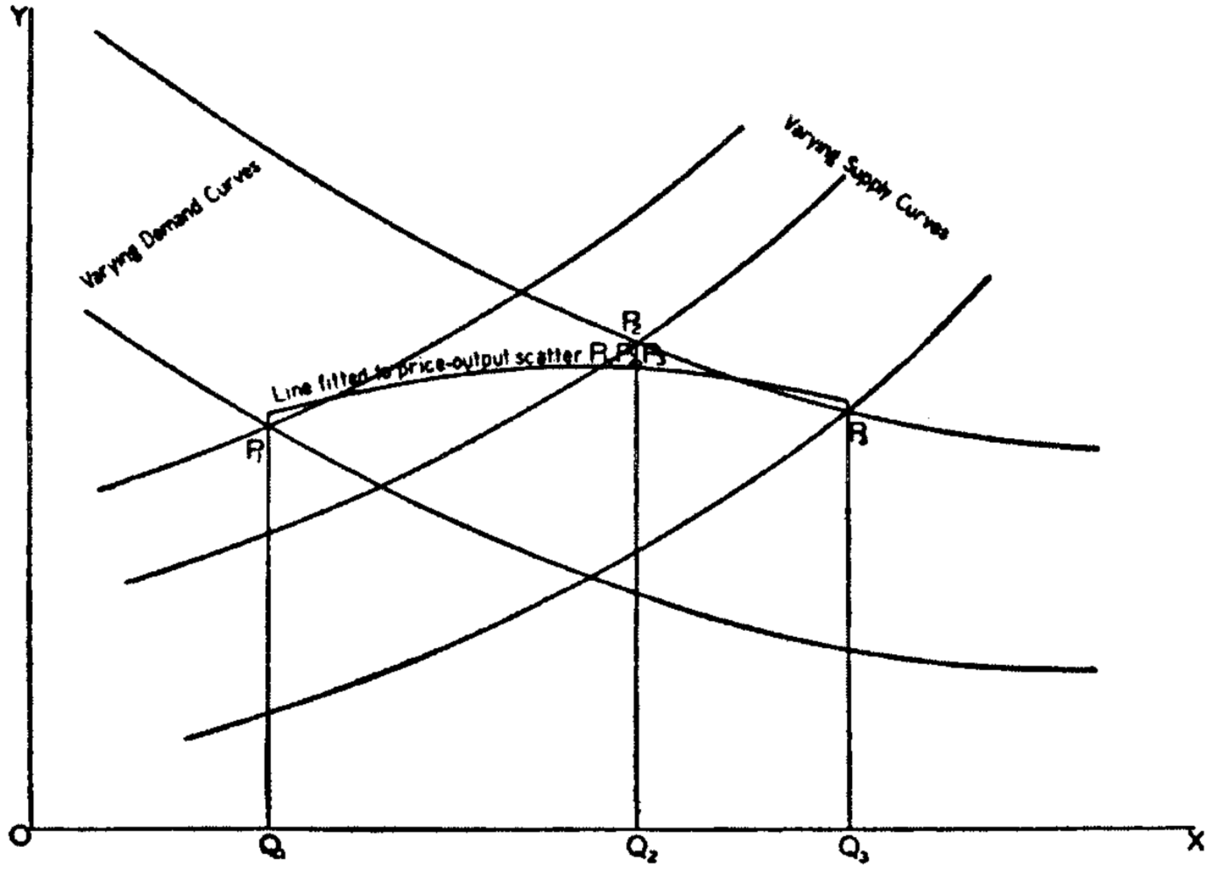
7 Instrumental Variables

History and background

IV estimator was first developed in the 1920s by father and son Phillip and Sewall Wright (Sewall Wright made major contributions in the field of genetics, including the path analysis that we discussed in the lecture on DAGs) who were studying agricultural markets. In 1928, Philip Wright wrote a book about animal and vegetable oils in which he argued that recent tariff increases harmed international relations. The challenge in his analysis was how to estimate the supply and demand curves if the equilibrium prices and quantities are determined simultaneously. When we see a scatter plot of prices and quantities, how can we use it to infer the supply and demand curves, on which curve do those points lie? Without further assumptions, we cannot say much.

Figure 4. Price-output Data Fail to Reveal Either Supply or Demand Curve.



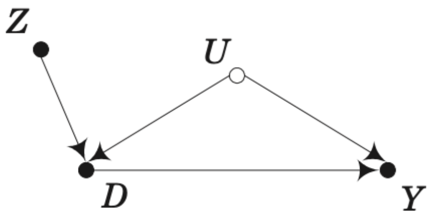
The method proposed by Phillip Wright (and inspired by his son's work on path analysis) solves the simultaneity problem by using the variables that appear in one equation to "shift" this equation and trace out the other. These variables are now called the *instrumental variables*. Stock and Trebbi (2003) provide a historical perspective on the development of IV by Phillip and Sewall Wright. The modern use of IV still uses the language developed for the simultaneous equations models application, however, the actual use is different.

Separately from the application to simultaneous equations, IV methods were used to solve the problem of bias from measurement error in regression models (see, e.g., Wald (1940)). In linear models, a regression coefficient is biased towards zero when the regressor of interest is measured with random errors. IV methods can eliminate this bias.

Homogenous treatment effects

In this section, we assume that the treatment effect (TE) is the same for every individual and equal to δ.

Consider the following causal diagram.



We are interested in the causal effect of D on Y. However, there is a backdoor path $D \leftarrow U \rightarrow Y$. U is unobserved, so we cannot close this backdoor path. This is a situation of *selection on unobservables*. There is no conditioning strategy that will satisfy the backdoor criterion.

D varies because of Z and U. Under the assumption of TE homogeneity, it is not necessary to relate all the variation in D to all the variation in Y in order to estimate of the causal effect. The co-variation in D and Y that is generated by U can be ignored if we can find a way of isolating the variation in D and Y that is causal.

There is a mediated path from Z to Y via D. When Z varies, D varies, which causes Y to vary. Importantly, Y is only varying because D has varied. While there is another path from Z to Y via U, D is a collider along that path, so this path is closed. For this to work, Z has to cause Y only through D. It cannot, e.g., cause Y directly or indirectly through some other variable. Likewise, Z cannot be correlated with U.

In general, any paths between the instrument Z and the outcome Y must either pass through the treatment D or be closed. There might be some common causes of Z and Y or mediated paths between Z and Y that do not go through D, but they must be closed. IV just moves that responsibility of closing the back doors from the treatment to the instrument, and hopefully it is easier to do that for the instrument.

Consider now a simple regression model with homogeneous treatment effects, in which both D and Z are binary. Z causes D but is independent of ϵ.

$$Y = \alpha + \delta D + \epsilon.$$

We assume that D is correlated with ϵ and that we cannot fix this issue by any conditioning strategy. Let's compute the expected value of Y conditional on Z:

$$\mathbb{E}[Y \mid Z=1] = \alpha + \delta\mathbb{E}[D \mid Z=1] + \mathbb{E}[\epsilon \mid Z=1]$$
$$\mathbb{E}[Y \mid Z=0] = \alpha + \delta\mathbb{E}[D \mid Z=0] + \mathbb{E}[\epsilon \mid Z=0].$$

Taking the difference between the two, we get

$$\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0] = \delta\left(\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0]\right) + \mathbb{E}[\epsilon \mid Z=1] - \mathbb{E}[\epsilon \mid Z=0].$$

Notice that the difference $\mathbb{E}[\epsilon \mid Z=1] - \mathbb{E}[\epsilon \mid Z=0] = 0$ since Z and ϵ are independent. Hence, we have

$$\delta = \frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0]}.$$

The sample analogue of this quantity is called the *Wald estimator*. It says that we can estimate the causal effect δ by simply dividing the NTE of Z on Y by the NTE of Z on D. The NTE of Z on Y is often called the *reduced form* and the NTE of Z on D is called the *first stage*, due to the language of a two-stage least squares (2SLS) estimator we will introduce later.

✓ Math time: Private schools and vouchers

Suppose we are interested in the causal effect of attending a private school (D) on academic achievement (Y). A simple regression of Y on D will likely give a biased estimate because there might be some unobserved confounders, like family background, that affect both going to a private school and achievement. However, suppose that there is a voucher program that randomly assigns vouchers for attending a private school. We can use vouchers as an IV (Z).

Assume the following joint distribution of Z and D, $\mathbb{P}(D, Z)$.

	$D=0$	$D=1$	$\mathbb{P}(Z)$
$Z=0$	0.8	0.1	0.9
$Z=1$	0.08	0.02	0.1
$\mathbb{P}(D)$	0.88	0.12	

Overall, only 12% of the students attend a private school and vouchers are assigned to 10% of the students. Notice that not everyone who receives a voucher goes to a private school. But some students do.

It will be helpful to compute the conditional distribution of $\mathbb{P}(D \mid Z)$ using the Bayes rule:

$$\mathbb{P}(D \mid Z) = \frac{\mathbb{P}(D, Z)}{\mathbb{P}(\overset{2}{D})}$$

	$Z=0$	$Z=1$
$\mathbb{P}(D=0 \mid Z)$	8/9	4/5
$\mathbb{P}(D=1 \mid Z)$	1/9	1/5

	$Z = 0$	$Z = 1$

Getting a voucher does increase the likelihood of attending a private school.

Now assume the following expected outcomes, $\mathbb{E}[Y \mid D, Z]$

	$D = 0$	$D = 1$
$Z = 0$	50	60
$Z = 1$	50	58

Using this table and the conditional distribution $\mathbb{P}(D \mid Z)$, we can estimate the causal effect of private schools using the Wald estimator.

$$\begin{aligned}
\mathbb{E}[Y \mid Z = 1] &= \mathbb{E}[Y \mid Z = 1, D = 0]\mathbb{P}(D = 0 \mid Z = 1) + \mathbb{E}[Y \mid Z = 1, D = 1]\mathbb{P}(D = 1 \mid Z = \\
&= 50 \times 4/5 + 58 \times 1/5 = 51.6 \\
\mathbb{E}[Y \mid Z = 0] &= \mathbb{E}[Y \mid Z = 0, D = 0]\mathbb{P}(D = 0 \mid Z = 0) + \mathbb{E}[Y \mid Z = 0, D = 1]\mathbb{P}(D = 1 \mid Z = \\
&= 50 \times 8/9 + 60 \times 1/9 \approx 51.11 \\
\mathbb{E}[D \mid Z = 1] &= \mathbb{P}(D = 1 \mid Z = 1) = 1/5 \\
\mathbb{E}[D \mid Z = 0] &= \mathbb{P}(D = 1 \mid Z = 0) = 1/9
\end{aligned}$$

Plugging in all the values, we get

$$\delta = \frac{51.6 - 51.11}{1/5 - 1/9} \approx 5.51.$$

Homework 5

Show that the naive treatment effect is 9.67.

Non-binary treatment variable

When the treatment variable is non-binary, we can derive a similar result. Assume that the treatment variable is generated according to

$$D = \gamma + \beta Z + \eta.$$

Recall that

$$\beta = \frac{Cov(D, Z)}{V(Z)}.$$

Let $\hat{D} = \mathbb{E}[D \mid Z] = \gamma + \beta Z$ be the conditional expectation function (or predicted/fitted values) of D. Notice that $V(\hat{D})$ = $V(\beta Z)$ and $Cov(\hat{D}, X) = Cov(\beta Z, X)$ (for any random variable X).

Now consider the covariance between Y and Z:

$$Cov(Y, Z) = Cov(\alpha + \delta D + \epsilon, Z) = \delta Cov(D, Z) + Cov(\epsilon, Z).$$

Since ϵ and Z are independent by assumption, we get

$$\delta = \frac{Cov(Y,Z)}{Cov(D,Z)}.$$

This is a generalization of the Wald estimator to the case when D is non-binary. Interestingly, this formula also generalizes a regular OLS estimator: if you substitute D for Z, we are back to the usual expression for the regression coefficient. In other words, OLS treats D as its own IV.

If we divide both part by $V(Z)$, we get

$$\delta = \frac{Cov(Y,Z)/V(Z)}{Cov(D,Z)/V(Z)}.$$

The numerator is then the coefficient on Z in the simple regression of Y on Z (reduced form) and the denominator is the coefficient on Z in the simple regression of D on Z (first stage).

We can rewrite the expression for δ one more time.

$$\delta = \frac{Cov(Y,Z)/V(Z)}{\beta} = \frac{\beta Cov(Y,Z)}{\beta^2 V(Z)} = \frac{Cov(Y,\beta Z)}{V(\beta Z)} = \frac{Cov(Y,\hat{D})}{V(\hat{D})}.$$

This means that, effectively, we can estimate δ as a coefficient from a simple regression of Y on the *predicted* values of D. This is how the 2SLS estimator works: first, you get the predicted values of D by regressing it on the instrument Z, and second, you regress Y on the predicted values of D (in practice, the actual algorithm is a bit different). So in estimating the effect of D on Y using IV, we are only using the *exogenous* variation in D caused by Z.

Assumptions and weaknesses

We need two assumptions for IV to identify a homogenous TE: *relevance* and *validity*. Relevance means that our IV should have a causal effect on the treatment variable. This assumption is obvious: if we want to isolate the variation in D and Y due to Z, then Z must cause D to vary. Technically, if relevance is not satisfied, then the denominators in the IV formulas are zero and the estimators are undefined. Validity means that the instrument should affect the outcome *only through* the treatment variable (or, in general, that there are no open back door paths between the instrument and outcome). Technically, this assumption is what leads to the correlation between ϵ and Z to be zero and eliminates the conditional expectations of ϵ from the formula for δ. The validity assumption is also called the *exclusion restriction*.

In practice, a bigger issue is not that the covariance between the instrument and treatment is exactly zero. In a finite sample, it can be non-zero even if the instrument is completely unrelated to the treatment. The issue is that this covariance may be small. This is known as a *weak instrument* problem (Bound, Jaeger, and Baker (1995)). In general, weak instruments tend to produce biased estimates. The argument is the following:

1. In finite samples, IV point estimates can always be computed because sample covariances are never exactly equal to zero;
2. An IV point estimate can be computed even for an instrument that is not relevant
3. The formulas for calculating the standard errors of IV estimates fail in such situations, giving artificially small standard errors
4. The bias due to small violations of the validity assumption can explode

For example, for a non-binary treatment variable, the bias is equal to

$$\frac{Cov(\epsilon, Z)}{Cov(D, Z)}.$$

Even if the numerator is small (but non-zero), a weak instrument (small $Cov(D, Z)$) will blow up the bias. It can be shown that the expected bias of the 2SLS estimator is inversely related to the F statistic about the joint significance from the first stage. If F is large, the bias of the 2SLS goes to zero. However, as F gets smaller, the bias increases. In general, the F statistics decreases if the instrument becomes weaker or if you increase the number of weak instruments. How big does the first-stage F need to be in practice? A common rule of thumb is that it should be greater than 10, but Stock and Yogo (2005) provide more specific guidance.

What to do if your instrument is weak? One option is to use a single strongest IV instead of many weak IVs. Another option is to use alternative IV estimators, such as a limited-information maximum likelihood estimator (LIML). In terms of standard errors, one can use Anderson-Rubin confidence intervals (Anderson and Rubin (1949)) that provide valid standard errors even if the instruments are weak. But the ultimate solution would be to just get better instruments, if possible.

The validity assumption (or exclusion restriction), on the other hand, is untestable. One can think that you can test it empirically. Suppose you run a regression of Y on Z and D. If there is an association between Z and Y after conditioning on D, then the instrument must be invalid. The idea is that conditioning on D will block the indirect relationship between Z and Y through D. If the only association between Z and Y is through D, then there should be no association between Z and Y after conditioning on D. However, this logic is false. If the IV is invalid, then Z and Y will be associated after conditioning on D. But the converse is *not* true. Z and Y will always be associated after conditioning on D when validity holds, because D is a collider that is mutually caused by both Z and U. Conditioning on D creates dependence between Z and U, even if validity holds.

There are also tests for overidentification in case you have more than one instrument: Durbin-Wu-Hausman test and Sargan test. These tests can tell you if some instruments are likely to be invalid.

An informal test for the validity is that "good instruments should feel weird." If you introduce an instrument and tell someone that it affects the outcome, the effect should not be obvious, unless you explain the mechanism. Without knowledge of the treatment variable, the relationship between the instrument and the outcome should not make much sense. Because a valid instrument should be irrelevant to the determinants of the outcome, except for its effect on the treatment.

For example, suppose you tell someone that mothers whose first two children were the same gender were employed outside of the home less than those whose two children were a boy and a girl. That would be confusing. What does the gender composition have to do with whether a woman works outside the home? It does not seem to change the incentives. However, empirically if the first two children are the same gender, families are more likely to have a third compared to those who had a boy and a girl first. So you have an instrument (gender composition) that only changes the outcome (labor supply) through changing a treatment variable (family size), which allows us to identify the causal effect of family size on labor supply.

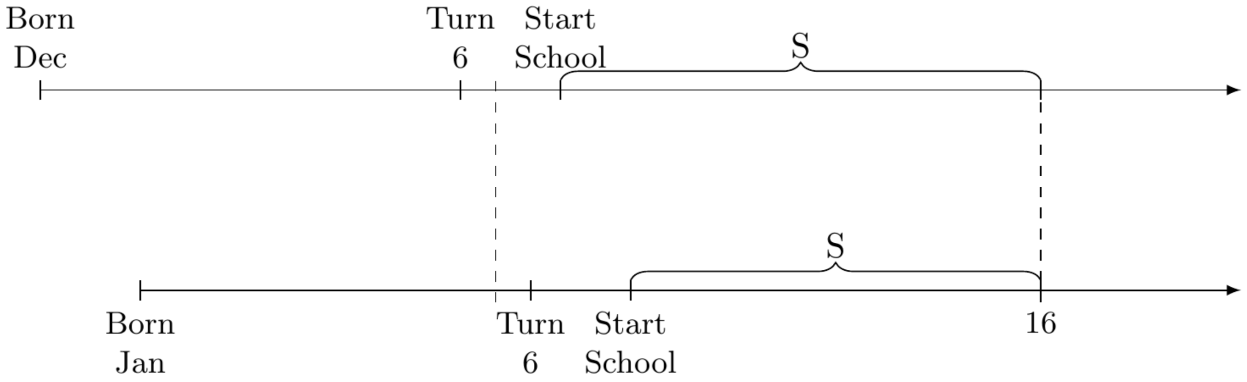
IV estimators are also noisy, i.e., they have large standard errors. By using only a portion of the available co-variation in the treatment and outcome, IV estimators use only a portion of the information in the data. This represents a loss in statistical power, and as a result, IV estimators tend to exhibit substantially more sampling variance than other estimators.

IV in action

IV has been used to estimate the causal effect of education on labor market earnings. As reviewed by Card (1999) and by Angrist and Krueger (2001), the causal effect of years of schooling on subsequent earnings has been estimated with a variety of IVs, including proximity to college, regional and temporal variation in school construction, tuition at local colleges, temporal variation in the minimum school-leaving age, and quarter of birth. The argument here is that each of these variables predicts educational attainment but has no direct effect on earnings.

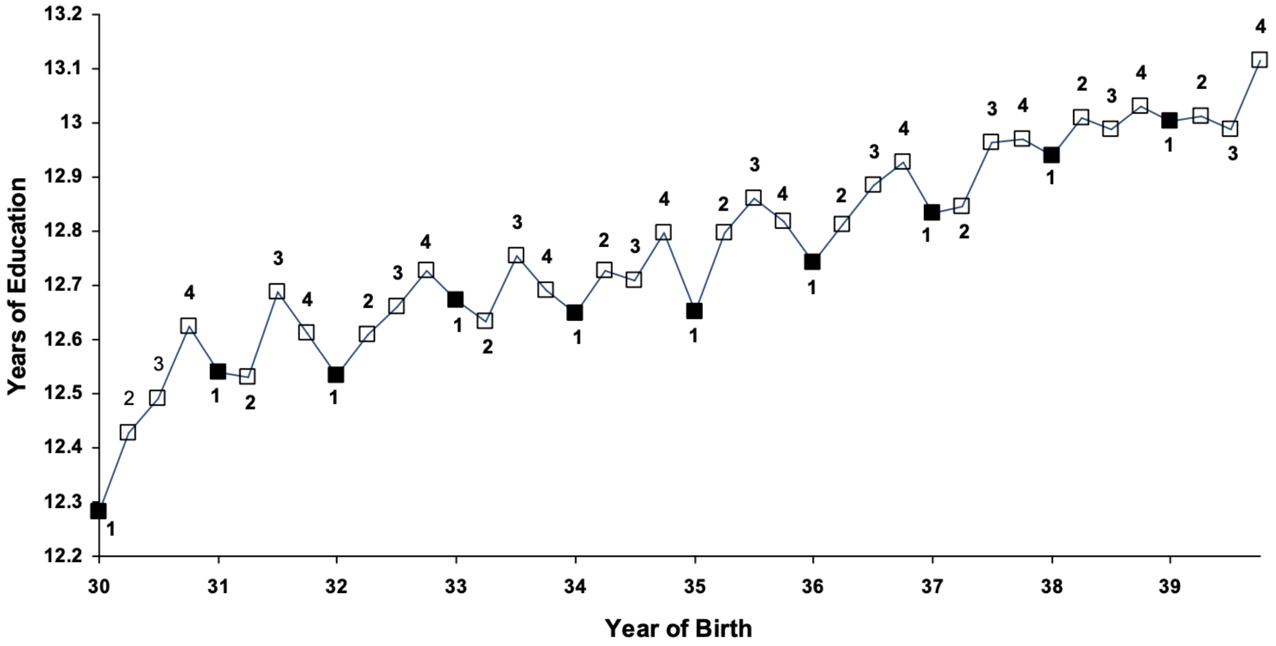
Consider the quarter of birth instrument used by Angrist and Krueger (1991). In the United States educational system, a child enters a grade on the basis of his or her birthday. For a long time, that cutoff was late December. If children were born on or before December 31, then they were assigned to the first grade. But if their birthday was on or after January 1, they were assigned to kindergarten. Thus, two people—one born on December 31 and one born on January 1—were exogenously assigned different grades. Also, for most of the twentieth century, the US had compulsory schooling laws that forced a person to remain in high school until age 16. After age 16, one could legally stop going to school.

The following figure illustrates this idea



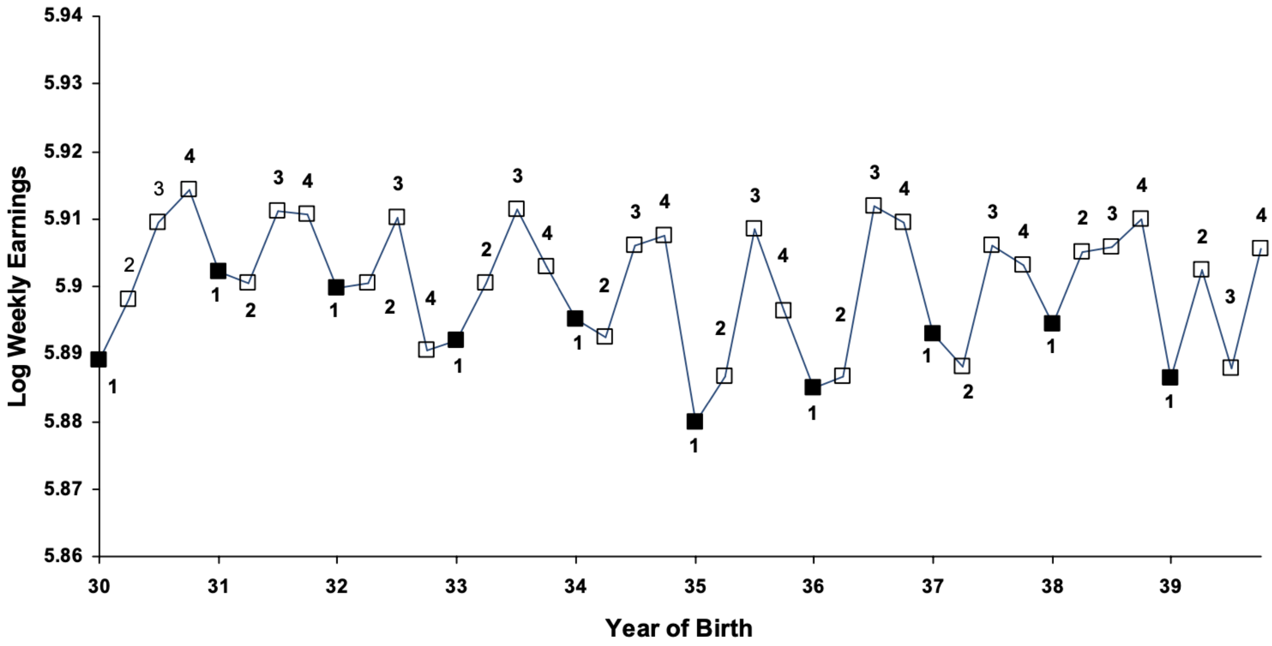
Angrist and Krueger (1991) show that quarter of birth does affect schooling. The following figure is the first stage.

A. Average Education by Quarter of Birth (first stage)



The picture below shows that the quarter of birth affects earnings (reduced form).

B. Average Weekly Wage by Quarter of Birth (reduced form)



Combining these two effects, they estimate that an 8.9% return for every additional year of schooling.

Heterogeneous treatment effects

The assumption of homogenous treatment effects is rather strong, in this section we are going to relax it. If we allow treatment effects to be heterogeneous, then what kind of an average treatment effect would we estimate using IV? We will introduce a new treatment effect parameter: the local average treatment effect (LATE). The following discussion uses the ideas developed in Imbens and Angrist (1994).

Imbens and Angrist (1994) classify individuals into those who respond positively to an instrument, those who remain unaffected by an instrument, and those who rebel against an instrument. We can define potential treatment assignment variables, $D^{Z=z}$, for each state z of the instrument Z. When D and Z are binary, there are four possible groups of individuals in the population.

	$D^1 = 0$	$D^1 = 1$
$D^0 = 0$	Never takers ($\tilde{C} = n$)	Compliers ($\tilde{C} = c$)
$D^0 = 1$	Defiers ($\tilde{C} = d$)	Always takers ($\tilde{C} = a$)

Here, variable $\tilde{C}$ represents compliance status and takes one of the four possible values.

The observed treatment assignment (D) can now be represented using its own switching equation (analogous to the switching equation for potential outcomes):

$$D = D^0 + (D^1 - D^0)Z.$$

The difference $D^1 - D^0$ is the individual-level causal effect of the instrument on D, which can vary across individuals. If the instrument represents encouragement to take the treatment, such as the randomly assigned school voucher in our earlier example, then $D^1 - D^0$ can be interpreted as the individual-level compliance inducement effect of the instrument. This difference equals 1 for compliers, −1 for defiers, and 0 for always takers and never takers (because none of them respond to the instrument).

We need three assumptions to characterize the treatment effect we would obtain using the Wald estimator:

1. Independence: $(Y^1, Y^0, D^1, D^0) \perp Z$
2. Relevance: $D^1 - D^0 \neq 0$ for all i
3. Monotonicity: $D^1 - D^0 \geqslant 0$ for all i or $D^1 - D^0 \leqslant 0$ for all i

The independence assumption states that the potential outcomes and potential treatment assignments (but not the observed ones) are independent of the instrument. Knowing the value of the instrument for individual i must not yield any information about the potential outcomes of individual i or potential treatment status. It implies the validity assumption (or exclusion restriction) that we discussed earlier and that the reduced form gives us the causal effect of Z on Y. It also implies that the first stage can be used to estimate the causal effect of Z on D:

$$\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0] = \mathbb{E}[D^1 \mid Z=1] - \mathbb{E}[D^0 \mid Z=0] = \mathbb{E}[D^1 - D^0].$$

The relevance assumption (nonzero effect of Z on D) means that the instrument must cause treatment assignment for at least some individuals. There must be at least some compliers or some defiers in the population of interest. The monotonicity assumption further specifies that the effect of Z on D must be either weakly positive or weakly negative for all individuals i. There may be either defiers or compliers in the population but not both.

If the assumptions hold, then an instrument identifies a local average treatment effect (LATE): the average causal effect of the treatment for the subset of the population whose treatment selection is induced by the instrument. In particular, if there are no defiers, then

$$LATE = \frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0]} = \mathbb{E}[Y^1 - Y^0 \mid \tilde{C} = c]$$

Proof

Consider the numerator. By independence, we have

$$\mathbb{E}[Y \mid Z=1] = \mathbb{E}[Y^0 + D^1(Y^1 - Y^0)]$$
$$\mathbb{E}[Y \mid Z=0] = \mathbb{E}[Y^0 + D^0(Y^1 - Y^0)]$$

Taking the difference between the two, we get

$$
\begin{aligned}
\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0] &= \mathbb{E}[Y^0 + D^1(Y^1 - Y^0)] - \mathbb{E}[Y^0 + D^0(Y^1 - Y^0)] \\
&= \mathbb{E}[(Y^1 - Y^0)(D^1 - D^0)] \\
&= \mathbb{E}[(Y^1 - Y^0)(D^1 - D^0) \mid D^1 > D^0]\mathbb{P}(D^1 > D^0). \\
&+ \mathbb{E}[(Y^1 - Y^0)(D^1 - D^0) \mid D^1 = D^0]\mathbb{P}(D^1 = D^0) \\
&= \mathbb{E}[Y^1 - Y^0 \mid D^1 > D^0]\mathbb{P}(D^1 > D^0)
\end{aligned}
$$

Now consider the denominator. Again, by independence we have

$$
\begin{aligned}
\mathbb{E}[D \mid Z=1] &= \mathbb{E}[D^1] \\
\mathbb{E}[D \mid Z=0] &= \mathbb{E}[D^0].
\end{aligned}
$$

Taking the difference, we get

$$
\begin{aligned}
\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0] &= \mathbb{E}[D^1 - D^0] \\
&= \mathbb{E}[D^1 - D^0 \mid D^1 > D^0]\mathbb{P}(D^1 > D^0) \\
&+ \mathbb{E}[D^1 - D^0 \mid D^1 = D^0]\mathbb{P}(D^1 = D^0) \\
&= \mathbb{P}(D^1 > D^0).
\end{aligned}
$$

Dividing the two parts, we get

$$
\frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0]} = \mathbb{E}[Y^1 - Y^0 \mid D^1 > D^0].
$$

The case when $D^1 > D^0$ is only possible when $D^1 = 1$ and $D^1 = 0$, in other words, when $\tilde{C} = c$ (compliers).

The difference in the average value of Y, when examined across Z, is not a function of the outcomes of always takers and never takers. Defiers and compliers contribute all of the variation that generates the IV estimate because only their behavior is responsive to the instrument. If compliers are present but defiers are not, then the causal estimate is interpretable as the average causal effect for compliers. If defiers are present but compliers are not, then the causal estimate is interpretable as the average causal effect for defiers. If both compliers and defiers are present, then the estimate generated by the ratio does not have a well-defined causal interpretation.

Using our assumptions, we can actually count all the four groups of individuals. The proof of the LATE theorem, in particular, shows that the probability of being a complier is

$$
\mathbb{P}(\tilde{C} = c) = \mathbb{P}(D^1 > D^0) = \mathbb{E}[D \mid Z=1] - \mathbb{E}[D \mid Z=0] = \mathbb{P}(D=1 \mid Z=1) - \mathbb{P}(D=1 \mid Z=0),
$$

which is simply the first stage.

To find the probabilities of being an always taker or a never taker, we can use the following observations.

$$
\begin{aligned}
\mathbb{P}(D=1 \mid Z=0) &= \mathbb{P}(D^0=1, D^1=1 \mid Z=0) + \mathbb{P}(D^0=1, D^1=0 \mid Z=0) \\
&= \mathbb{P}(\tilde{C}=a) + \mathbb{P}(\tilde{C}=d) \\
&= \mathbb{P}(\tilde{C}=a),
\end{aligned}
$$

where we use independence and the assumption of no defiers.

Similarly, we can find that

$$\begin{aligned}\mathbb{P}(D=0 \mid Z=1) &= \mathbb{P}(D^0=0, D^1=0 \mid Z=1) + \mathbb{P}(D^0=1, D^1=0 \mid Z=1) \\ &= \mathbb{P}(\tilde{C}=n) + \mathbb{P}(\tilde{C}=d) \\ &= \mathbb{P}(\tilde{C}=n).\end{aligned}$$

Depending on the values of the treatment status and the instrument, we can construct the following table, which shows where different individuals are located based on their compliance status.

	$D=0$	$D=1$
$Z=0$	never takers + compliers	always takers
$Z=1$	never takers	always takers + compliers

We can count the number of individuals in each of these cells (or rather probabilities of belonging to each type). For example, the joint probability of being a never taker and being in the top left cell is

$$\mathbb{P}(\tilde{C}=n, D=0, Z=0) = \mathbb{P}(\tilde{C}=n, Z=0) = \mathbb{P}(\tilde{C}=n)\mathbb{P}(Z=0).$$

Similarly, we can find the probability of being an always taker and being in the bottom right cell.

$$\mathbb{P}(\tilde{C}=a, D=1, Z=1) = \mathbb{P}(\tilde{C}=a, Z=1) = \mathbb{P}(\tilde{C}=a)\mathbb{P}(Z=1).$$

The probabilities of being a complier in each of the diagonal cells are

$$\begin{aligned}\mathbb{P}(\tilde{C}=c, D=0, Z=0) &= \mathbb{P}(\tilde{C}=c)\mathbb{P}(Z=0), \\ \mathbb{P}(\tilde{C}=c, D=1, Z=1) &= \mathbb{P}(\tilde{C}=c)\mathbb{P}(Z=1).\end{aligned}$$

> Notice, however, that we cannot identify which individuals are never takers and which are compliers in the top left cell. Likewise, we cannot identify which individuals are always takers and which are compliers in the bottom right cell.

Having these probabilities lets us find the average treatment effect for compliers in a more explicit, albeit longer, way. Consider the observed outcomes in the diagonal cells. For example, in the top left cell, the average outcome would be the the weighted average of outcomes for never takers and compliers.

$$\begin{aligned}\mathbb{E}[Y \mid D=0, Z=0] = {} & \mathbb{E}[Y \mid D=0, Z=0, \tilde{C}=n]\mathbb{P}(\tilde{C}=n \mid D=0, Z=0) \\ & + \mathbb{E}[Y \mid D=0, Z=0, \tilde{C}=c]\mathbb{P}(\tilde{C}=c \mid D=0, Z=0).\end{aligned}$$

Notice that the conditional expectations on the RHS can be rewritten as

$$\begin{aligned}\mathbb{E}[Y \mid D=0, Z=0, \tilde{C}=n] &= \mathbb{E}[Y^0 + (Y^1 - Y^0)D \mid D=0, Z=0, \tilde{C}=n] \\ &= \mathbb{E}[Y^0 \mid \tilde{C}=n], \\ \mathbb{E}[Y \mid D=0, Z=0, \tilde{C}=c] &= \mathbb{E}[Y^0 + (Y^1 - Y^0)D \mid D=0, Z=0, \tilde{C}=c] \\ &= \mathbb{E}[Y^0 \mid \tilde{C}=c]\end{aligned}$$

The conditional probabilities are given by the Bayes' rule:

$$\begin{aligned}\mathbb{P}(\tilde{C}=n \mid D=0, Z=0) &= \frac{\mathbb{P}(\tilde{C}=n, D=0, Z=0)}{\mathbb{P}(D=0, Z=0)} \\ &= \frac{\mathbb{P}(\tilde{C}=n)\mathbb{P}(Z=0)}{\mathbb{P}(D=0 \mid Z=0)\mathbb{P}(Z=0)} \\ &= \frac{\mathbb{P}(\tilde{C}=n)}{\mathbb{P}(D=0 \mid Z=0)}, \\ \mathbb{P}(\tilde{C}=c \mid D=0, Z=0) &= \frac{\mathbb{P}(\tilde{C}=c)}{\mathbb{P}(D=0 \mid Z=0)}.\end{aligned}$$

Hence, the control-state outcome for compliers is

$$\mathbb{E}[Y^0 \mid \tilde{C}=c] = \frac{\mathbb{E}[Y \mid D=0, Z=0]\mathbb{P}(D=0 \mid Z=0) - \mathbb{E}[Y^0 \mid \tilde{C}=n]\mathbb{P}(\tilde{C}=n)}{\mathbb{P}(\tilde{C}=c)}.$$

Following similar steps, we can find the treatment-state outcome for compliers

$$\mathbb{E}[Y^1 \mid \tilde{C}=c] = \frac{\mathbb{E}[Y \mid D=1, Z=1]\mathbb{P}(D=1 \mid Z=1) - \mathbb{E}[Y^1 \mid \tilde{C}=a]\mathbb{P}(\tilde{C}=a)}{\mathbb{P}(\tilde{C}=c)}.$$

The difference between the two will give the average treatment effect for compliers.

Example: Private Schools and Vouchers, part 2

Recall our example. We will assume that there are no defiers. First, let's find the proportions of the remaining three groups.

$$\begin{aligned}\mathbb{P}(\tilde{C}=a) = \mathbb{P}(D=1 \mid Z=0) = 1/9(0.111) \\ \mathbb{P}(\tilde{C}=n) = \mathbb{P}(D=0 \mid Z=1) = 4/5(0.8).\end{aligned}$$

The proportion of compliers can be found as either one minus the previous two proportions or using the first stage:

$$\mathbb{P}(\tilde{C}=c) = \mathbb{P}(D=1 \mid Z=1) - \mathbb{P}(D=1 \mid Z=0) = 1/5 - 1/9 = 4/45(0.0889).$$

Now let's find the proportions of individuals in each of the four cells defined by Z and D.

$$\begin{aligned}\mathbb{P}(\tilde{C}=c, D=0, Z=0) = \mathbb{P}(\tilde{C}=c)\mathbb{P}(Z=0) = 4/45 \times 9/10 = 0.08 \\ \mathbb{P}(\tilde{C}=c, D=1, Z=1) = \mathbb{P}(\tilde{C}=c)\mathbb{P}(Z=1) = 4/45 \times 0.1 \approx 0.0089 \\ \mathbb{P}(\tilde{C}=n, D=0, Z=0) = \mathbb{P}(\tilde{C}=n)\mathbb{P}(Z=0) = 4/5 \times 0.9 = 0.72 \\ \mathbb{P}(\tilde{C}=a, D=1, Z=1) = \mathbb{P}(\tilde{C}=n)\mathbb{P}(Z=1) = 1/9 \times 0.1 \approx 0.0111\end{aligned}$$

	$D = 0$	$D = 1$
$Z = 0$	never takers (0.72) + compliers (0.08)	always takers (0.1)
$Z = 1$	never takers (0.08)	always takers (0.0111) + compliers (0.0089)

Finally, the outcomes for compliers are

$$
\begin{aligned}
\mathbb{E}[Y^1 \mid \tilde{C} = c] &= \frac{\mathbb{E}[Y \mid D = 1, Z = 1]\mathbb{P}(D = 1 \mid Z = 1) - \mathbb{E}[Y^1 \mid \tilde{C} = a]\mathbb{P}(\tilde{C} = a)}{\mathbb{P}(\tilde{C} = c)} \\
&= \frac{58 \times 1/5 - 60 \times 1/9}{4/45} = 55.5 \\
\mathbb{E}[Y^0 \mid \tilde{C} = c] &= \frac{\mathbb{E}[Y \mid D = 0, Z = 0]\mathbb{P}(D = 0 \mid Z = 0) - \mathbb{E}[Y^0 \mid \tilde{C} = n]\mathbb{P}(\tilde{C} = n)}{\mathbb{P}(\tilde{C} = c)} \\
&= \frac{50 \times 8/9 - 50 \times 4/5}{4/45} = 50
\end{aligned}
$$

The average treatment effect for compliers is then 5.5, which is the number we found earlier using the Wald estimator. This value, on the other hand, yields no information about the effect of private schooling for the always takers or the never takers. We have no way to estimate the counterfactuals for them: the mean outcome in public schools for always takers and the mean outcome in privates schools for never takers.

Discussion of LATE

LATE estimators depend on the instrument under consideration. As a result, different instruments define different average treatment effects for the same group of treated individuals. And, when this is possible, the meanings of the labels for $\tilde{C}$ depend on the instrument, such that some individuals can be never takers for one instrument and compliers for another. This also means that different IVs will in general produce different estimated causal effects.

Although from one perspective this is a weakness, from another it is a feature. In our example, the IV estimate does not provide any information about the average effect for individuals who would attend private schooling anyway (i.e., the always takers) or those who would still not attend the private schools if given a voucher (i.e., the never takers). Instead, the IV estimate is an estimate of a narrowly defined average effect only among those induced to take the treatment by the voucher policy intervention. However, this is precisely what should be of interest for policy evaluation purposes. If the policy question is "What is the effect of vouchers on school performance?" then they presumably care most about the average effect for compliers.

If more than one instrument is available, the traditional econometric literature suggests that they should all be used to overidentify the model and obtain a more precise treatment effect estimate. Overidentified models generate a mixture-of-LATEs problem. Since each instrument defines a LATE for a different group of individuals, the estimated causal effect would be averaged across these different groups, even though one would probably be more interested in each group separately.

LATE with one-sided non-compliance

In general, LATE is not identical to either ATT or ATU. However, in a special case of one-sided non-compliance LATE coincides with ATT. In many randomized trials, participation is voluntary among those randomly assigned to receive treatment. On the other hand, no one in the control group has access to the experimental intervention. This means that $\mathbb{P}(D = 1 \mid Z = 0) = 0$.

Since the group that receives (i.e., complies with) the assigned treatment is a self-selected subset of those offered treatment, a comparison between those actually treated and the control group is misleading. The selection bias in this case is almost always positive: those who take their medicine in a randomized trial tend to be healthier. IV using the randomly assigned treatment intended as an instrumental variable for treatment received solves this sort of compliance problem. Moreover, LATE is the effect of treatment on the treated.

Formally, let the assumption of the LATE theorem hold and let $\mathbb{P}(D = 1 \mid Z = 0) = 0$. Then

$$\frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[D \mid Z = 1]} = \mathbb{E}[Y^1 - Y^0 \mid D = 1]$$

Proof
First, consider

$$\mathbb{E}[Y \mid Z = 1] = \mathbb{E}[Y^0 + (Y^1 - Y^0)D \mid Z = 1].$$

Since $\mathbb{P}(D = 1 \mid Z = 0) = 0$ (or equivalently, $\mathbb{P}(D = 0 \mid Z = 0) = 1$), we have that

$$\mathbb{E}[Y \mid Z = 0] = \mathbb{E}[Y^0]$$

Taking the difference between the two, we obtain

$$\begin{aligned}\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] &= \mathbb{E}[(Y^1 - Y^0)D \mid Z = 1] \\ &= \mathbb{E}[(Y^1 - Y^0)D \mid D = 1, Z = 1]\mathbb{P}(D = 1 \mid Z = 1) \\ &+ \mathbb{E}[(Y^1 - Y^0)D \mid D = 0, Z = 1]\mathbb{P}(D = 0 \mid Z = 1) \\ &= \mathbb{E}[Y^1 - Y^0 \mid D = 1]\mathbb{P}(D = 1 \mid Z = 1)\end{aligned}$$

We know that

$$\mathbb{E}[D \mid Z = 1] = \mathbb{P}(D = 1 \mid Z = 1).$$

The result obtains after dividing both parts.

In other words, the average treatment effect on the treated is equal to the *intention-to-treat* (ITT) divided by compliance.