LECTURE 6. COMPLIANCE, INSTRUMENTAL VARIABLES & NATURAL EXPERIMENTS

Dr. Tom O'Grady

The method of instrumental variables has a relatively long history in the social sciences. It is used in two different settings that turn out to be very conceptually related: experiments with non-compliance, and in many natural experiments where some but not all units end up receiving the treatment, or units receive different amounts of the treatment. We'll begin with one-sided non-compliance, then we'll shift to two-sided non-compliance, and then we'll examine natural experiments, with the complexity increasing as we go along.

1. One-Sided Non-Compliance

One-sided non-compliance occurs in an experiment when all of the control units comply with their assignment, but some of the treated units do not take the treatment or take the control instead. By far the most common scenario is one where the treatment is available only through the experiment and the control consists of not receiving anything, or receiving a placebo. One-sided non-compliance is very prevalent in these settings because taking the treatment is much more onerous than taking the control (e.g. in a medical setting, taking the treatment may be risky or could induce unwanted side-effects), and because by definition the control units simply cannot access the treatment. Another common example is in experiments on 'get out the vote' techniques with political parties that measure whether techniques such as phone canvassing increase voter turnout. The control group typically receive no intervention whilst some of the treatment group do not answer the phone call. Of course, in either situation we cannot compel anyone to comply with the experiment against their will. This means that dealing with one-sided non-compliance is a completely normal situation for experimenters in the social sciences. Happily, it is also quite easy to deal with statistically.

As a running example, we will look at a well-known study in public health that helped raise awareness of how to deal with one-sided non-compliance.[1] In this experiment, 450 Indonesian villages were randomly divided into treatment and control groups. In treatment villages, children were provided with Vitamin A supplements, whereas in control villages they were provided with them a year later. During the intervening year, mortality in the treatment and control villages was compared. The authors speculated that Vitamin A deficiency could be linked to childhood mortality because it increases susceptibility to infections, amongst other effects. However, around 20% of children assigned to receive the supplement did not take it, whereas

[1] Alfred Sommer *et al* (1986). "Impact of Vitamin A Supplementation on Childhood Mortality: A Randomised Controlled Community Trial." *The Lancet* i, 1196-1173. Alfred Sommer and Scott L. Zeger (1991). "On Estimating Efficacy from Clinical Trials." *Statistics in Medicine* 10 (1): 45-52.

all control children did not take it for the first year, as intended. This was therefore a case of one-sided non-compliance.

1.1. Types of Subjects under One-Sided Non-Compliance

In this situation, the experimental units must be one of two *types* that are defined by their potential outcomes, “compliers” or “never-takers”:

- **Compliers** are units that always take the treatment when assigned to the treatment group, and always don’t take the treatment when assigned to the control group
- **Never-Takers** are units that always do not take the treatment, regardless of whether they are assigned to treatment or control

In the Vitamin A example, compliers always take the vitamin A supplement when offered and always do not take it when not offered it. Note that these groups are defined as innate characteristics. Being a complier or a never-taker is akin to being male or female, or black or white. They are separate sub-populations, who may be very different from one another. Importantly, observed data does not fully reveal to us which units are compliers and which are never-takers. We’ll say that Z_i defines i’s treatment *assignment* and D_i defines their actual treatment *status*. Suppose we have a situation like the following:

i	Y_i	Z_i	$D_i(0)$	Type
1	1	1	1	Complier
2	3	1	1	Complier
3	2	1	0	Never-Taker
4	7	0	0	Complier or Never-Taker
5	3	0	0	Complier or Never-Taker
6	6	0	0	Complier or Never-Taker

We observe three groups:

	$Z_i = 1$	$Z_i = 0$
$D_i = 1$	Complier	
$D_i = 0$	Never-Taker	Complier or Never-Taker

Amongst the group that were randomly assigned to treatment ($Z_i = 1$), we know that all those who actually *received* treatment must be compliers. But amongst the group assigned to control, ($Z_i = 0$) we do not know which units are compliers and which are never-takers, because we do not also observe them in the treatment condition.

1.2. Treatment Effects under One-Sided Non-Compliance with Instrumental Variables Estimation

Up to now in this course, we’ve been considering how to find the Average Treatment Effect for all units, defined in terms of potential outcomes (for N experimental units) as:

$$\tau_{ATE} \; = \; \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes of units that *take* the treatment and control. We would estimate this in a standard experiment using the difference in means between treated and control units, since under random assignment they are interchangeable, having the same potential outcomes under treatment and control, in expectation. The control group serves as a counterfactual for the treated units.

But we can't do this under non-compliance. A simple ATE that compared those who actually took the treatment to those who did not is likely to be biased, because we are no longer comparing apples to apples. Instead, we would be comparing compliers in the former group to a mixture of compliers and never-takers in the latter, who may have very different average potential outcomes. Those who received the control no longer tell us what would have happened to the units who took the treatment, had they not taken it. For instance, children who always refuse to take vitamins even when offered them may not be a good counterfactual for those who always comply with instructions. An obvious possibility is that the compliers have higher potential outcomes than never-takers. They may be more healthy in many unobserved ways, in another example of the 'healthy user bias' that we talked about in week 2.

Instead, under non-compliance it turns out that we can in fact estimate a slightly different quantity known as the Complier Average Causal Effect.

> **Complier Average Causal Effect (CACE)**: The causal effect of the treatment on the outcome amongst compliers only

While not the same as the ATE for all units, the ATE for the sub-group of compliers remains a useful quantity because it tells us the average treatment effect for the units that do in fact take the treatment when it's offered to them. It is defined as:

$$CACE = \frac{1}{Q}\sum_{i=1}^{Q}[Y_i(1) - Y_i(0)] \mid i \; is \; a \; complier \tag{1}$$

where there are Q compliers amongst the N units. How can we estimate this? It isn't directly estimable using observed outcomes (or even in theory), because we don't know which units are compliers and which are never-takers. The group assigned to control consists of both types mixed together. However, it turns out that we can use a neat trick to find this quantity indirectly by using a technique known as Instrumental Variables estimation. The trick is based on the fact that the CACE is actually the ratio of two quantities that we can estimate from the data in an unbiased way. Z (the treatment assignment) is known as an Instrumental Variable.

> **Instrumental Variable:** A variable Z that *encourages* units to take a treatment, but does not force them to do so.

These are often simply referred to as an "instrument" for short. They should be thought of as a randomised encouragement to receive treatment. The key thing is that due to non-compliance, treatment *uptake* D cannot be considered to be randomly assigned, but treatment *assignment* Z was randomly assigned. Using this, we can estimate a quantity from our data that is known in the medical literature as the Intent-to-Treat Effect.[2]

> **Intent-to-Treat (ITT) Effect:** The causal effect of being assigned to treatment versus being assigned to control

In other words, this estimates the following quantity for unit i:

$$\tau_{ITT} \;=\; \frac{1}{N}\sum_{i=1}^{N}[Y_i(1_a) - Y_i(0_a)]$$

where $Y_i(1_a)$ and $Y_i(0_a)$ refer to potential outcomes when i is *assigned* to treatment and control. Hence it is just the expected difference in potential outcomes between being encouraged (being assigned to treatment, $Z = 1$) and unencouraged (being assigned to control, $Z = 0$).

This is called the intent-to-treat effect because it measures the causal effect of intended treatments Z, not actual treatment outcomes D. The ITT effect could be interesting in its own right as an estimate of the average total effect of rolling out a treatment, especially if the experimental setting closely mimics the real-world situation where we would do the roll-out. But that is rare. Usually it is not the quantity that we ultimately wish to estimate, because it does not tell us what impact the treatment has on those who actually take it. In the Vitamin A example, a doctor deciding whether or not to prescribe Vitamin A to a new patient needs to know what effect it is likely to have if the patient actually takes the drug (complies).

To estimate the CACE, note a key fact: *only compliers contribute to the ITT effect.* By definition, for all never-takers $Y_i(1_a) - Y_i(0_a) = 0) = 0$: they always reveal the same potential outcome (the potential outcome under control) whether encouraged or unencouraged, because they always take the control. However, compliers reveal their potential outcome under treatment when encouraged and their potential outcome under control when unencouraged. Therefore they alone contribute anything to the ITT, and τ_{ITT} actually consists of a series of terms like:

$$\begin{aligned}\tau_{ITT} \;&=\; \frac{1}{N}\sum_{i=1}^{N} 0 + [Y_2(1) - Y_2(0)] + 0 + [Y_4(1) - Y_4(0)] + \ldots \\ &=\; \frac{1}{N}\sum_{i=1}^{Q} Y_i(1) - Y_i(0) \mid \textit{i is a complier}\end{aligned}$$

where in the first line, units 1 and 3 are never-takers and units 2 and 4 are compliers. We are getting close to the quantity we want, the CACE in equation 1. However, the ITT is clearly too small. Instead of dividing by Q (the total number of compliers), we are dividing by N (the total number of all units). To get the CACE in equation 1, therefore, all we have to do is multiply the ITT by (N/Q), or equivalently, divide by (Q/N), the Proportion of Compliers.[3]. This makes intuitive sense. Because not all of the group who are encouraged to take the treatment actually take it, we should expect the ITT effect to be strictly smaller than the CACE. All we are doing here is up-weighting the ITT effect to get back the CACE.

[2] Gerber and Green call this the *ITT*

[3] Gerber and Green call this ITT_D

Proportion of Compliers: The proportion of the population that is comprised of compliers, for a given experiment

This means that the Complier Average Causal Effect can be estimated with an important equation known as the **Wald Estimator**:

$$CACE = \frac{Intent\ to\ Treat\ Effect}{Proportion\ of\ Compliers}$$

Why does this trick work? Because we have managed to uncover a causal effect for compliers due to the fact that the ITT Effect is only affected by compliers. And crucially, although we never know which units are compliers and which are not, we can estimate both the ITT Effect and the Proportion of Compliers in an unbiased way from our data. Thanks to the random assignment of Z, the unencouraged units are a valid counterfactual for the encouraged units, and so we can just estimate the ITT as the difference in means between encouraged and unencouraged units in terms of observed outcomes:

$$\hat{\tau_{ITT}} = \frac{1}{m}\sum_{i=1}^{m}(Y_i \mid Z_i = 1) - \frac{1}{N-m}\sum_{i=m+1}^{N}(Y_i \mid Z_i = 0)]$$

where m units are randomly assigned to encouragement ($Z = 1$) and $N-m$ to non-encouragement ($Z = 0$). The proportion of compliers can be estimated simply as:

$$\frac{1}{m}\sum_{i=1}^{m}(D_i \mid Z_i = 1)$$

because by definition under one-sided non-compliance, we know that everyone who actually took the treatment ($D = 1$) must be a complier.

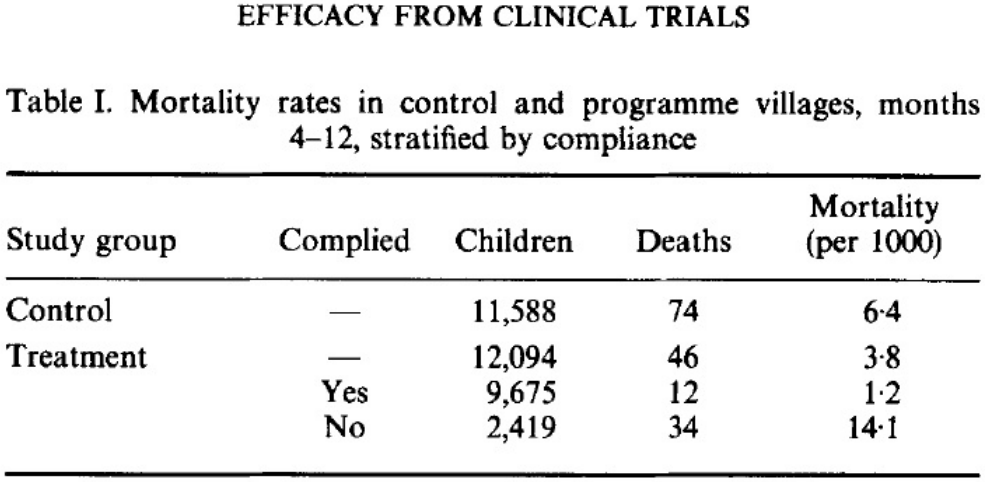
As an example, we can now calculate these quantities for the Vitamin A study introduced earlier. The results of the experiment are shown in Figure 1 from Sommer *et al* (1986).

Y here is a binary variable equalling 1 if the child died and 0 otherwise. In the group assigned to control, 74/11588 children died, a mortality rate of 0.639%. In group assigned to treatment, the equivalent figure is 46/12094, a mortality rate of 0.380%. Therefore:

$$\begin{aligned} Intent\ to\ Treat\ Effect &= \frac{1}{m}\sum_{i=1}^{m}(Y_i \mid Z_i = 1) - \frac{1}{N-m}\sum_{i=m+1}^{N}(Y_i \mid Z_i = 0)] \\ &= \frac{46}{12094} - \frac{74}{11588} \\ &= 0.38 - 0.639 \\ &= -0.26 \end{aligned}$$

The intent-to-treat effect is -0.26 percentage points, or a 40% (0.26/0.639) reduction in mortality amongst those who were assigned to treatment. However, this underestimates the

Figure 1: **Results of the Vitamin A Experiment**



EFFICACY FROM CLINICAL TRIALS

Table I. Mortality rates in control and programme villages, months 4–12, stratified by compliance

Study group	Complied	Children	Deaths	Mortality (per 1000)
Control	—	11,588	74	6·4
Treatment	—	12,094	46	3·8
	Yes	9,675	12	1·2
	No	2,419	34	14·1

treatment effect for those who actually complied. Instead, we form the Complier Average Causal Effect:

$$\begin{aligned} CACE &= \frac{ITT}{Proportion\ of\ Compliers} \\ &= \frac{-0.26}{(9675/12094)} \\ &= \frac{-0.26}{0.8} \\ &= -0.32 \end{aligned}$$

Therefore the treatment of Vitamin A reduced mortality by 0.32 percentage points (or 50%) amongst those who actually took the supplement. It was extremely effective.

2. Two-Sided Non-Compliance

The analysis of experiments with two-sided non-compliance is extremely similar to the case of one-sided non-compliance, but with some added complexity. Here, some control units do not comply with their assignment to control by taking the treatment instead, in addition to some treatment units failing to take the treatment. Such a situation is common in field experiments of policies like training programs that may be highly desirable and available to the control group outside the experiment, perhaps from another provider. Now, in addition to Compliers and Never-Takers we must add two more groups of subjects:

- **Always-Takers** are units that always take the treatment, regardless of whether they are assigned to treatment or control

- **Defiers** are units that take the control if assigned to treatment and the treatment if assigned to control

Suppose we have a situation like the following:

i	Y_i	Z_i	$D_i(0)$	Type
1	1	1	1	Complier or Always-Taker
2	3	1	1	Complier or Always-Taker
3	2	1	0	Never-Taker or Defier
4	7	0	0	Complier or Never-Taker
5	3	0	1	Always-Taker or Defier
6	6	0	0	Complier or Never-Taker

Again, observed data cannot tell us which unit is of which type. We observe four groups:

	$Z_i = 1$	$Z_i = 0$
$D_i = 1$	Complier or Always-Taker	Always-Taker or Defier
$D_i = 0$	Never-Taker or Defier	Complier or Never-Taker

Always-Takers and Defiers could not exist under one-sided non-compliance because it was impossible to take the treatment if assigned to control. We cannot make any progress here without invoking an assumption, monotonicity.

Monotonicity: There are no defiers in the population

Montonicity gets its name because it restricts the instrument Z to having a positive impact on treatment uptake only. It must not be the case that receiving an encouragement to take the treatment makes someone more likely to take the control, or vice versa. In general, this is not a particularly implausible assumption. It is hard to dream up scenarios where people always try to do the opposite of what they are encouraged to do, unless we decide to experiment exclusively on teenagers. Under this assumption, suppose we have a situation like the following:

i	Y_i	Z_i	$D_i(0)$	Type
1	1	1	1	Complier or Always-Taker
2	3	1	1	Complier or Always-Taker
3	2	1	0	Never-Taker
4	7	0	0	Complier or Never-Taker
5	3	0	1	Always-Taker
6	6	0	0	Complier or Never-Taker

Now, we are left with:

	$Z_i = 1$	$Z_i = 0$
$D_i = 1$	Complier or Always-Taker	Always-Taker
$D_i = 0$	Never-Taker	Complier or Never-Taker

Monotonicity matters because it allows us to once again estimate the Wald Estimate as:

$$CACE = \frac{Intent\ to\ Treat\ Effect}{Proportion\ of\ Compliers}$$

The intent-to-treat effect can still be measured under random assignment as the average difference in outcomes between the encouraged ($Z = 1$) and the non-encouraged ($Z = 0$). The proportion of compliers can be calculated as:

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

In words, this is simply the difference in treatment uptake between the encouraged and unencouraged. The only units that contribute to the two sums that make up this estimate are those for whom $D = 1$. The left-hand sum counts the top-left cell of the table above, compliers and always-takers, because some of those assigned to treatment would have taken it even if they had been assigned to control. The right-hand sum counts the top right of the table, always-takers, because under monotonicity they are the only people who can take the treatment when assigned to control. Due to random assignment, in expectation there should be an equal number of always-takers in the $Z = 1$ and $Z = 0$ groups. This is just another statement of balance under random assignment. Therefore the difference between the two sums is an unbiased estimate of the proportion of compliers under two-sided non-compliance.

As an example, we'll look at another study from public health, this time looking at the impact of flu vaccines on flu-related hospitalizations by Hirano *et al* (2000).[4] They examined an experiment where a set of doctors were randomly assigned to receive a letter reminding them to offer flu vaccines, with the control group receiving no letter. This is a situation of two-sided non-compliance, because some doctors who received the letter did not provide the vaccine, and other doctors who did not receive the letter did provide the vaccine. These are the results (by doctor):

The outcome Y here is hospitalization rates. Therefore the ITT is:

$$\begin{aligned} Intent\ to\ Treat\ Effect &= \frac{1}{m}\sum_{i=1}^{m}(Y_i \mid Z_i = 1) - \frac{1}{N-m}\sum_{i=m+1}^{N}(Y_i \mid Z_i = 0)] \\ &= 0.078 - 0.092 \\ &= -0.014 \end{aligned}$$

[4] Keisuke Hirano, Guido W. Imbens, Donald B. Rubin and Xiao-Hua Zhou (2000). "Assessing the effect of an influenza vaccine in an encouragement design." *Biostatistics* 1 (1): 6988

Figure 2: **Results of the Influenza Vaccine Experiment**

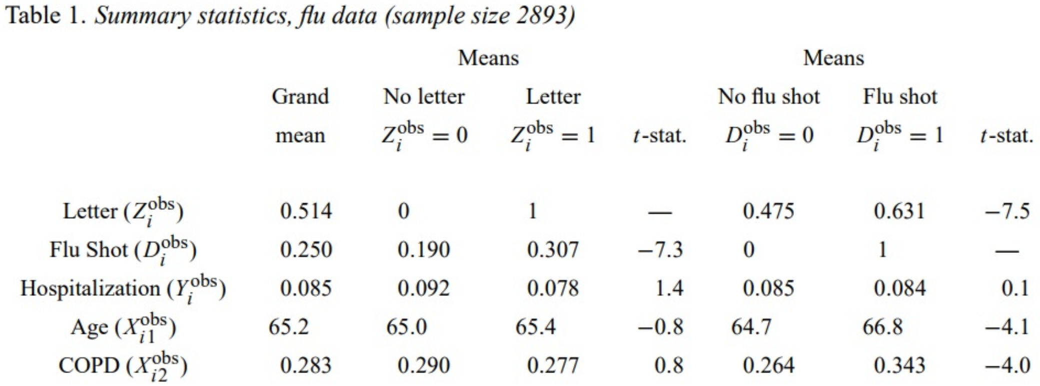


Table 1. *Summary statistics, flu data (sample size 2893)*

		Means			Means		
	Grand mean	No letter $Z_i^{\text{obs}} = 0$	Letter $Z_i^{\text{obs}} = 1$	t-stat.	No flu shot $D_i^{\text{obs}} = 0$	Flu shot $D_i^{\text{obs}} = 1$	t-stat.
Letter (Z_i^{obs})	0.514	0	1	—	0.475	0.631	−7.5
Flu Shot (D_i^{obs})	0.250	0.190	0.307	−7.3	0	1	—
Hospitalization (Y_i^{obs})	0.085	0.092	0.078	1.4	0.085	0.084	0.1
Age (X_{i1}^{obs})	65.2	65.0	65.4	−0.8	64.7	66.8	−4.1
COPD (X_{i2}^{obs})	0.283	0.290	0.277	0.8	0.264	0.343	−4.0

a 15% (0.014/0.092) reduction in hospitalization as a result of the encouragement to provide the vaccine. The proportion of compliers is:

$$\begin{aligned} Prop.\ of\ Compliers &= \frac{1}{m}\sum_{i=1}^{m}(D_i \mid Z_i = 1) - \frac{1}{N-m}\sum_{i=m+1}^{N}(D_i \mid Z_i = 0)] \\ &= 0.307 - 0.19 \\ &= 0.117 \end{aligned}$$

The way to interpret this is that 30.7% of the sample are a mix of compliers and always-takers, while 19% are always-takers only: 19% of doctors would have provided the flu vaccine regardless of receiving a letter. Thus 11.7% of the population are estimated to be compliers since 19% of the 30.7% are always-takers. We can do this thanks to random assignment, assuring that the groups assigned to treatment and to control contain the same number of always-takers in expectation.

Therefore the CACE is:

$$\begin{aligned} CACE &= \frac{ITT}{Proportion\ of\ Compliers} \\ &= \frac{-0.014}{(0.117)} \\ &= -0.079 \end{aligned}$$

Therefore amongst the doctors who are compliers, the flu vaccine led to a 7.9 percentage-point (85%) reduction in hospitalization: far larger than the intent-to-treat effect. The reason for this large disparity is that only a small fraction of the population are estimated to be compliers in this case.

3. Natural Experiments and Instrumental Variables

So far we have covered instrumental variables as a solution to non-compliance in experiments. But in the social sciences today, instrumental variables estimation is much more famous (or perhaps infamous, as we'll see) as a way to analyse natural experiments. Natural experiments occur when some treatment is randomised without the intervention of the analyst. Sometimes this occurs because a government deliberately randomises something, for example by providing an over-subscribed treatment by lottery, as we will see below in the example of the Hajj Pilgrimage. But more often, natural experiments occur because some treatment is assigned arbitrarily or accidentally. In these cases, the mechanism by which treatment is assigned is just an instrumental variable, i.e. an encouragement to take the treatment. Some examples include:

1. **The Number of Children and Labor Supply:** Are people likely to stop work when they have more children? A simple comparison of parents and non-parents or people with more vs. less children would be problematic, because the two groups have different potential outcomes. Instead, Angrist and Evans (1998)[5] use having two first children of the same sex as an instrument that randomly encourages people to have larger families. Childrens' sexes are randomly assigned, and having two children of the same sex may encourage parents to try for a third child of a different sex.

2. **GDP and Civil War Onset:** Are civil wars more likely to start in poorer countries? A simple comparison of poorer and wealthier countries would be problematic. Miguel, Satyanath and Sergenti (2004)[6] use high rainfall as an instrument that randomly 'encourages' some third-world countries to have high GDP. Rainfall is randomly assigned, and in high-rainfall years, countries that are heavily agricultural tend to experience higher economic growth due to greater crop yields.

3. **Wind Farms and 'Electoral Backlash':** Do climate-change mitigation policies like wind farms generate local electoral backlashes against politicians who enact them? A simple comparison of places with and without wind farms would be problematic. Stokes (2016)[7] uses high localised wind speed as an instrument that randomly encourages Canadian politicians to build wind farms in communities that happen to have high prevailing winds. Wind speed is probably randomly assigned, as an accident of geography.

[5] Joshua Angrist and Bill Evans (1998). "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review* 88 (3):450-477

[6] Edward Miguel, Shanker Satyanath and Ernest Sergenti (2004) "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy* 112 (4): 725-753

[7] Leah Stokes (2016). "Electoral Backlash against Climate Policy: A Natural Experiment on Retrospective Voting and Local Resistance to Public Policy." *American Journal of Political Science* 60 (4): 958-974

We can summarise these examples in a table:

Study	Outcome	Treatment	Instrument
Angrist and Evans (1998)	Labour supply	Having more children	Two children of same sex
Miguel, Satyanath and Sergenti (2004)	Civil war onset	GDP growth	Rainfall
Stokes (2016)	Backlash against incumbent politicians	Building wind farms	Wind speed

The reason why natural experiments with instrumental variables have often been controversial in the social sciences is that they rely on a number of key assumptions that are often violated. We will look at those in the next lecture in a lot of detail. For today, the key point is that these natural experiments can be analysed in exactly the same way as in randomised experiments. In fact, they are just like randomised experiments with two-sided non-compliance. In each case, some of the group that randomly receive the encouragement will not take the treatment. Likewise, some of the group that do not randomly receive the encouragement will take the treatment. For instance, in Angrist and Evans (1998) some parents with two children of the same sex will not choose to have a third child, while some parents with two children of different sexes will still have a third child. Parents who would always have a third child when their first two are of the same sex and would always not have a third child when their first two are of a different sex are compliers. In Miguel, Satyanath and Sergenti (2004), some countries with high rainfall will not have high GDP growth, and some countries with low rainfall will have high GDP growth. Countries that would always have high GDP after high rainfall and would always not have high GDP after low rainfall are compliers.

One difference, though, is that because the use of instrumental variables in natural experiments grew up quite separately from the literature on randomised experiments, different terminology is often used for the same phenomena. In studies of natural experiments in the social sciences:

- **"First-Stage" = proportion of compliers**. The reason for this will become clear below, when we cover the method of two-stage least-squares
- **"Reduced Form Estimate" = intent-to-treat effect**
- **"Local Average Treatment Effect" = complier average causal effect**, because the estimated treatment effect is "local" only to those units that change their treatment status as a result of the encouragment

Now we'll look in more detail at a nice example of a natural experiment from Pakistan.[8] Every year, around two million Muslims participate in the Hajj pilgrimage to Mecca. For many pilgrims, particularly those from non-Western countries, the Hajj may be their first experience of foreign travel or mingling with people from other countries and strands of Islam. An obvious

[8] Daniel Clingingsmith, Asim Ijaz Khwaja and Michael Kremer. "Estimating the Impact of The Hajj: Religion and Tolerance in Islam's Global Gathering." *Quarterly Journal of Economics 124 (3): 1133-1170*

question is how the experience affects them: does it, for example, make them more tolerant of other cultures and belief systems? Answering this question with observational data is unlikely to provide valid causal effects because the potential outcomes of those who choose to go on the Hajj are likely to be very different to those who do not. Instead, Clingingsmith, Khwaja and Kremer take advantage of a natural experiment arising from Pakistani government policy. Due to concerns about over-crowding and crowd safety, Pakistan strictly limits the number of visas it provides each year for its citizens to go on the hajj in Saudi Arabia. In fact, it only issues 150,000 personal visas per year. Because demand outstrips supply, it allocates them via a lottery. This is a natural experiment that can be analysed with instrumental variables:

- The lottery is an instrument because winning it provides a randomised encouragement to go on the Hajj
- Actually going on the Hajj is the treatment
- This is a natural experiment because the government introduced the visa lottery not in order to study the Hajj, but as a way to cope with demand hugely out-stripping supply
- This is a situation of two-sided non-compliance, because not everyone who wins the lottery actually goes on the Hajj, and not everyone who loses the lottery fails to go on it. Some visas are available through other means, such as via private tour operators
- Compliers are those who always go on the Hajj when they win the lottery and always do not go on the Hajj when they lose it
- The authors conducted a survey of 1600 Pakistanis recording whether they won the lottery, whether they went on the Hajj, as well as a range of opinion questions about their beliefs, conducted after the Hajj
- The outcome variables are a series of opinion measures including measures of tolerance for other countries and belief systems

4. Instrumental Variables with Continuous Treatments or Instruments: The Method of Two-Stage Least Squares

So far in this module we have only considered *binary* treatments that can only equal 0 or 1, depending on assignment to treatment or control. For the rest of the course, we will be leaving that behind, to also consider *continuous* treatments that can take a range of values.

One of the great things about instrumental variables analysis is that it naturally incorporates both treatment variables and instruments with a range of values. In such a case, *higher values* of the instrument encourage people to take on *higher values* of the treatment variable. Having one or both of the instrument or treatment as continuous is a common scenario in social science applications, and two of the examples above use continuous treatments or instruments in their ultimate models:

- Stokes (2016) actually uses a continuous instrument: wind speed. Faster wind speeds make it more likely that an area will receive treatment (a wind farm). Higher wind speeds can still be considered to be “as-if” randomly assigned to areas, because there is no reason to believe that areas with higher wind speeds have different potential outcomes (in terms of voting against the incumbent government) than areas with lower wind speeds.

- Miguel, Satyanath and Sergenti (2004) actually use both a continuous treatment *and* a continuous instrument: GDP growth and annual change in rainfall. Larger rises in rainfall make it more likely that countries will experience higher values of the treatment (higher GDP growth). Higher rainfall can still be considered to be "as-if" randomly assigned to countries, because there is probably no reason to believe that countries experiencing high changes in rainfall have different potential outcomes (in terms of susceptibility to civil war) than countries experiencing low changes in rainfall.[9]

With continuous treatments and/or instruments, we must use a slightly different method of analysis called two-stage least squares. Very often, authors will say that the instrument "induces exogenous variation" in their treatment variable. "Exogenous" means that the instrument causes the treatment to change in a way that is akin to randomisation. It induces some units to increase the values of their treatment in a way that is unconnected with (exogenous to) the units' potential outcomes. This is the key to understanding two-stage least squares. It proceeds in two stages, each of which is a regression:

1. In the **first stage**, we regress the **treatment** variable on the **instrument** and extract **predicted treatment values**:

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

2. In the **second stage**, we regress the **outcome** variable on the **predicted treatment values from the first stage**:

   $$Y_i = \gamma_0 + \gamma_1 \hat{D}_i$$

The idea is that in the second stage, we end up regressing Y on only the variation in D that is induced by random variation in the instrument. With a continuous treatment and a continuous randomly-assigned instrument:

- $\hat{\beta}_1$ measures the average amount of change in D induced by a one-unit change in Z
- Variation across units in the fitted values $\hat{D}_i$ reflect only the predicted amount of D, for each unit, that is caused by that unit's value of Z.
- $\hat{\gamma_1}$ provides an unbiased estimate of the effect of D on Y for those people for whom a shift of the instrument causes higher values of D.
- Thus we estimate a Local Average Treatment Effect only for those people whose treatment value is shifted by the instrument: akin to compliers in the binary case.
- As in any regression, $\hat{\gamma_1}$ should be interpreted as the predicted effect of a one-unit increase in $\hat{D}_i$ on the outcome Y.

[9] Although this is not as clear-cut as in the case of Stokes (2016). Perhaps you may be able to think of reasons why countries experiencing unusually high rainfall are actually more likely to have civil wars. You might also question whether higher rainfall always leads to higher growth (what about flooding?). We'll come back to these issues in Lecture 6.

$\hat{\gamma}_1$ is the key parameter of interest, akin to the Complier Average Treatment Effect. Here it is called the **Local Average Treatment Effect**. The first stage tells us how strongly the instrument affects the treatment. The higher is $\hat{\beta}_1$, the more strongly the instrument encourages people to take higher values of the treatment.

In fact, in the case where both the treatment and the instrument are binary, it is still perfectly possible to use two-stage least squares to estimate the CACE. This gives a completely identical answer to the Wald Estimate, with the added advantage of giving us a standard error for the CACE to enable hypothesis testing. The first stage is exactly identical to the proportion of compliers. Hence, the term "first stage estimate" is often used in the social sciences. The proportion of compliers is itself just a measure of the instrument's strength: the more that people comply with the randomised encouragement, the stronger it is. Therefore there is a very close analogy between the continuous and binary cases. Finally, although it is not necessary to do so in the case of two-stage least-squares, one can also estimate the ITT effect – typically called the **reduced form** estimate by regressing the outcome on the instrument.

Finally, two-stage least squares is even more flexible than this. It can accomodate two further extensions of instrumental variables analysis:

1. We can include **additional covariates**, just as we sometimes did when analysing experiments. This may be useful when we think that the instrument is only randomly assigned conditional on some covariate(s). Including covariates in the first-stage regression is just like fixing failures of randomisation in experiments. Covariates will make very little difference, however, if the instrument is genuinely randomly assigned.

2. We can include **multiple instruments**, if we think there is more than as-if-randomly-assigned variable that encourages higher values of the treatment. In the first stage, variation in the estimated fitted values reflects the influence of these multiple instruments on the treatment

One final complication is that the standard error on $\hat{\gamma}_1$ will be slightly wrong if we actually carry out two-stage least-squares manually in two stages. Modern econometric software includes a correction for the standard errors and carries out both stages at once. It should always be used. In R, we use the `ivreg()` function in the `AER` package with the code:

```
ivreg(outcome ~ treatment | instrument)
```

With additional covariates, they need to be added to both parts of the code:

```
ivreg(outcome ~ treatment + covariates | instrument + covariates)
```

5. The Next Lecture

Instrumental variables estimation is both complex as well as controversial in the social sciences. Because of that, we are spending two lectures discussing it. In the next lecture, we will focus on where the controversy lies. Valid estimation in the instrumental variables framework actually relies on four key assumptions, only one of which we have mentioned in any detail:

1. The instrument is **relevant**, meaning that it affects the treatment quite strongly. Angrist and Pischke call this the first-stage assumption.

2. The instrument is genuinely **as-if randomly assigned**. Angrist and Pischke call this the ‘independence’ assumption.

3. The **exclusion restriction** holds, meaning that the instrument affects the outcome *only* via its influence on the treatment.

4. **Monotonicity**, as defined earlier.

Instrumental variables estimation is controversial because in natural experiments, some of these assumptions may be invalid. Next time we’ll look at a range of studies that, to varying degrees, do and do not meet these assumptions. You should come to class ready to discuss them, having read in advance about the four assumptions and having thought about whether the papers meet the assumptions, or not.