t.a.stoerk@lse.ac.uk
May 2019

Methods of empirical ex post policy evaluation

This note gives a non-technical overview over methods of empirical policy evaluation.

0. Causal policy evaluation

The main purpose of empirical ex post policy evaluation is to answer: what was the *causal* effect of a policy? Or, to use terminology from medical research, to determine the effect of a *treatment*. For a given treatment, an individual can experience two potential outcomes (y_i): the outcome without treatment ($D_i = 0$) and the outcome with treatment ($D_i = 1$):

$$y_i = \begin{cases} y_{1i} \;\; if \; D_i = 1 \\ y_{0i} \;\; if \; D_i = 0 \end{cases}$$

The *treatment effect* is the difference between these two potential outcomes:

$$treatment \; effect = \; y_{1i} - \; y_{0i}$$

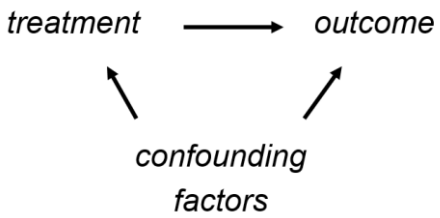
To calculate the treatment effect, we therefore need to know both outcomes. The outcome with treatment (y_{1i}) is observable from data; it is possible to observe the patient after treatment.

However, the outcome without treatment (y_{0i}), the *counterfactual*, is not observable. We can not know what would have happened to an individual had they not been given treatment. The same is true when we talk about policies instead of medical treatments: a firm is either regulated through the EU ETS, or it is not. The treatment effect for an individual can therefore never be estimated.

However, it is possible to find the *average treatment effect (ATE)* in a population. Social scientists generally believe that individuals or firms are similar, so that comparable untreated individuals or firms can be used to form a valid counterfactual.

1. The logic of randomized experiments

How can a comparable group of untreated individuals be *identified*? The fundamental problem of estimating ATEs is to isolate confounding factors with would result in *bias* – an under- or overestimate of the true treatment effect. For example, emissions of EU ETS-regulated firms declined in 2009, but was that due to the EU ETS, or the economic recession, or perhaps another factor?



The traditional solution is to use *randomized controlled experiments* in which a targeted population gets assigned to treatment and *control group* by randomization. In other words, both the treatment and the control group are identical on average before the treatment. Any statistically significant differences in outcomes after the treatment can therefore be due to the effect of the treatment, the ATE.

The ATE can be estimated as:

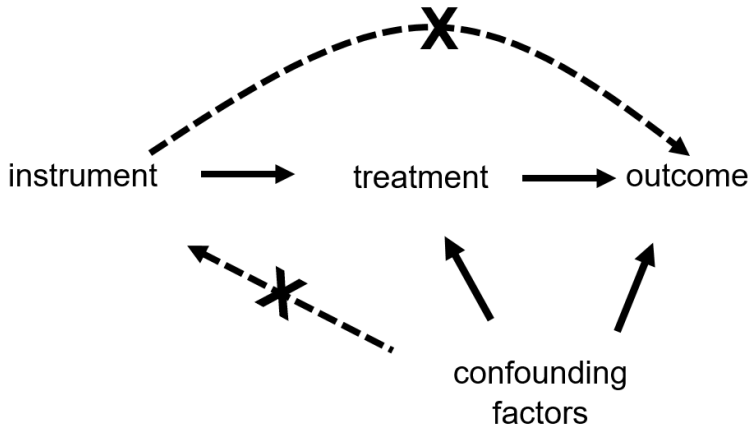
$$\widehat{ATE} = \; \bar{y}_1 - \; \bar{y}_0 = \; \frac{1}{N_1} \sum_{i|D_i=1} y_i \; - \; \frac{1}{N_2} \sum_{i|D_i=2} y_i$$

In other words, the ATE is the difference between the mean outcome of both groups. Note that both outcomes are observable.

Most policies, however, are not randomly assigned. The remainder of this note shows quasi-experimental techniques to estimate the counterfactual in such cases. All methods use the tools of microeconometrics, but the technical specifics are not key. Instead, what matters is a researcher's way of identifying a suitable counterfactual, or their *identification strategy*.

2. Instrumental variables (IV)

An *instrumental variable (IV)* is a variable that is independent of the confounding factors, but correlated with the treatment. In other words, it only affects the outcome through the treatment. Apart from correcting for confounding factors, instruments can be used to overcome measurement error and reverse causality (i.e., the outcome itself affects the treatment).



Whether an IV is a good identification strategy depends on two conditions, *exogeneity* and *relevance*. Exogeneity means that the IV is uncorrelated with the confounding factors, and has no direct effects on the outcome. Exogeneity can not be tested for but instead depends on a qualitative argument as to whether the setting is convincing in practice. Relevance is a measure for the strength of the correlation between the instrumental and the treatment, and it can generally be tested.

There are two caveats regarding IV estimates. Firstly, IV generally does not identify the ATE for the whole treated population. Instead, IV estimates a *local average treatment effect (LATE)*, i.e. the effect of the treatment on a subsample of those treated. Which subsample this is in practice, and how informative about the ATE it is, depends on the setting. Secondly, IV estimates should be interpreted with caution even if the identification strategy makes intuitive sense: technical considerations can make IV estimates biased under some circumstances (Young, 2018).

Example: What is the effect of state aid for job creation in the EU? (Criscuolo, Martin, Overman and Van Reenen, Forthcoming)

The paper evaluates a British policy to create jobs through public investment funding. EU state aid rules govern the eligibility of a given area to receive public funding. The authors use a change in the eligibility criteria for a region to receive the public funds to construct an instrumental variable.

This identification is credible because it fulfils the exogeneity assumption: all effects on job creation can be expected to go through the increased public funds as a result of the policy. The eligibility criteria, by contrast, do not influence jobs themselves. The authors test instrument relevance by showing the F statistic of the coefficient on the instrument in the *first stage regression*, a regression of treatment on instrument that measures the strength of the association between the instrument and the

treatment (marked in red boxes below). As a rule of thumb, an F statistic beyond 10 is a necessary condition for instrument relevance.

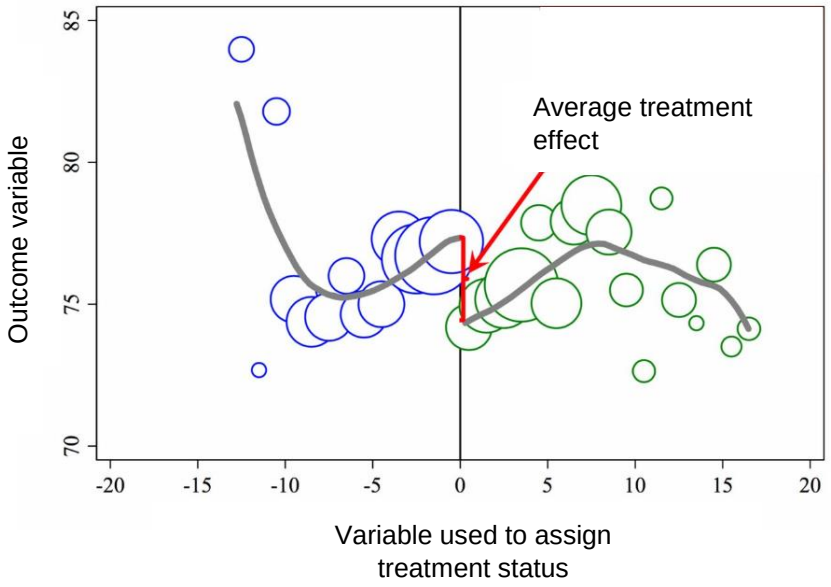
Table 4: ln(Employment) Regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	**Ln(EMP)**	**ln(EMP)**	**RSA**	**ln(EMP)**	**ln(EMP)**	**ln(EMP)**	**RSA**	**ln(EMP)**
	OLS	Reduced Form	First Stage	IV	FE	Reduced Form	First Stage	IV
RSA	0.370***			4.658***	0.166***			0.646***
	(0.020)			(0.530)	(0.013)			(0.154)
NGE = 10%		0.197	0.063			-0.062*	0.015	
		(0.127)	(0.041)			(0.032)	(0.020)	
NGE = 15%		0.368***	0.122***			-0.005	0.068***	
		(0.049)	(0.023)			(0.019)	(0.014)	
NGE = 20%		0.382***	0.050***			0.027***	0.015***	
		(0.017)	(0.008)			(0.003)	(0.004)	
NGE = 30%		0.406***	0.091***			0.033***	0.027***	
		(0.031)	(0.009)			(0.006)	(0.005)	
NGE = 35%		0.236***	0.197***			0.033	0.098***	
		(0.043)	(0.024)			(0.023)	(0.014)	
Observations					157771			
Number of firms					28882			
F-stats for excluded instruments			29.52				16.90	
Fixed effects	NO	NO	NO	NO	YES	YES	YES	YES

Source: Criscuolo et al. (Forthcoming)

3. Regression discontinuity design (RDD)

Regulation often uses arbitrary size thresholds. A *regression discontinuity design (RDD)* compares regulated individuals just exceeding the threshold to individuals just below it. Identification in an RDD comes from the idea that who ends up on one side of the threshold versus the other is as good as random for observations close to the threshold. For example, installations are only regulated in EU ETS if they are larger than a certain capacity threshold. Installations just below the threshold would therefore be an instructive control group. The differences in the outcome just at the threshold will therefore show the average treatment effect at the threshold.



Source: Adapted from Ebenstein, Fan, Greenstone, He and Zhou (2017)

Put differently, the identifying assumption is that all factors except for treatment status vary continuously across the threshold. This assumption can be tested for observable individual characteristics. However, such a test is only a necessary condition for identification; unobservable confounding factors could still exist.

Identification through an RDD only works if regulated individuals cannot manipulate the assignment. In the case of the EU ETS, for instance, an RDD would require that firms cannot change the capacity of their installation in response to the announcement of the EU ETS. A statistical test, the McCrary test, can be used to test for suspicious patterns in the variable used to assign individuals to treatment status. Such a test is, however, only an indication. In some situations it is also not necessary, as manipulation is ruled out by the setting itself.

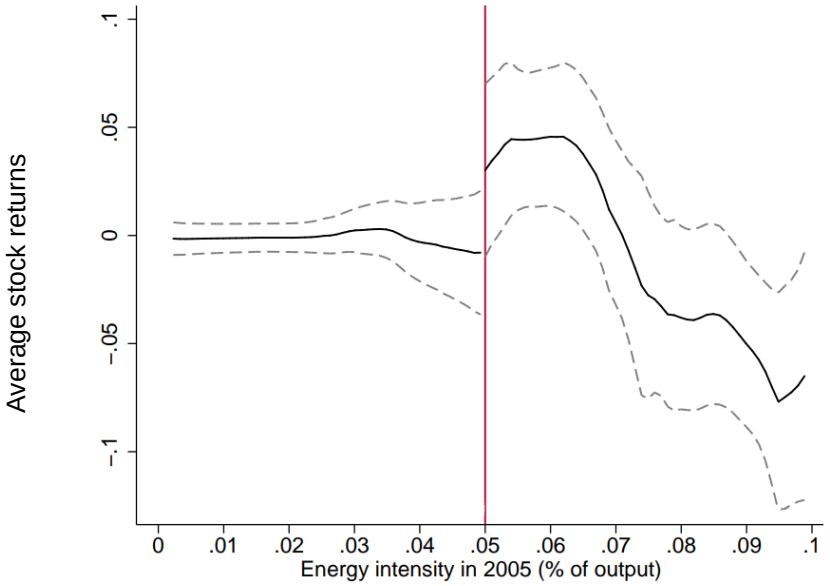
RDD is a transparent identification strategy, though it gives the researchers two parameters that can greatly affect the results. *Bandwidth* decides how far away from a threshold to look. Generally speaking, identification is best when only data points very close to the threshold are used. On the other hand, statistical precision increases the larger the sample. For this reason, it is generally considered good practice to show the ATE for different bandwidths. Similarly, the estimated *ATE* depends on how the line is fitted across the data points on either side of the threshold. It is again good practice to show the ATE for different ways of doing so.

RDD estimates come with one caveat: they only identify a LATE for individuals at the threshold. In the EU ETS example, an RDD would tell us the LATE for small installations. Whether this LATE is informative of the ATE for larger installations is not clear.

Example: What is the real word marginal abatement cost of climate policy? (Meng, 2017)

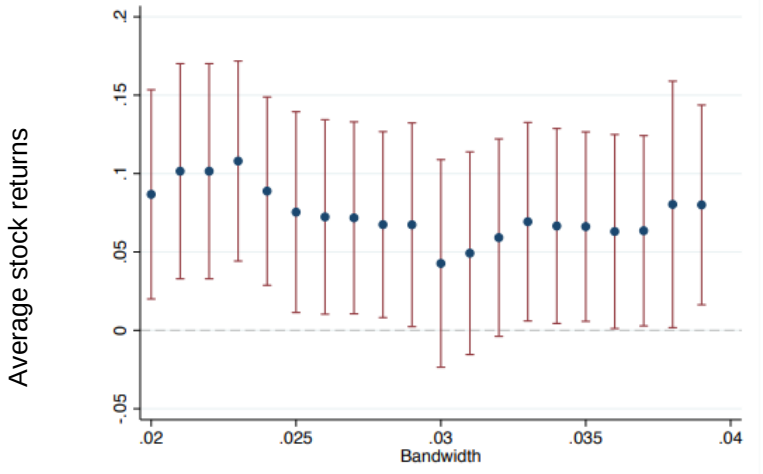
Meng (2017) evaluates what would have been the marginal abatement cost for firms had the Waxman-Markey Bill established a cap-and-trade system for greenhouse gas emissions in the US in 2009. The proposed bill would have granted regulated firms free allowances if their energy intensity exceeded 5 percent. This threshold makes it possible to study the effect of obtaining free allowances on stock returns through an RDD.

The graph below shows a clear discontinuity in average stock returns at the threshold, and this discontinuity can be causally attributed to the free permit rule. Meng (2017) then translates these changes in stock returns into marginal abatement costs using a rich arsenal of methods, but we focus on just the RDD aspect here.



Source: Adapted from Meng (2013), itself an earlier version of Meng (2017).

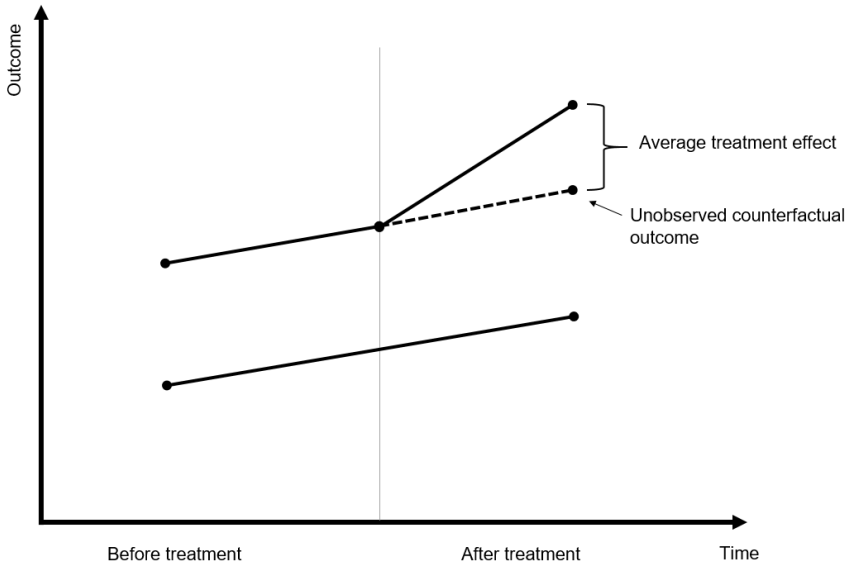
The effects are likely causal as sorting was unlikely: the free permit rule was unexpected, and the energy intensity was calculated retrospectively. Meng (2017) furthermore includes a robustness check to show that his ATE estimate is not driven by his choice of bandwidth. The graph below shows that ATE estimates differ little independent of bandwidth.



Source: Meng (2017)

4. Differences-in-differences (DID)

A *differences-in-differences (DID)* strategy is used when there is no obvious control group that is similar in every way to the treatment group except for the receipt of treatment. Instead, DID assumes that the difference between two similar groups is constant over time. Put differently, as shown below, DID assumes that the outcome for the treatment and the control group would have followed the same time trend in the absence of treatment.

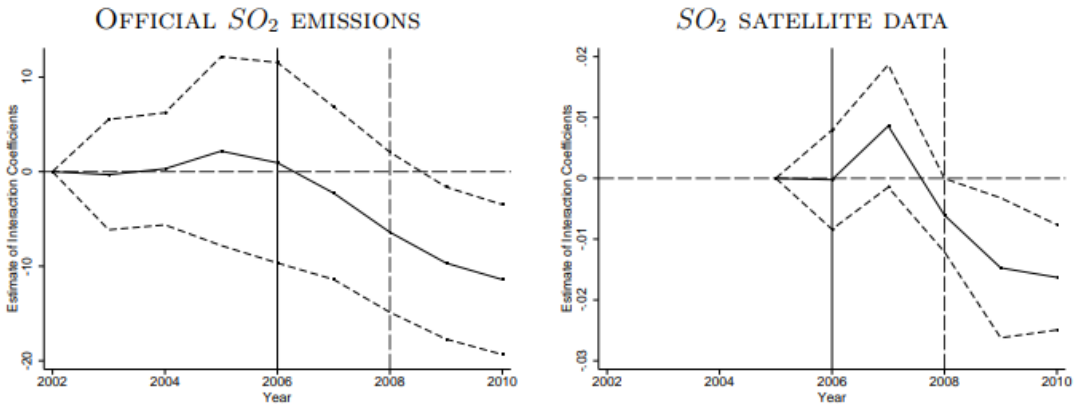


This identification strategy is reasonable when the likely confounding factors affect both groups alike and the treatment is unrelated to those factors. The common evolution of the difference across treatment and control group prior to the start of the treatment can be tested based on data. However, such a test is only supportive and never conclusive and may not be informative about trends after the treatment. For instance, identification still fails if another policy changes at the same time as the one that is studied.

Example: How effective is air pollution control in China? (Stoerk, 2018)

In my own research, I evaluate whether more stringent air pollution targets actually decreased air pollution in China. To do so, I use a DID identification strategy that compares provinces with a high reduction target to provinces with a low reduction target using two different data sources for SO_2 emissions.

The graphs below illustrates how DID works in practice. The solid black line in each graphs shows the estimated average treatment effect per year, compared to the first year for which I have data.



Source: Stoerk (2018)

There are two take-aways from this graph:

1. The data confirm the assumption of common pre-trends: the estimated ATE is 0 for all periods up to 2006, which is when the policy was announced.
2. The policy had an effect in reducing air pollution eventually: the ATE is statistically different from zero in 2009 and 2010 (indicated by the dashed lines not including zero).

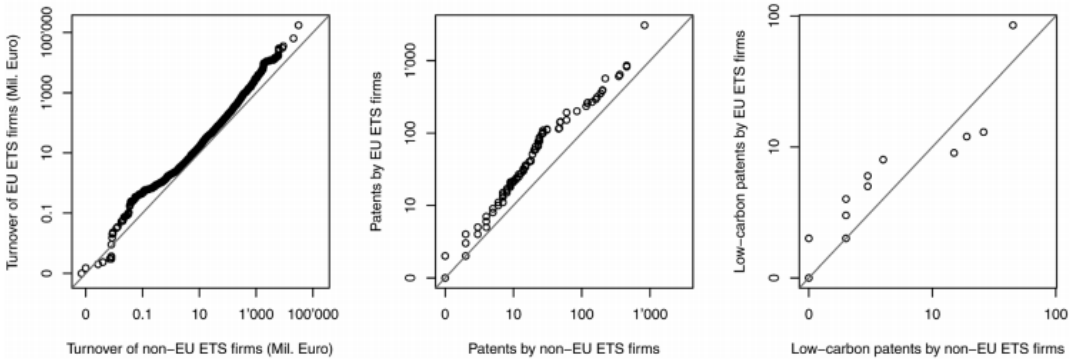
5. Matching with differences-in-differences

Matching is used when there is no clear control group for a DID. Typically, this is because the potential control group is more heterogeneous than the treatment group. Matching constructs a control group composed of individuals that look similar to those in the treatment group before the start of the treatment. The more data on individual characteristics, the better the match.

Note, however, that even if both groups appeared perfectly similar in observable characteristics, a correct estimate of the ATE is not guaranteed. Unobservable confounding factors could still bias the estimated ATE. Matching is for this reason typically combined with a DID approach to account for fixed differences across individuals.

<u>Example: Did the EU ETS cause low-carbon innovation? (Calel and Dechezleprêtre, 2016)</u>

Did the EU ETS cause low-carbon innovation by regulated firms? To answer this question, Calel and Dechezleprêtre (2016) use firm-level data to match EU ETS-regulated firms to firms that are observationally similar at the start of the EU ETS. They use data on firm-level revenue and patenting to do so. The graph below shows that the matches worked reasonably well: the observable characteristics of EU ETS firms (on the horizontal axes) are similar to those of unregulated firms (vertical axes).



Source: Calel and Dechezleprêtre (2016)

To convince us that the matched control group is a valid counterfactual, they furthermore show a *placebo test* in which they show that firms in both the treatment and the control group are similar even for a variable that was not used in the matching process. This test is used to give the reader confidence that other unobservable characteristics are equally similar. However, this is just an indication rather than proof.

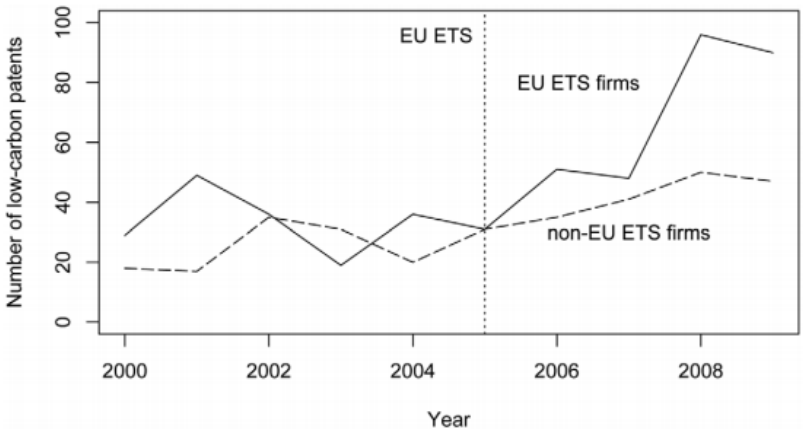
Source: Calel and Dechezleprêtre (2016)

After they have shown us that their identification strategy is convincing, the authors conduct a DID estimation to find the average treatment effect of the EU ETS on low-carbon innovation.

The graph below again offers two take-aways:

1. The data confirm the assumption of common pre-trends before the start of the policy.
2. The effect of the EU ETS on low-carbon innovation is visible in the divergence of the number of patents of EU ETS firms compared to non-EU ETS firms.

FIGURE 5.—LOW-CARBON PATENTS BY MATCHED EU ETS AND NON–EU ETS FIRMS



Source: Calel and Dechezleprêtre (2016)

Bibliography

Calel, Raphael and Antoine Dechezleprêtre (2016): "Environmental Policy and Directed Technical Change: Evidence from the European carbon market", *Review of Economics and Statistics*, 98(1): 173-191.
https://drive.google.com/open?id=1hyNxNP3upgID1Fd_VO5nac08Oug5OiMZ

Criscuolo, Chiara, Ralf Martin, Henry G. Overman, and John Van Reenen (Forthcoming): "Some causal effects of an industrial policy", *American Economic Review*.

Ebenstein, Avraham, Maoyong Fan, Michael Greenstone, Guojon He, and Maigeng Zhou (2017): "New evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River Policy", *Proceedings of the National Academic of Sciences*, 110(32): 12936-12941.
http://www.pnas.org/content/early/2017/09/05/1616784114.full

Meng, Kyle (2013): *Essays in the Economics and Political Economy of Climate Change*, Dissertation at Columbia University.

Meng, Kyle (2017): "Using a Free Permit Rule to Forecast the Marginal Abatement Cost of Proposed Climate Policy", *American Economic Review*, 107(3): 748-784.
https://www.dropbox.com/s/506wa9t2kvxop5y/Meng_aer_2017_manuscript.pdf?dl=1

Stoerk, Thomas (2018): "Effectiveness and cost of air pollution control in China", LSE Grantham Research Institute on Climate Change and the Environment Working Paper No. 273.
http://84.89.132.1/~tstoerk/Stoerk_2018_Effectiveness%20and%20Cost%20of%20Air%20Pollution%20Control%20in%20China.pdf

Young, Alwyn (2018): "Consistency without inference: Instrumental Variables in Practical Application", mimeo. https://personal.lse.ac.uk/YoungA/ConsistencyWithoutInference.pdf

Materials consulted in the writing of this note

Slides by Steve Pischke (LSE)
http://econ.lse.ac.uk/staff/spischke/ec533/

Slides by Kurt Schmidheiny (U Basel)
https://www.schmidheiny.name/teaching/upf/econometrics/

Slides by Fabian Waldinger (LSE)
https://www.fabianwaldinger.com/applied-econometrics

Further reading

Nontechnical
Angrist, Joshua D. and Jörn-Steffen Pischke (2015): *Mastering 'Metrics: The Path from Cause to Effect*, Princeton University Press.

Technical
Angrist, Joshua D. and Jörn-Steffen Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.