Andreas Kotsadam´s Lecture Notes on Empirical Methods in Development Economics.

These notes start with slide 8 from lecture 3 and go to slide 71.
Slide: The fundamental problem of causal inference
In order to answer any causal questions such as what is the effect of opening banks in rural areas on poverty or do institutions of private property cause growth, we must answer a counterfactual question: What would have happened if banks were not opened in areas where they were opened and what would happen if they were opened in areas where they were not opened?
The problem is of course that at any given time, places were either exposed to the program or they were not.
It is the same thing if we want to know the effect of primary education on women's bargaining power. At any point in time, the individual woman either has primary education or she has not. We do not have a counterfactual for the individual woman.

We solve this by using statistics, that is by collecting data on several individuals. If we compare groups of individuals that are otherwise similar, except for being exposed to the treatment, we can get the average effect of the program.

Slide We need a comparison group

- ...that would have had similar outcomes as the treatment group if there was no treatment.
- In general, however, those receiving treatment and those that do not usually differ due to:

In general, social programs target poorer individuals: Think of the banking example again, banking programs are targeting poorer areas.
people are often screened on the basis of need or motivation,
and decision to participate is often voluntary creating self-selection. Families chose to send their girls to school and different countries chose to have different institutions: Often these choices are correlated with outcomes measured at a later date.

Slide: This implies
- ...that those not exposed to a treatment are often a lousy comparison group.
- It is often impossible to disentangle treatment effects from selection bias.

Slide Example:
- A fertilizer program where fertilizers are given for free to some farmers.

Slide:
Effect=
Yield for the farmers who got fertilizer
-
Yield at *the same point in time* for *the same farmers* in absence of the program.

Slide Problem
We never observe the same individual with and without program at the same point in time.

Slide Not before after:

- Other things may happen over time so that we cannot separate the effect of the treatment and the effect of those other things.

The rest of the world moves on and you are not sure what was caused by the program and what by the rest of the world

- Even if *you* know ”nothing else happened” it is hard to convince others.
- The burden of proof is on you.

Slide: not compare with those without fertilizers

- Some may *choose* not to participate.

Those *not offered* the program may differ. Participants will differ from non-participants in observable and unobservable ways

- Again, the burden of proof is on you.

Slide Solution

- Find a good proxy for what would have happened to the outcome in the absence of program
- Compare the farmer with someone who is exactly like her but who was not exposed to the intervention
- In other words, we must find a valid Counterfactual

So that the only reason for different outcomes between treatment and counterfactual is the intervention

Slide: The potential outcomes framework

A very influential framework for thinking about these problems is offered in the potential outcomes framework, or Rubin’s causal model, after Rubin.

Suppose we want to know the effects of textbooks on test scores.

Let

Y_i = Observed test scores for school i. This is the outcome observed for the researcher.

Y_i^T = Average test scores of children in school i if the school has textbooks.

Y_i^C = Average test scores of children in *the same* school i if the school has no textbooks.

Knowing the causal effect of having textbooks on test scores in school i imply measuring the difference:

$$Y_i^T - Y_i^C$$

Slide: The problem
The problem is that every school has two potential outcomes and we only observe one of them.

This is just a restatement of the fundamental problem: We are obviously not able to observe school i both with and without textbooks at the same time. Therefore we cannot estimate individual treatment effects.

Slide: By using data on many schools we can do better

We can then hopefully learn the expected average effect of textbooks on test results:
$$E\left[Y_i^T - Y_i^C\right]$$

Ok, so some schools have textbooks and some schools have not. If we take the average of both types of schools and examine the difference in text scores we get:
$$D = E\left[Y_i^T \middle| T\right] - E\left[Y_i^C \middle| C\right]$$

Difference= Expected test scores for the treated conditional on being in the treatment group - Expected test scores for the control conditional on not being treated.

Slide:

You can always subtract and add a term to an expression and the result stays exactly the same. This is sometimes a useful trick.

Subtracting and adding $E\left[Y_i^C \middle| T\right]$: The expected outcome for a school in the treatment group that is not treated. This term is not observed but it is logically well defined and helps us fixing ideas. This gives:

$$D = E\left[Y_i^T \middle| T - Y_i^C \middle| T\right] - E\left[Y_i^C \middle| C - Y_i^C \middle| T\right]$$

This is the same as:
$$D = E\left[Y_i^T - Y_i^C \middle| T\right] + E\left[Y_i^C \middle| T\right] - E\left[Y_i^C \middle| C\right]$$

Slide: Let's have a closer look:
The first term here, $E\left[Y_i^T - Y_i^C \middle| T\right]$ is the treatment effect we are interested in. It is ToT, treatment effect on the treated. It answers the question: What difference did the textbooks make in the treatment schools.

The second term, $E\left[Y_i^C \middle| T\right] - E\left[Y_i^C \middle| C\right]$, is the selection bias (note that E(Yc given T is interpreted as "If the treatment group people would not have been treated"). It captures the idea that the treatment schools may have had different test scores on average even if they had not been treated. It is the difference in potential untreated outcomes between treatment and comparison schools.

Slide:
Give examples of such selection effects:

E.g. richer schools have textbooks, parents in some schools consider education to be more important and want the schools to have textbooks, and these parents may be more likely to help their children with homeworks or to prepare for the exam.
E(Yc given T) would be larger than EYc given C.
The bias can also work in the other direction if the government decides to give textbooks to particularly disadvantaged areas.
It could also be the case that textbooks are part of a larger policy intervention so that all schools that get textbooks also get more teachers. Then the effect of having more teachers would also be embedded in D.
Slide
So, the general point: In addition to the effect of textbooks there may be other systematic differences between schools with and without textbooks.
The problem is of course that $E\left[Y_i^C \middle| T\right]$ is not observed. Therefore it is in general impossible to assess the magnitude or even the sign of the selection bias and hence, the extent to which selection bias explains the difference in outcomes between treated and not treated schools.

The goal is to find situations where selection bias does not exist or where we can correct for it.
Slide: Randomization
When individuals, or schools, or countries, are randomly assigned to treatment and comparison groups, the selection bias disappears.

Take a sample of N individuals from a population of interest. Divide the sample randomly into a treatment and a control group.

Slide
Then give the treatment group a treatment so that their treatment status is T and nothing to the control so that their treatment status is C. Collect outcome data Y and compare the treatment average to the control average.

Slide
The average treatment effect can then be estimated as the difference in empirical means of Y between the two groups. For a large enough sample the difference becomes:

$$D = E\left[Y_i^T \middle| T\right] - E\left[Y_i^C \middle| C\right]$$

Since treatment is randomly assigned, individuals assigned to treatment and control are only expected to differ through their exposure to treatment.

Had neither received treatment, their expected outcomes would have been the same.
This implies that the selection bias, $E\left[Y_i^C \middle| T\right] - E\left[Y_i^C \middle| C\right]$, is equal to zero.

Slide:
Assuming that there are no externalities/spillovers so that the treatment of one individual affects the behavior of another, SUTVA (the Stable Unit Treatment Value Assumption) we get:

$$E\left[Y_i \middle| T\right] - E\left[Y_i \middle| C\right] = \left[Y_i^T - Y_i^C \middle| T\right] = E\left[Y_i^T - Y_i^C\right]$$

That is our causal parameter of interest is obtained by simply comparing the means.

Slide: In a regression this is simply:

$$Y = \alpha + \beta T + \varepsilon$$

Where T is a dummy for belonging to the treatment group.

Slide: A detour on the law of large numbers.
The law of large numbers, says that if you randomly split *enough* people or villages or playing cards into groups, the groups will converge statistically.

- ” For a large enough sample… ”
- If we were to draw a line in the middle of India and randomly (e.g. by flipping a coin) provide microcredit in one part this would be a randomized field experiment.
- ”Large enough” depends on the variance and magnitude of the effects.

Slide
What is being estimated?

Note that what we get in a randomized experiment is the overall impact of a particular treatment on an outcome. Note in particular that we allow other things to change as a response to the program. It is not the all else equal effect. Ceteris paribus effect is something else...

So, in our textbook example, we do not get the effect of having textbooks all else equal, but rather we get the total effect of having textbooks. It may be that textbooks help people on the exam but that people compensate by not going to class and listen to the lecture and that they therefore get a bad exam result.

This is still an effect of the intervention, it is still a causal effect of providing textbooks.

Think of it this way: Everything that happens differently across the two groups after the treatment is the effect of the treatment.

This difference is very important and often causes confusion, even in published papers you observe that people have got this wrong. This difference is also part of the ongoing debate between reduced form econometricians and those that want more theory. This is the dividing line between e.g. Imbens and the ones he is writing against, in particular Deaton.

So, what we get is the "reduced form estimates" and these are total derivatives. Partial derivatives can only be obtained if you specify a model linking inputs to outputs and collect data on intermediary inputs.

Slide: So, main advantages of randomization:

A randomized evaluation provides internally valid estimates = It provides an unbiased estimate of the impact of the program in the sample under study.
The effect we claim to have is indeed caused by the intervention in the sample.

They are also easy to understand. Convincing: e.g. Progresa, the Mexican conditional cash transfer program has spread around the world.

Very good for testing theories. And doing so in an iterative process. First test broadly, simply is it working or not, then test all the mechanisms. Since you have control yourself when conducting a field experiment the individual elements can be varied in unique ways.

Slide: Critiques of experiments
There is a debate between those who think that there is too much focus on experimental and quasi experimental studies in economics in general, and in development economics in particular.
Deaton argues that development economics runs the risk of ignoring serious thinking about how and why things work and how things are related.
The trend towards credible causal inference makes researchers avoid certain questions.
I agree with this point, not everything can be randomized and it is important to have smart people thinking about questions that we cannot evaluate. Nonetheless, that should be a complement.

Most of this critique of experiments actually applies to most empirical research and they are often highlighted in dealing with experiments since the most important concerns are dealt with by randomization.

External validity: = Is the effect generalizable to other samples?
Internal validity is not a sufficient condition for external validity.

a) Environmental dependence: Would we get the same result if we conducted the same experiment in another setting? Would providing free school lunch have the same effect in Norway and in Kenya? Obviously not, but the trickier question is where to draw the line: Is Argentina more like Norway or Kenya? The same issues often apply within a country.

The thing is: Without assumptions, results from experiments cannot be generalized beyond their context.
It is often argued that environmental dependence is a more serious problem with experiments. E.g. Rodrik 2008.

Slide:
b) Implementer effects. The results may not generalize to other NGO's for example.
More problematic, not every NGO wants to be evaluated: Probably a selection of more competent NGOs and better programs!

Slide: But these issues apply to all empirical work

- Argentina is not more like Norway because we build a model.

Also implementer effects in general, e.g. countries with better institutions have better data for example. Nonetheless, if the program is scaled up, the competence may be lost.

Slide: More critique

- General equilibrium effects: What happens if we scale up a successful program?

For example, say we provide some children education. If we scale up and give all children education the returns to education would be lower, and we would not get the effects of having relatively more education.

Randomization bias: The fact that the program is evaluated using randomization affects behavior.

- Hawthorne effect. Being monitored changes behavior.

But these are too general, these say that behavior change since people know they are evaluated, in fact the Hawthorne experiment on working conditions was not randomized.

Solution: Replication studies and meta studies!

Slide ethics:

- Is randomization unfair?

  It may also cause troubles, but given that the program is limited randomization should not be worse? Sometimes, it is better to give to the worst off, we will come back to this point.

Why so many experiments from developing countries? Generous interpretation: The questions merit it and there is not a lot of data to work with. More cynical interpretation: It is cheap and feasible. Not that many issues with ethical committees and you can easily provide high stakes incentives etc.
Probably a combination of the two.

Why not more?

- Ignorance may have political advantages.
- Technical capacity may be limited.
- Benefits are not clearly appropriated to those who bear the costs: Evaluations as a public good.
- Or randomization is simply not feasible. We do not want to randomize diseases to people for instance.

Slide: If randomization is not possible
Other methods can be used to handle selection bias but they all require more assumptions. But the issue is still the same: Try to find comparison groups that are valid.

But, as opposed to randomization, we must now invoke some identifying assumptions.

These identifying assumptions are not testable and the validity of any particular study depends on how convincing these assumptions appear.

Slide:

The first method is simply controlled regression analysis:

If there exists some vector X such that,

$$E\left[Y_i^C \middle| X, T\right] - E\left[Y_i^C \middle| X, C\right] = 0$$

Then we can estimate the causal effect by including X as control variables in a regression.

That is, we may think it is obvious that some things are correlated with both institutions and growth, like trade, and then we control for this.

The problem is that we then have to assume that the relationship is causal once we have put in our controls.

Slide: This approach is therefore only valid if there is no difference in potential outcomes between treated and untreated individuals once we have controlled for the observable differences.

It is generally unlikely that this is enough since X must account for all the relevant observed and unobserved differences between the treatment and control groups.

IV:

- Very common method in empirical economics.
- We saw it B&P in lecture 2 and we will see it in several other papers during the course.
- A very good reference for IV is Murray (2006) ”The Bad, the Weak, and the Ugly: Avoiding the Pitfalls of Instrumental Variables Estimation”

Slide:

- What’s the problem?
- How can it be solved by IV?
- How is it done in practice? Examples.
- Instruments can be:

i) Bad,
ii) Weak,
iii) Ugly.

Slide: The problem:

- IV solves the problem of ”endogeneity”. OK, all clear then.
- Endogeneity: An explanatory variable is correlated with the error term.
- It is very common in social science.
- Most common reasons:

i) Omitted variables
ii) Measurement error
iii) Simultaneity (reversed causation)

Slide: A very common example in economics

- We want to estimate the returns to education.
- Wage= a+B1education+ B2X+ ei

We cannot measure ability. Ability is likely to increase the amount of education people take and the wage people gets. Since we do not measure it ends up up in ei

What does B1, the coefficient for education, then tell us?

An individual with higher ability has a more positive error term since ability is not included in the regression and it affects wages positively.

The individual also has higher education due to higher ability.

Hence, there is a correlation between an explanatory variable and the error term.

What is the problem with this? Well, we put part of the effect of ability on B1, B1 is then biased. We would say that education is more important than it actually is.

Slide: An example with simultaneity.
Example: We want to show that conflict is bad for GDP. So we estimate…

(i) $\mathrm{GDP} = \alpha + \beta_1 \mathrm{conflict} + \beta x + \varepsilon_1$ (1)

The problem is that GDP may affect conflict so that we actually have two equations:

(ii) $\mathrm{Conflict} = \delta + \mu_1 \mathrm{GDP} + \mu x + \varepsilon_2$ (2)

Suppose we have a random shock: Since ε_1 affects *GDP* in (1) which in turn affects *Conflict* through (2), it follows that ε_1 will be correlated with conflict in (1), and hence we have an endogeneity problem.

Again: The explanatory variable is correlated with the error term.

Slide: What can be done:

- To overcome the endogeneity problem we can use the Instrumental Variables (IV) approach.

Slide:
How does it work:
Let's take the education example again (equation 1 below). We need at least one variable, Z, that is correlated with education, but uncorrelated with the wage received other than through education.

- Frequently, regressions requiring IV estimation have a single troublesome explanator (education) and several non-troublesome explanators (Xi):

  $$Wage = b0 + b1Education + b2\,Xi + ei \qquad (1)$$

Once we have found Z, we estimate a regression for education with all unproblematic variables and Z as explanatory variables and we get predicted education.

This predicted value is then put in the original equation instead of actual education and gives a non biased estimate.

Slide: HENCE:

- To use the IV approach we need at least one additional variable, referred to as an instrument. The instrument has to satisfy two conditions:
- i) Relevance (easy to test)
- ii) Validity (cannot be tested)

The problem is that this is extremely difficult to find!

Slide: Proposed IV for edu

- Distance to college. This one was used early but would not work today. Argument is that wages are only affected by the higher probability of going to college induced by living close to one.
  It is likely that you live close to an educational institution since your parents are academics and such areas are often correlated with class. Maybe you learn middle class manners etc.

Quarter of birth with compulsory schooling. This instrument uses the fact that people had to stay in school until they are 16 years old.

Since different people turn 16 at different times, they will drop out of school having been there different amount of years. We will come back to this instrument.

Slide: Bad instruments:

- When the instruments are not valid.
- Remember that this cannot be tested.
- Overidentification (If you have more instruments, use only one of the instruments in 1^{st} stage and use the other as controls in the 2^{nd} stage. Then test if the "control" instruments are equal to zero, which they should be) tests are always used when possible but they can only help prove that an instrument is bad. E.g. all instruments may be bad, then the test is useless!

Slide: Weak instruments:

- We call an instrument weak if the correlation with the troublesome variable is low.
- One consequence is that the variance of 2SLS becomes greatly inflated.

Let us look at this with the help of so called Venn diagrams:

Slide
Here we see 2 circles. Y shows the variation in Y and the Circle X shows the variation in X. Where the circles overlap, the purple part, shows the correlation between X and Y.
(Conceptual regression: Y=a+Bx+e)
This can be interpreted as the information used to calculate Bx in a regression. A larger purple area means more information and hence less variance in the estimate of Bx. The black area is what we cannot explain, that is, the error term.

Slide: multiple regression:
Now we have 3 circles and we can imagine a multiple regression (Conceptual regression: Y=a+Bx+Gw+e). The area blue+red+green is the variation in Y that is explained by X and W. This area divided by the yellow area gives R^2.
OK, note the red area: When we only include X this information is used to estimate the effect of X on Y. But this is wrong, and this is precisely the omitted variables bias, we do not include all relevant information.

We actually do not know if the red area arises due to X or W so in a multiple regression we throw away this information.

So, we should include W as well. But including W costs, and this is the logic behind multicoliniarity: The more X and W overlap, the less information is left to estimate their effects and hence we get larger variance in the estimates.

Slide:
What does this have to do with instruments?
In the figure here we see that Y is determined by X and an error term. Disregard Z for a moment.
X is endogenous so that part of the error term overlaps with X, the red area.
How can we get rid of the red area? We use an instrument:

The circle Z is an instrument for X: It is correlated with X but uncorrelated with the error term.

In the first stage we run a regression where Z shall explain X. (Conceptual regression: X=L+BZ+u). We then take the predicted X value: Purple + Green.

In step 2 we then run a regression where the predicted value explains Y. The B value we get then is our IV estimate. (Conceptual regression: $Y=a+Bx_{predicted}+e$)

Where Y and predicted X overlap is the purple area so it is only information from this area that is used. The estimator is not correlated with the error term and hence it is consistent. But we see that the amount of information is dramatically reduced, which leads to higher variance.
With weak instruments, this area is even smaller and hence the variance higher.
Slide: Clear?
Think about it and we will come back to all of this several times in discussing the different papers.

Slide: Ugly instruments:

It is not really the instruments that are ugly, but rather the interpretation of the instruments. It is important to realize what the instrument is really measuring.

In particular, if the causal effect is not the same for all individuals and the instrument works differently for different groups, we must be aware of this.

Imagine that we have two types of individuals that are affected differently by the same intervention, A and B, with Coefficients Ba and Bb. OLS (i.e. “normal regression”) gives a weighted average of these two effects, which is often what we want. If we run an IV estimation and the instrument only applies to group B, then we only get the true effect for Bb. This is called the local average treatment effect, and it is not always the most relevant measure. In particular, this effect is not the same as the average treatment effect!

- What are we really measuring?
- If heterogeneity is present IV estimation may reveal results for a specific group which may differ from the average effect.
- LATE: Local Average Treatment Effect.

Slide: Example

- Effect of education.

Those affected by school laws have little education and therefore have a high marginal utility of an extra year. Thereby we are not measuring the returns to education in general. Not even the effect of education for those with low education. The effect is rather one for those who would not have studied the extra year absent the schooling law.
Another example is the credit expansion studied by Burgess and Pande. Their IV estimates gives us the effects of having more banks via the program and not of having more banks in general.

- So, we must know what we are measuring!

Diff in diff:

- Requires that data is available both before and after treatment.
- Basic idea: Control for pre-period differences in outcomes between T and C.
- Crucial assumption. Absent the treatment, the outcomes would have followed the same trend.
- Main practical issue: Omitted variable… you must argue your case strongly!

We compare the difference before (1^{st} diff) to the difference after (2^{nd} diff).

2 slides with figures: The counterfactual trend is taken from the control. In the second slide the control group changes, we could not possibly pick up the correct effect in that case.

Slide problems
The most compelling DD studies report outcomes for treatment and control observations for a period long enough to show the underlying trends, with attention to how the deviations from the trend relate to changes in policy.

Slide: Real world example: Homicide rates and the death penalty.
Donohue and Wolfers (2005) “Uses and abuses of empirical evidence in the death penalty debate”.

Slide Figure

This figure plots homicide rates in Canada and the US for over half a century, indicating periods when the death penalty was in effect in the two countries.
The point of the figure is not to focus on Canada's consistently lower homicide rate, but instead to point out that Canadian and US homicide rates move roughly in parallel, suggesting that changes in death penalty policy were of little consequence for murder.

1) If looking at only Canada it seems as if the abolishment of the death penalty had an effect.
2) Not if we look at both (more likely a regional trend).
3) Abolishing in the US nothing happens *to the trend* (if only looking year before and after we would conclude there was an effect) first but then there is a dip (but also in Canada!)
4) Re-establishing the death penalty in the US does not reduce homicides
5) The figure also suggests that the deterrent effect of capital punishment would have to be large to be visible against the background noise of yearly fluctuations in homicide rates.
6) In sum, convincingly showing that there is no effect. How? By using many periods and several reforms.

Slide Regression discontinuity:
It is not a new method but it is increasingly popular…

Basic idea: Exploit that the probability of treatment is a discontinuous function of at least one observable variable.
Clear right ☺
Examples may be that a poverty relief program is only given to those with less than 40 dollars per month.
We can't then compare those who got the program to those that do not, because one group is poor and the other group less poor. But we could compare those that have exactly 40 dollars and are precisely out of the program to those that have 39.99 dollars per month. These two groups should be so similar that it is random who receives treatment and who doesn't.
Other examples may be that you get into a good university if your exam score is at least 207, comparing those who got 207 points to those that got 206 points may give a causal interpretation.
So, the idea is to estimate the treatment effect using individuals just below the threshold as a control for those just above.

2 Slides with figures.

Slide: Another example:

- Pension program in rural Mexico:
- Rural: Only in places with less than 30 000 inh.
- Let p be the ”forcing/running variable”
- p= population – 30 000 so that:

$$Treatment_i = \begin{cases} 1 if \text{ population} < 30\,000 \rightarrow \text{p} < 0 \\ 0 if \text{ population} \geq 30\,000 \rightarrow \text{p} \geq 0 \end{cases}$$

Slide

So, how do we estimate this?

- Say we want to estimate the effects on poverty. See the 2 figures drawn during class.

1) Poverty=a+BT+e, gives the diff in means. In our case, the continuous effect of *population* is controlled for by estimating:
2) Put in a line and Poverty= a+BT+gammap+ e, Controls for the continues effect of population. So that we get the jump.
3) It is the same as estimating the two regression functions below and calculating the difference in intercepts (*a1-a2*):

   Poverty= a1+ gammap+ e, if p<0

   Poverty= a2+ gammap+ e, if p>=0

Slide You can also use RD in physical space as Michalopoulos and Papaioannou

They look at outcomes within ethnicities across national boundaries to assess the effects of national level institutions. We will look more at this in our last lecture!

Slide

- Very popular.
- Often a much closer cousin of randomization than the other methods.
- Also ethical advantage if distribution is based on needs.
- Crucial assumption: No manipulation or sorting around the threshold.

Slide

RD is underexploited:
Burgess and Pande:
“Banks were required to select unbanked locations for branch expansion from a list circulated by the Central Bank. This list identified all unbanked locations with a population above a certain number. As the same population cut-off was applied across India...The list was updated, with a lower population cutoff, every three years.”