pressure. If there were a perfect correlation between being assigned to the treatment group and taking the pill (i.e. everyone in the treatment group took the pill, and nobody in the control group took the pill), then the reduced form estimate would be the effect of taking the pill on blood pressure. In that case the first stage would give us $\hat{\gamma}_1 = 1$, and the IV would be $\hat{\beta}_{1IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \hat{\pi}_1$. In other words, the IV would be the same as the reduced form (which is what we would expect, since both are supposed to be estimating the same thing in this case, i.e., the effect of the pill on blood pressure).

In our example, however, there is not a perfect correlation between being assigned to the treatment group and taking the pill, which is why our first stage estimate is $\hat{\gamma}_1 = 0.5$, not $\hat{\gamma}_1 = 1$. So in our case, the reduced form is estimating the effect on your blood pressure of increasing the probability that you take the pill by 50 percentage points. This means that the reduced form is not going to be estimating the full effect of taking the pill. Instead, it's estimating half of the effect of taking the pill. If it helps, imagine that there are 10 people in the treatment group, 5 of whom take the pill and 5 of whom do not, and 10 people in the control group, 0 of whom take the pill. The (expected) mean blood pressure for the treatment group will be $\frac{5\cdot\beta_0+5\cdot(\beta_0+\beta_1)}{10} = \beta_0 + \frac{\beta_1}{2}$, while the (expected) mean blood pressure for the control group will just be β_0. So the reduced form coefficient, $\hat{\pi}_1$, will be the difference of means between the treatment and control groups, or $\frac{\beta_1}{2}$. This is, of course, half the effect of taking the pill.

Therefore, the (plim of the) IV estimate will be $\beta_{1IV} = \frac{\pi_1}{\gamma_1} = \frac{0.5\beta_1}{0.5} = \beta_1$, which is exactly what we want. We can see that the IV estimate gives us a consistent estimate precisely because it is rescaling the reduced form by the first stage. In our example what this means in practice is that we are rescaling the reduced form to account for the fact that being in the treatment group only increases your probability of taking the pill by 50 percentage points, not by a full 100 percentage points. So the reduced form only represents half the effect of

taking the pill, and it must be rescaled by (divided by) 0.5 in order to estimate the full effect of taking the pill.

More generally, what this example demonstrates is that IV functions by taking the estimated causal effect of z_i on y_i (the reduced form) and rescaling it by the estimated causal effect of z_i on d_i (the first stage).

Before we move on, I should note how IV is different than simply taking the mean of y_i for the people in the treatment group who took the pill and subtracting the mean of y_i for the people in the control group who did not take the pill (which, in our example, is the entire control group). The estimator I just described, which I will refer to as the naïve estimator, is affected by the same selection issues as a simple OLS regression of y_i on d_i. Specifically, it may be the case that the people in the treatment group who choose not to take the pill do so because their blood pressure was not very high to begin with. Thus the group of people that actually took the pill are the ones that all had high blood pressure to begin with, and we will tend to estimate that the pill does not have much of an effect (because its downward effect is being counteracted by the fact that the people who select to take it all had high blood pressure to begin with).

The IV estimator does not suffer from this selection problem because it does not release the people in the treatment group who choose not to take the pill. To understand this, imagine for the moment that there are two types of people in our sample: high blood pressure types and low blood pressure types. Assume that they occur with equal frequency, so that when we randomly assign our sample to the treatment and control groups, half of the treatment group is high blood pressure, half of the treatment group is low blood pressure, half of the control group is high blood pressure, and half of the control group is low blood pressure. The half of the treatment group that takes the pill all have high blood pressure, so when we apply the naïve estimator and compare their average blood pressure to the average blood pressure of the control group, we underestimate the effect of the pill because we are comparing a group of high blood pressure people (who took the pill) to a group that is a 50/50 mix of high blood pressure and low blood pressure people (who did not take the pill). In

contrast, what IV does is compare the mean of the treatment group (which is half high blood pressure people and half low blood pressure people) to the mean of the control group (which is half high blood pressure people and half low blood pressure people) in the reduced form. It then rescales this difference in means by the first stage to account for the fact that not all of the treated group took the pill. So unlike the naïve estimator, which deceptively compares a high blood pressure group to a half-high/half-low blood pressure group, IV compares two comparable groups, and that is why it gives us a consistent estimate of the effect of the pill.

5.2 Quarter of Birth

The quarter of birth application is perhaps the most-studied example of IV in the economics literature. This example is taken from Angrist and Krueger (1991). I will discuss the basic framework and idea; for more details see the article itself. The purpose is to demonstrate a nice application of IV/2SLS (which may help you think about what a good instrument looks like) and to familiarize you with the canonical example used in the "weak instruments" literature.

The question addressed with this instrument is a familiar one: what is the return to an additional year of schooling? One way to answer this question is to run a standard regression, $y_i = \beta_0 + \beta_1 d_i + x_i\beta_2 + \varepsilon_i$, where y_i is log wages, d_i is years of school, and x_i is a vector of covariates. However, as we know, this regression is likely to give us a biased estimate of β_1 for a variety of reasons, including selection bias and measurement error. The problem is that $\text{Cov}(d_i, \varepsilon_i) \neq 0$; one way to address this problem is to find an instrument z_i that is correlated with d_i (schooling) but uncorrelated with ε_i.

Angrist and Krueger suggest quarter of birth as the instrument. Why use this as an instrument for schooling? The idea is that states have mandatory schooling laws stipulating that students must stay in school until a given age (say age 16, for simplicity). However, the key thing is that these laws dictate the *age* at which a student may leave school, not how many *years of schooling* a student must get. Therefore, if a student starts school at age

Quarter of Birth QoB

FS: $d_i = \gamma_0 + \gamma_1 \overset{Q_1}{z}_{1i} + \gamma_2 \overset{Q_2}{z}_{2i} + \gamma_3 \overset{Q_3}{z}_{3i} + u_i$

SS: $y_i = \beta_0 + \beta_1 \hat{d}_i + \varepsilon_i$

① A & K stated w/ IV only comparing 1st Qu. w/ 2, 3, & 4] used Wald estimator

② Then used dozens of instruments by interacting quarters w/ state & decade dummies

Many weak instruments

- overfitting 1st stage w/ too many weak instruments
  - ↳ FS fits d well and SS just recovers OLS estimate instead of IV

6, she will be legally required to receive 10 years of schooling. However, if she starts school at age 5, she will be legally required to receive 11 years of schooling. Thus, variations in the age at which a student starts school will result in variations in the amount of schooling that student is legally required to receive. While this will not make a difference for most people (because most people do not drop out of high school as soon as they are no longer required to be there), it will make a difference for some people, so there should be a nonzero correlation (albeit a modest one) between the age one starts school and how many years of schooling one receives.

How does this all pertain to quarter of birth? The quarter in which a student is born can have a large effect on what age the student starts school because the academic calendar begins in September regardless of quarter of birth. Many states require children to start school in the calendar year in which they turn 6. So, for example, a child born in December (fourth quarter) might start school at age 5.7, and thus be required by law to receive a minimum of 10.3 years of schooling (16 minus 5.7). However, a child born in January (first quarter) might start school at age 6.7, and thus be required by law to receive a minimum of only 9.3 years of schooling (16 minus 6.7). Quarter of birth is thereby correlated with legally required years of schooling, and thus quarter of birth is also correlated with actual years of schooling. Quarter of birth therefore satisfies the first property of a good instrument, $\text{Cov}(d_i, \varepsilon_i) \neq 0$.

Does quarter of birth satisfy the second property of a good instrument, i.e. $\text{Cov}(z_i, \varepsilon_i) = 0$? Potentially, yes (though it turns out not). It seems plausible that the quarter in which one is born might not causally affect one's future wages, except through its effect on schooling (though there could be some strange weather effect on young babies). There is also no obvious reason to think that quarter of birth should be spuriously correlated with anything that affects future wages, particularly if we think that the time of conception is determined in a random manner. It is therefore plausible that quarter of birth and future wages are uncorrelated (except through changes in schooling).

How would we implement the quarter of birth instrument in practice? We would probably

use three instruments: a dummy for the first quarter (z_1), a dummy for the second quarter (z_2), and a dummy for the third quarter (z_3) (we exclude the fourth quarter to avoid the dummy variable trap, i.e. to avoid perfect colinearity with the constant term). So the first stage would be to regress schooling on quarter of birth (assuming there are no additional covariates that we are including):

$$d_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + u_i$$

Then take the predicted $\hat{d}_i$ from the first stage and use them in the second stage to run the regression:

$$y_i = \beta_0 + \beta_1 \hat{d}_i + u_i$$

The value of $\hat{\beta}_1$ from this regression is our estimate of the effect of schooling on wages. If the two IV assumptions are true ($\text{Cov}(d_i, \varepsilon_i) \neq 0$ and $\text{Cov}(z_i, \varepsilon_i) = 0$), then this will be a consistent estimate of the effect of schooling on wages.

There are a couple of things to note in this application. First, Angrist and Krueger implement IV in a couple of different ways. They begin with a Wald estimator which compares only two groups, people born in the first quarter and people born in the second through fourth quarters. The Wald estimator divides the difference in mean earnings for the two groups by the difference in mean schooling. Given our previous discussion of IV, it should be clear that this procedure is equivalent to doing IV with a binary instrument and no covariates. With this estimator, Angrist and Krueger estimate the return to schooling to be around 0.10 (i.e., one additional year of schooling raises wages by 10 percent) in the 1980 Census. This is higher than the OLS estimate from the same sample, which they find to be around 0.07. However, Angrist and Krueger also implement a 2SLS procedure in which they use dozens of instruments. They produce these instruments by interacting quarter of birth with year of birth (since the effect of quarter of birth on schooling might vary across years). In their 2SLS regressions, they frequently find coefficients closer to the OLS estimate of 0.07

than to the Wald estimate of 0.10. Unbeknown to them, the culprit behind this pattern is the "weak instruments" problem, which we will discuss in a subsequent section.

Second, while the quarter of birth instrument is much better than most instruments you will come across (at least in terms of satisfying the exclusion restriction), it is still not impervious to criticism. For example, many babies are conceived shortly after people get married. Some couples are likely to wait until the summer to get married, while other couples are more likely to get married quickly or when it is most convenient. Therefore, couples of the first type would be more likely to have children in the first or second quarter, whereas couples of the latter type would be equally likely to have children in any quarter. If couples of the first type are different in some important way (e.g., perhaps they have higher income on average) than couples of the second type, then that could introduce a correlation between quarter of birth and future wages. Any nonzero correlation between z_i and ε_i would be particularly problematic in this case because the first stage is relatively weak (again, we will discuss this issue in a subsequent section).