Instrumental variables

Christoph Hanck

Summer 2023

Isolating variation

Definition: Instrumental variable

Instead of trying to strip away all the undesirable variation using controls, an instrumental variable (IV) is a source of variation that allows to isolate just the front-door path of interest. IV designs attempt to mimic a randomized experiment, but using statistics and opportune settings instead of actually being able to influence or randomize anything.

For instrumental variables to work, two assumptions must be satisfied:

- Relevance of the instrument
- Validity of the instrument

Isolating variation

A typical setting with randomization

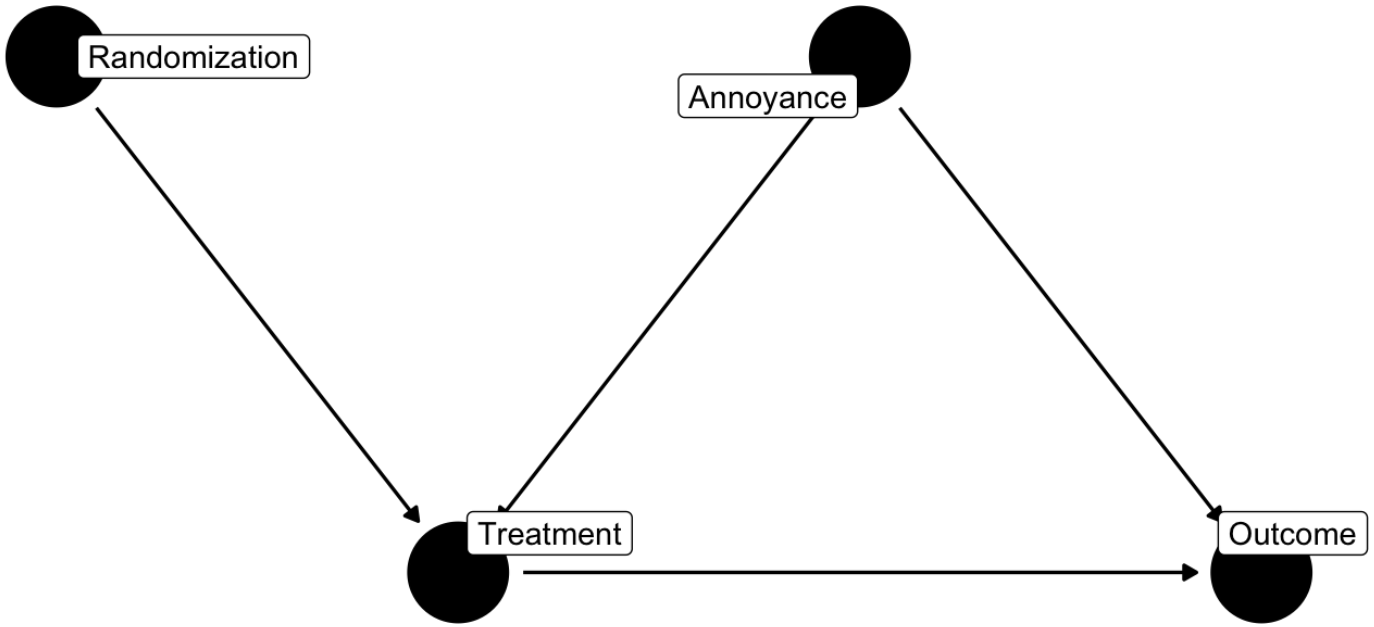

Figure 1: Causal Diagram

Isolating variation

An IV design works as above:

- An IV design does not remove the requirement to closing back doors in order to identify an effect
- IV shifts the requirement to something simpler: Instead of needing to close the back doors between $Treatment$ and $Outcome$ it becomes enough to

  ...close the back doors between $Randomization$ and $Outcome$.

  ...close the front doors between $Randomization$ and $Outcome$ that do not go through $Treatment$.

  (If there are any, respectively.)

Isolating variation

Steps

1. Use the IV to explain the treatment
2. Remove any part of the treatment that is not explained by the instrument
3. Use the instrument to explain the outcome
4. Remove any part of the outcome that is not explained by the instrument
5. Look at the relationship between the remaining, instrument-explained part of the outcome and the remaining, instrument-explained part of the treatment

Isolating variation

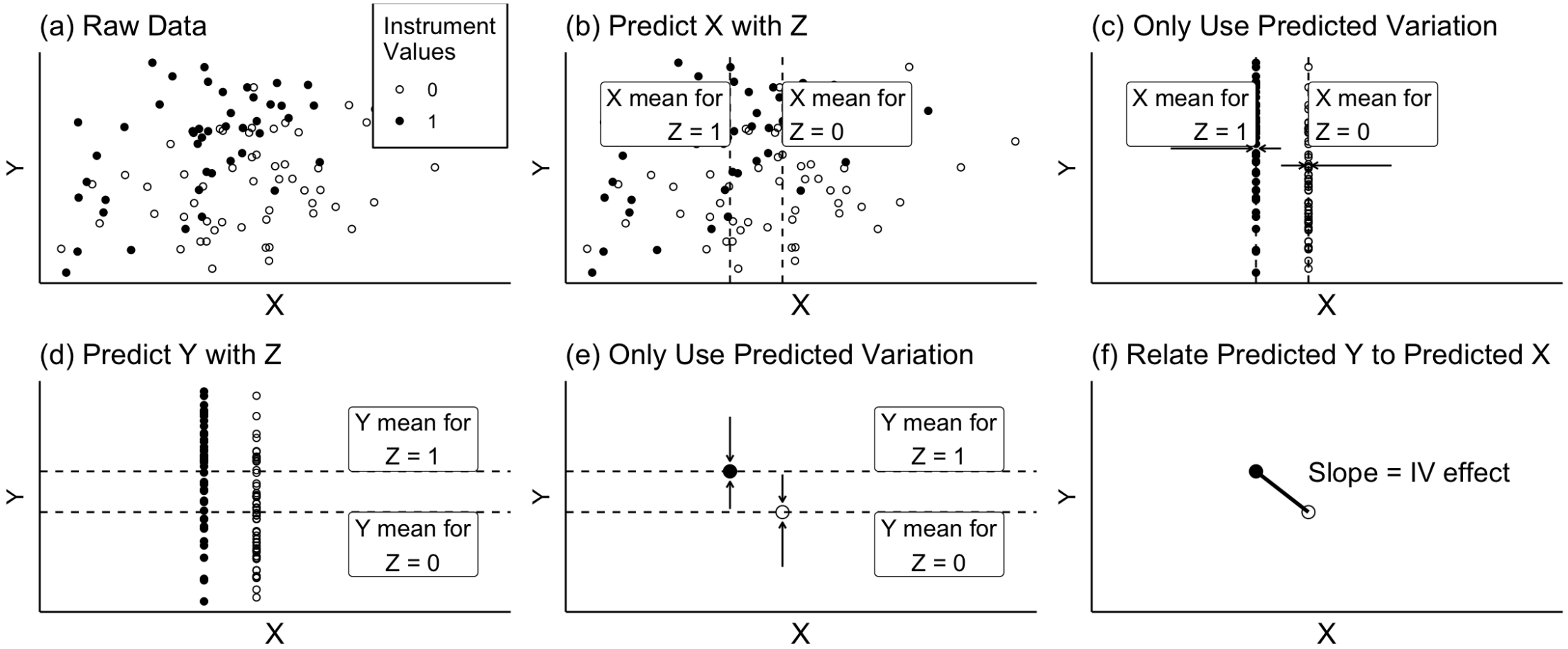

Figure 2: Instrumental variables with a binary instrument

Assumptions for IV

Relevance of the instrument

- The key idea of IV is to use that part of treatment X that is explained by the instrument Z to explain the outcome Y. In its basic form, IV gives us $\frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)}$.
- If no part of X is explained by Z then $\text{Cov}(Z, X) = 0$ and hence we get $\frac{\text{Cov}(Z, Y)}{0}$

  $\rightarrow$ IV does not work!

Validity of the instrument

Validity is the assumption that the instrument Z is has no open back doors of its own:

Any paths between the instrument Z and the outcome Y must either pass through the treatment X or be closed.

Assumptions for IV

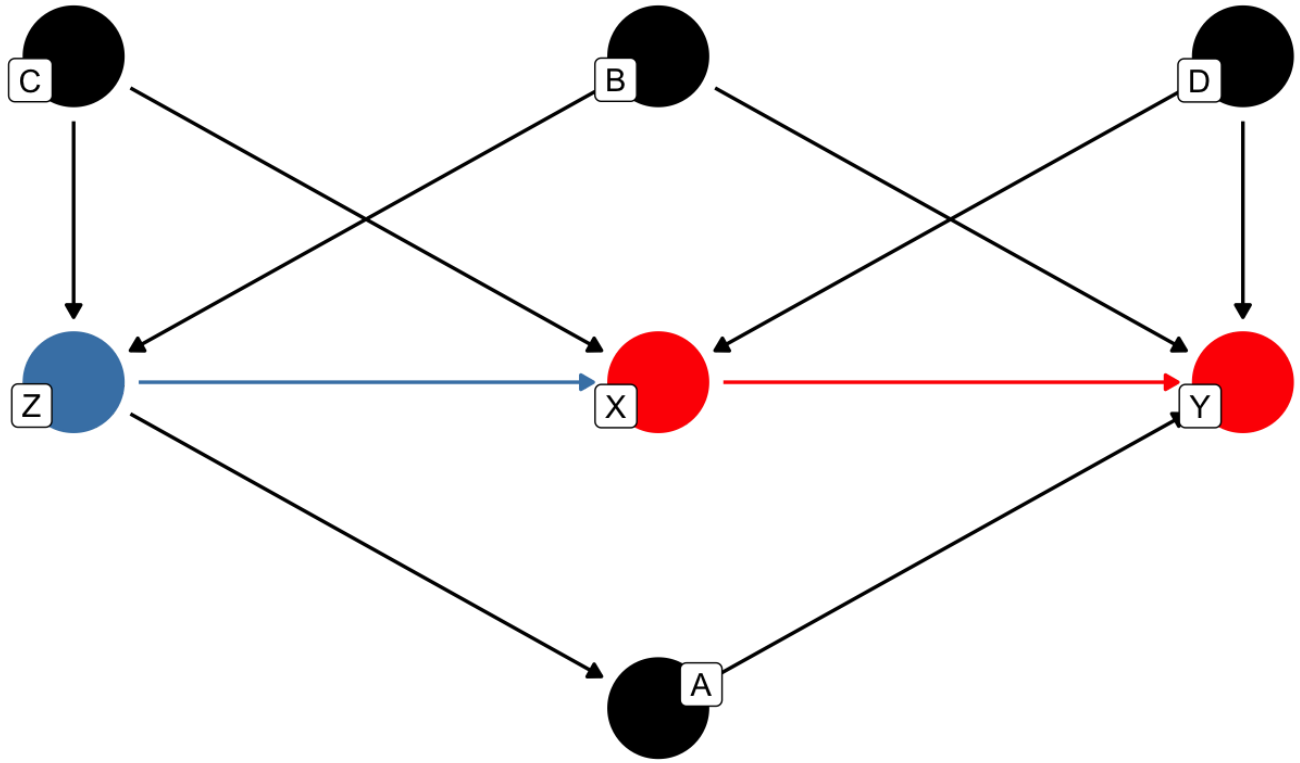

Figure 3: A causal diagram amenable to IV estimation

Assumptions for IV

Validity of the instrument

There are a few paths from Z to Y

- $Z \rightarrow X \rightarrow Y$
- $Z \rightarrow X \leftarrow D \rightarrow Y$
- $Z \leftarrow C \rightarrow X \rightarrow Y$
- $Z \leftarrow C \rightarrow X \leftarrow D \rightarrow Y$
- $Z \leftarrow B \rightarrow Y$
- $Z \rightarrow A \rightarrow Y$

Assumptions for IV

Validity of the instrument

- Assume that we cannot control for the endogeneity through D and we want to use Z as an IV

  $\rightarrow$ We require all open paths from Z to Y to contain X.
- Putting paths that contain X to the side, that gives
  - $Z \leftarrow B \rightarrow Y$
  - $Z \rightarrow A \rightarrow Y$

  $\rightarrow$ These two paths are problematic (why?) and must be shut down for Z to be valid!
- So for the validity assumption to hold for Z, we need to control for A and B.

Assumptions for IV

How realistic is validity?

Example: Effect of population's health on economic growth in that country

- Acemoglu and Johnson (2007) use the timing of when new medical technology (like vaccines for certain diseases) is introduced as an instrument for a country's health (specifically, mortality from certain diseases)
- They find that changes in health in a given year driven by changes in medical technology that year have a negative effect on a country's economic growth

Assumptions for IV

How realistic is validity?

Example: Effect of population's health on economic growth in that country

- Bloom et al. (2014) state that the original study did not account for the possibility that changes in health may have long-term effects on growth:

  The healthier the population, the less a new medical technology can further improve mortality rates.

- Effectiveness of medical technology is related to pre-existing health, which is related to pre-existing economic factors, which is related to current economic growth.

  → Back door!

Fixing the validity issue by controlling for pre-existing factors, Bloom et al. find a *positive* effect!

Canonical designs

The use of IV means requires *clever design*: finding causal inference using good instruments in situations with numerous back doors.

Clever instruments are often highly context-dependent. One good example of a canonical IV design is **judge assignment**.

Example: Judge assignment (Aizer and Doyle Jr. 2015)

- In many court systems, the process of assigning a judge to your trial is more or less random. This is important because some judges are harsher than others.
- This means $JudgeHarshness$ can act as an instrument for $Punishment$

Canonical designs

Example: Judge Assignment

- By estimating the harshness of each judge using their prior rulings, the harshness of the judge becomes an instrument for the punishment.
- This can be used anywhere where there is the system of randomly-assigned judges and the interest is to know the impact of harsher punishments (or even being judged guilty) on some later outcome.

IV estimators

IV estimators can be computed using OLS

Instrumental Variable can be estimated using Two-stage least squares (2SLS). The method uses two regressions to estimate an IV regresssion model

- The **first stage** uses the instrument Z (and controls W) to predict the treatment variable X,

$$X = \gamma_0 + \gamma_1 Z + \gamma_2 W + \nu$$

- In the **second stage**, the predicted (explained) values of the treatment variable $\widehat{X}$ from the first stage are used to predict the outcome Y in the second stage (along with the controls),

$$Y = \beta_0 + \beta_1 \widehat{X} + \beta_2 W + \epsilon$$

IV estimators

Interpretation of 2SLS

- 2SLS produces a *ratio of effects*, dividing the effect of Z on Y by the effect of Z on X
- It asks for each movement in X we get by changing Z, how much movement in Y does that lead to?

  $\rightarrow$ Since Z has no back doors, this should give us the causal effect of X on Y.

Advantages and disadvantages of 2SLS

- Easy to estimate and flexible: Adding more instruments to the first stage is a no-brainer (adding more treatment variables is less easy)
- Bad small-sample performance. For real data, the relationship between Z and the 'cleansed' part of Y is going to be somewhat nonzero, just by random chance
- 2SLS also does not perform particularly well when the errors are heteroskedastic

IV estimators

Generalized Method of Moments (GMM)

Using assumptions and theory, construct statistical moments (means, variances, covariances, etc.) that should have certain values. Use sample versions of these moments to estimate the quantity of interest

Example: The sample mean is a GMM estimator of the expected value

- To use GMM to estimate the expected value of a variable Y, it can be said that the difference between the estimator of μ and the population expected value $E(Y)$ should be zero
  $\rightarrow \mu - E(Y) = 0$ is a moment condition for the estimation.
- Replace $E(Y)$ with its sample value $\frac{1}{N}\sum Y$ and solve the equation to get $\widehat{\mu} = \frac{1}{N}\sum Y$
- GMM will pick the μ that makes the moment condition true, on average.

IV estimators

Overidentification

- When the number of instruments is bigger than the number of treatment/endogenous variables, we say that the model is **overidentified**

  When the model is overidentified, 2SLS and GMM will diverge!

- GMM will have less sampling variation (and thus smaller standard errors) under heteroskedasticity than 2SLS, even if heteroskedasticity-robust standard errors are used

IV estimation and treatment effects

For a standard estimator like 2SLS or GMM with one treatment/endogneous variable and *one* instrument, the weights are what the individual effect of the instrument would be in the first stage.

Example: Effect of exercise on blood pressure

A recent set of TV ads encourages people to exercise more. Exposure to the advertisements is used as an instrument for how much a person exercises.

Name	Effect of Ads on Exercise Hours	Effect of Exercise Hours on Blood Pressure
Jakeila	0.5	-2
Kyle	0.25	-8
Li	0.00	-10

IV estimation and treatment effects

Example: Effect of exercise on blood pressure — ctd.

- 2SLS is used to determine the effect of exercise hours on blood pressure.
- Jakeila responds the strongest to the ads, so the -2 effect of exercise that she gets will be more heavily weighted. Specifically, it gets the $.5$ weight she has on the effect of ads. Similarly, Kyle gets a weight of .25.
- Li, on the other hand, does not respond to the ads at all — they make no difference to him. So it turns out he makes no difference to the 2SLS estimate. He gets a weight of 0.

IV estimation and treatment effects

Example: Effect of exercise on blood pressure — ctd.

- The estimated LATE is

$$\frac{0.5 \times (-2) + 0.25 \times (-8) + 0 \times (-10)}{(0.5 + 0.25 + 0)} = \frac{-1 + (-2)}{0.75} = -4$$

- This is in contrast to the ATE,

$$\frac{(-2) + (-8) + (-10)}{3} = -6.67$$

IV estimation and treatment effects

Some more terminology

- Different IVs tend to produce different weighted average treatment effects
- LATE is estimated in most cases, although the specifics on *which* LATE change from estimator to estimator
- There is a common terminology for these weights. The sample can be divided into three groups.
  - **Compliers:** For compliers, the effect of the instrument on the treatment is in the expected direction.
  - **Always-takers/never-takers:** Always-takers/never-takers are completely unaffected by the instrument.
  - **Defiers:** Defiers are affected by the instrument in the opposite of the expected direction.

IV estimation and treatment effects

Some more terminology

- From the above terminology, we obtain
  - **... a result:** If all of the compliers are affected by the instrument to the same degree, then 2SLS gives the ATE among compliers.
  - **... an assumption:** For all of this to work, it is *important to assume that there are no defiers*.
- The assumption that there are no defiers is also known as the **monotonicity assumption**
- If there is an instrument that has an effect on average, then it is advisable to think carefully about whether that effect is likely to be in the same direction for everyone!

IV estimation and treatment effects

Example: Children and their parents (Angrist and Evans 1998)

- Families appear to have a preference for having both a boy and a girl
- A family that happens to have two boys as their first two children, or two girls, would be more likely to have a third child so as to try for a mix
- 'Your first two kids being the same gender' has been used as an instrument for 'having a third child' in lots of studies, following on from Angrist and Evans

IV estimation and treatment effects

Example: Children and their parents — ctd.

- But for this to work, *there have to be no defiers!*
- Even if most people would be more likely to have the third kid if the first two are the same gender (or not base their third-kid decision on that at all), if there are *some* people who would be less likely to have the third kid because the first two are the same gender, then monotonicity is violated!

Checking the IV assumptions

Definition: Weak instrument

A weak instrument is one that is valid and does predict the treatment variable, but it only predicts the treatment variable a little bit. *It predicts weakly*.

Checking the IV assumptions

First-stage F-test

An F-test on *all* instruments in the first stage regression is the most common way to check for relevance.

Steps

1. Estimate the first stage of the model (regress the treatment/endogenous variable on the controls and instruments)
2. Do a joint F-test on the instruments
3. Decide if the instruments are relevant

Checking the IV assumptions

Cut-off for F-test

- Weak instruments lead to bias because, even if the instrument is truly valid, in a sample the instrument will have a nonzero relationship with the error term just by random chance, *worsening validity*
- There is no single correct cutoff for F-statistic
- However, there is a *tradeoff*: The bigger the F-statistic, the less bias one gets

  → The F-statistic we want will depend on how much bias we are willing to accept

Checking the IV assumptions

Tests for the validity assumption

- There are several very well-known tests that are designed to test the validity assumption
- The main idea of the tests are to check whether there are any open back doors between the instrument Z and the outcome Y
- This can be done by checking whether Z is related to the *second-stage* error term ϵ
- If Z and ϵ are related, the validity assumption is violated
- Testing for validity has some obvious hurdles:

Checking the IV assumptions

Problems in testing for validity

ϵ cannot be observed, hence the relationship between Z and ϵ cannot be seen!

Approach 1

- Run the second-stage model but include the instrument as a control, i.e., estimate

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$$

- A nonzero coefficient on the instrument Z suggests a violation of the validity assumption

Checking the IV assumptions

Approach 2, more principled

- Run a **Durbin-Wu-Hausman test**:

  The Durbin-Wu-Hausman test compares the results from two models, one of which may have inconsistent results if some assumptions are wrong, while the other does not rely on those assumptions but is less precise.

- If the results are similar, it suggests that the assumptions at least are not so wrong that they mess up the results, and it means a 'green light' to use the more precise model

Checking the IV assumptions

Durbin-Wu-Hausman test

In the context of IV estimation, Durbin-Wu-Hausman is used in two ways:

1. It can be used to compare OLS — which is inconsistent if X is related to ϵ — to IV (which is less precise). If the results are different, that means that X has open back doors.
2. Durbin-Wu-Hausman can be used to compare two different *IV models* in an **overidentification test**

Either way, you need an estimator whose validity you are certain about.

Also recall what a non-rejection tells us.

Also recall LATE effects.

Fixing instrument weakness

There are several ways to cope with (possibly) weak instruments

- In the case of one treatment/endogenous variable and one instrument, the standard errors can be adjusted to account for the possibility that the instruments are weak
- **Anderson-Rubin** confidence intervals provide valid measures of uncertainty in the estimate of the effect even if the instruments are weak
- A common estimation method that attempts to perform better when the instrument is weak is **limited-information maximum likelihood (LIML)**

Fixing instrument weakness

Limited-information maximum likelihood (LIML)

- Remember that IV regression uses the parts of the treatment/endogenous variable X and the outcome Y that are predicted by the instrument Z
- LIML in the context of instrumental variables does the same thing, except that it scales down the prediction (making it weaker) using a parameter κ which is generally estimated from the data

Fixing instrument weakness

Limited-information maximum likelihood (LIML)

- If $\kappa = 1$, then the prediction is scaled by 1, and we end up with 2SLS
- If $\kappa < 1$, then a little of the original endogenous variable is brought back, that would have been used in OLS, relying less on that weak instrument
- **Fuller's α adjustment** can be made to κ as

$$\kappa = \hat{\kappa} - \frac{\alpha}{(N - N_I)},$$

where N is the number of observation and N_I is the number of instruments

Nonlinear IV estimation

Binary treatments

There is a handy and easy-to-implement method for incorporating a binary treatment variables into 2SLS popularized by Wooldridge (2010)

Steps:

1. Estimate the first stage using nonlinear regression (probit) of the treatment/endogenous variable on the instruments and controls
2. Get the predicted values
3. Use those predicted values *in place of the instrument* in 2SLS

Nonlinear IV estimation

What if the *outcome* is binary?

A common fix is the **control function approach**.

- The control function approach is a lot like 2SLS, except that instead of isolating the explained part of X, and using that in the second stage, one uses X and also *controls for the unexplained part* of X
- In regular linear instrumental variables, 2SLS and the control function give the same point estimates

Validity violation

There are two popular approaches which can be taken if the validity of the assumptions does *not perfectly* hold:

- The first tries to do what it can with a mostly-valid instrument:

  As long as the instrument impacts treatment, we can get useful partial inferences. The inferences we get are not as narrow as when you do have a 'fully valid' IV.

- The second makes up for invalidity by using a large number of instruments

Validity violation

Partial identification

- Instead of treating an invalid instrument as a reason to throw out the analysis, instead think about how bad that validity violation is — how strongly the instrument is related to the second stage error term — and construct a *range* of plausible estimates based on the plausible range of violations.
- This approach gives something called *partial identification*, since instead of giving a single estimate (identifying an estimate *exactly*), it gives a range, based on a range of different assumptions.

Lasso in IV estimation

- Lasso is a method that modifies OLS to tell it to minimize a sum of the absolute values of the coefficients, in addition to the sum of squared residuals, in effect encouraging it to toss variables out of the model entirely
- It can be used as it normally is, here, selecting the most important predictors among both control variables and instruments
- Lasso can also be used to help *spot invalid instruments*