Causal inference with observational data

A brief review of quasi-experimental methods

Austin Nichols

July 30, 2009

Why should you care?

Virtually every set of estimates invites some kind of causal inference.

Most data is observational and estimates are biased. May even have the wrong sign!

Selection and Endogeneity

In a model like $y = Xb + e$, we must have $E(X'e) = 0$ (**exogeneity**) for unbiased estimates of b. Without random assignment of X, we have observational data, and biased estimates are the norm. The assumption of $E(X'e) = 0$ fails in the presence of measurement error in X, simultaneous equations or reverse causality, omitted variables in X, or selection (of X) based on unobserved or unobservable factors. The **selection problem** is my focus, though it can also be framed as an omitted variables problem. The general term for $E(X'e) \neq 0$ is **endogeneity** of the error e.

A classic example is the effect of education on earnings, where the highest ability individuals may get more education, but would have had higher earnings regardless, leading us under this simple assumption to guess that the effect of education is overestimated by a comparison of mean income conditional on education.

Following standard practice, I will refer to the columns of X whose effect we are trying to measure as the **treatment** variables.

Solutions

There are three kinds of solutions:

1. control for all important observables directly (may require you to observe unobserved factors),
2. run an experiment (may not be possible, or may be prohibitively expensive),
3. use a quasi-experimental (QE) method.

Also used to address other causes of endogeneity; see e.g. Hardin, Schmiediche, and Carroll (2003) on measurement error.

I will discuss four classes of these methods:

1. Matching or reweighting,
2. Panel methods,
3. Instrumental variables (IV), and
4. Regression discontinuity (RD).

and some hybrids. Angrist and Pischke (2009) provide a good overview of a few approaches, and Imbens and Wooldridge (2007) cover most.

A Simple Example

Treatment	Success 0	1	Total
P	.1743 [.1621,.1872]	.8257 [.8128,.8379]	1
O	.22 [.2066,.234]	.78 [.766,.7934]	1
Total	.1971 [.188,.2066]	.8029 [.7934,.812]	1

Key: row proportions
[95% confidence intervals for row proportions]

A Simple Example, cont.

-------Large Stones----------------------------------

Treatment	Success 0	Success 1	Total
P	.3125	.6875	1
	[.2813,.3455]	[.6545,.7187]	
O	.27	.73	1
	[.2533,.2873]	[.7127,.7467]	
Total	.2799	.7201	1
	[.2651,.2952]	[.7048,.7349]	

---

v.

-------Small Stones----------------------------------

Treatment	Success 0	Success 1	Total
P	.1333	.8667	1
	[.121,.1467]	[.8533,.879]	
O	.069	.931	1
	[.0539,.0878]	[.9122,.9461]	
Total	.1176	.8824	1
	[.1075,.1286]	[.8714,.8925]	

The Rubin Causal Model

Rubin (1974) gave us the model of identification of causal effects that most econometricians carry around in their heads, which relies on the notion of a hypothetical **counterfactual** for each observation. The model flows from work by Neyman (1923,1935) and Fisher (1915,1925), and perhaps the clearest exposition is by Holland (1986); see also Tukey (1954), Wold (1956), Cochran (1965), Pearl (2000), and Rosenbaum (2002).

To estimate the effect of a college degree on earnings, we'd like to observe the earnings of college graduates had they not gone to college, to compute the gain in earnings, and to observe the earnings of nongraduates had they gone to college, to compute their potential gain in earnings.

The Fundamental Problem

The Fundamental Problem is that we can never see the counterfactual outcome, but randomization of treatment lets us estimate treatment effects. To make matters concrete, imagine the treatment effect is the same for everyone but there is heterogeneity in levels—suppose there are two types 1 and 2:

$Type$	$E[y\|T]$	$E[y\|C]$	TE
1	100	50	50
2	70	20	50

and the problem is that the treatment T is not applied with equal probability to each type. For simplicity, suppose only type 1 gets treatment T and put a missing dot in where we cannot compute a sample mean:

$Type$	$E[y\|T]$	$E[y\|C]$	TE
1	100	.	?
2	.	20	?

The difference in sample means overestimates the ATE (80 instead of 50); if only type 2 gets treatment the difference in sample means underestimates the ATE (20 instead of 50).

The Solution

Random assignment puts equal weight on each of the possible observed outcomes:

Type	$E[y\|T]$	$E[y\|C]$	*TE*
1	100	.	?
2	.	20	?
1	.	50	?
2	70	.	?

and the difference in sample means is an unbiased estimate of the ATE.

For all of this, we are assuming treatment only affects outcomes for the unit treated (the Stable Unit Treatment Value Assumption, or SUTVA), so the number of people treated has no impact on the efficacy of any one treatment. In practice, this assumption is usually violated—there are spillover effects, so it is useful to bear in mind what they might be and how it affects the interpretation of estimates.

The Gold Standard

To control for unobservable factors, the gold standard is a randomized controlled trial, where individuals are assigned X randomly. In the simplest case of binary X, where $X = 1$ is the treatment group and $X = 0$ the control, the effect of X is a simple difference in means, and all unobserved and unobservable selection problems are avoided. In fact, we can always do better (Fisher 1926) by conditioning on observables, or running a regression on more than just a treatment dummy, as the multiple comparisons improve efficiency.

In many cases, an RCT is infeasible due to cost or legal/moral objections. Apparently, you can't randomly assign people to smoke cigarettes or not. You also can't randomly assign different types of parents or a new marital status, either. Still, it is useful to imagine a hypothetical experiment, which can guide our estimation strategy.

All That Glitters

Even where an experiment is feasible, the implementation can be quite daunting. Often, the individuals who are randomly assigned will agitate to be in another group—the controls want to get treatment if they perceive a benefit, or the treatment group wants to drop out if the treatment feels onerous—or behave differently.

Even in a double-blind RCT, there may be leakage between treatment and control groups, or differing behavioral responses. Those getting a placebo may self-medicate in ways the treatment group do not (imagine a double-blind RCT for treatment of heroin addiction), or side effects of treatment may induce the treatment group to take some set of actions different from the control group (if your pills made you too sick to work, you might either stop taking the pills or stop working—presumably the placebo induces fewer people to give up work).

QE Methods in Experiments

In practice, all of the quasi-experimental methods here are used in experimental settings as well as in observational studies, to attempt to control for departures from the ideal of the RCT.

Sometimes, the folks designing experiments are clever and build in comparisons of the RCT approach and observational approaches. See Orr et al. (1996) for one example where OLS appears to outperform the more sophisticated alternatives, and Heckman, Ichimura, and Todd (1997) where more sophisticated alternatives are preferred. Smith and Todd (2001,2005) pursue these comparisons further.

Another major problem with experiments is that they tend to use small and select populations, so that an unbiased estimate of a treatment effect is available only for a subpopulation, and the estimate may have large variance. This is mostly a question of scale, but highlights the cost, bias, and efficiency tradeoffs in choosing between an experiment and an observational study.

The Counterfactual Again

The mention of the placebo group self-medicating may also bring to mind what happens in social experiments. If some folks are assigned to the control group, does that mean they get no treatment? Generally not. A person who is assigned to get no job training as part of an experiment may get some elsewhere. Someone assigned to get job training as part of an experiment may sleep through it.

The treatment group may not get treated; the control group may not go untreated. The important thing to bear in mind is **the relevant counterfactual**: what two regimes are you comparing? A world in which everyone who gets treated gets the maximum intensity treatment perfectly applied, and those who don't get treated sit in an empty room and do nothing? What is the status quo for those not treated?

ATE

The assumption so far has been that the treatment effect is the same for everyone. If individuals may have different treatment effects, or marginal effects, of some endogenous X, a regression of Y on X will not in general recover the mean marginal effect of X, or average treatment effect (ATE).

(IV, which comes later, can get consistent estimates in some cases of heterogeneous effects, but not all; see e.g. Wooldridge 1997 and Heckman and Vytlacil 1997.)

ATE and LATE

For evaluating the effect of a treatment/intervention/program, we may want to estimate the ATE for participants (the average treatment effect on the treated, or ATT) or for potential participants who are currently not treated (the average treatment effect on controls, or ATC), or the ATE across the whole population (or even for just the sample under study).

Often, however, for interventions which we are thinking about expanding, we want only the ATE for the marginal participants, i.e. those to whom treatment will be extended. This quantity, one version of the Local Average Treatment Effect (LATE) where local means "local to marginal participants at the current size," is often exactly what is estimated by quasi-experimental methods, particularly IV and RD. See the classic, short, and well-written papers Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996), and see Heckman and Vytlacil (1999, 2000, 2004) for further discussion.

Outline

Outline

Matching and Reweighting Distributions

If individuals in the treatment and control groups differ in observable ways (selection on observables case), a variety of estimators are possible. One may be able to include indicators and interactions for the factors that affect selection, to estimate the impact of some treatment variable within groups of identical X (a fully saturated regression). There are also matching estimators (Cochran and Rubin 1973) which compare observations with like X, for example by pairing observations that are "close" by some metric. A set of alternative approaches involve reweighting so the distribution of X is identical for different groups, discussed in Nichols (2008).

Nearest Neighbor Matching

Nearest neighbor matching pairs observations in the treatment and control groups and computes the difference in outcome Y for each pair, then the mean difference across pairs. Imbens (2006) presented at last year's meetings on the Stata implementation `nnmatch` (Abadie et al. 2004). See Imbens (2004) for details of Nearest Neighbor Matching methods.

The curse of dimensionality, and other problems

The downside to Nearest Neighbor Matching is that it can be computationally intensive, and bootstrapped standard errors are infeasible owning to the discontinuous nature of matching (Abadie and Imbens, 2006).

Propensity score matching

Propensity score matching essentially estimates each individual's propensity to receive a binary treatment (via a `probit` or `logit`) as a function of observables and matches individuals with similar propensities. As Rosenbaum and Rubin (1983) showed, if the propensity were known for each case, it would incorporate all the information about selection and propensity score matching could achieve optimal efficiency and consistency; in practice, the propensity must be estimated and selection is not only on observables, so the estimator will be both biased and inefficient.

Morgan and Harding (2006) provide an excellent overview of practical and theoretical issues in matching, and comparisons of nearest neighbor matching and propensity score matching. Their expositions of different types of propensity score matching, and simulations showing when it performs badly, are particularly helpful.

Propensity score matching methods

Typically, one treatment case is matched to several control cases, but one-to-one matching is also common. One Stata implementation `psmatch2` is available from SSC (`ssc desc psmatch2`) and has a useful help file, and there is another Stata implementation described by Becker and Ichino (2002) (`findit pscore` in Stata). `psmatch2` will perform one-to-one (nearest neighbour or within caliper, with or without replacement), k-nearest neighbors, radius, kernel, local linear regression, and Mahalanobis matching.

As Morgan and Harding (2006) point out, all the matching estimators can be thought of as reweighting scheme whereby treatment and control observations are reweighted to allow causal inference on the difference in means. Note that a treatment case i matched to k cases in an interval, or k nearest neighbors, contributes $y_i - k^{-1} \sum_1^k y_j$ to the estimate of a treatment effect, and one could just as easily rewrite the estimate of a treatment effect as a weighted mean difference.

Common support

Propensity score methods typically assume a common support, i.e. the range of propensities to be treated is the same for treated and control cases, even if the density functions have quite different shapes. That way, there are close matches for all observations as the support is filled in (asymptotically)—in practice, of course, many closest matches may not be all that close. It is also rarely the case in practice that the ranges of estimated propensity scores are the same, but they do nearly always overlap, and generalizations about treatment effects are often limited to the smallest connected area of common support.

Often a density estimate below some threshold greater than zero defines the end of common support—see Heckman, Ichimura, and Todd (1997) for more discussion. This is because the common support is the range where both densities are nonzero, but the estimated propensity scores take on a finite number of values, so the empirical densities will be zero almost everywhere—we need a kernel density estimate in general, to obtain smooth estimated density functions, but then areas of zero density may have positive density estimates, so some small value is redefined to be effectively zero.

Limiting to common support

It is unappealing to limit the sample to a range of estimated propensity scores, since it is hard to characterize the population to which an estimate would generalize in that case. A more appealing choice if the distributions of propensity scores exhibit poor overlap, or if kernel density estimates of propensity scores for treatment or control groups exhibit positive density or nonzero slope at zero or one, is to limit to ranges of X variables, such that the distributions of propensity scores exhibit better properties. At least in this case, we can say "our estimates apply to unemployed native workers with less than a college education" or somesuch, together with an acknowledgement that we would like estimates for the population as well, but the method employed did not allow it.

Common support diagnostics

Regardless of whether the estimation or extrapolation of estimates is limited to a range of propensities or ranges of X variables, the analyst should present evidence on how the treatment and control groups differ, and which subpopulation is being studied. The standard graph here is an overlay of kernel density estimates of propensity scores for treatment and control groups, easy in Stata with `twoway kdensity`, but better with `kdens` (Jann 2007).

The assumption that p is bounded away from zero and one is important. In practice, the kernel density graph of propensity scores gives information about violations of the assumption that p is bounded (strictly) away from zero and one. Not only should the density be zero at the boundaries zero and one, but the slope of the density should be zero there. Unfortunately, kernel density estimators do not work very well at boundaries; but see **kdens** (Jann 2007) offering boundary corrections at zero and one.

Propensity score reweighting

The propensity score can also be used to reweight the treatment and control groups so the distribution of X looks the same in both groups: one method is to give treatment cases weight one and control cases weight $p/(1-p)$ where p is the probability of treatment. Additional choices are discussed in Nichols (2008).

Note how important is the assumption that p is bounded away from zero and one here. If estimated p is very close to one for a control case, the reweighting scheme above assigns infinite weight to that one control case as the counterfactual for every treatment case, and this control case should not even exist (as p approaches one for a control case, the probability of observing such a case approaches zero)!

Propensity scores, true and estimated

Part of the problem may be that propensity scores are estimated. If we had true propensity scores, they would certainly never be one for a control case. But it turns out that is really not the problem, at least for mean squared error in estimates of causal impacts. In fact, you can usually do better using an estimated propensity score, even with specification error in the propensity score model, than using the true propensity score (based on unpublished simulations). This arises because the variance of estimates using true propensity scores is very high, whereas using an estimated propensity score is effectively a shrinkage estimator, which greatly reduces mean squared error.

In fact, Hirano, Imbens, and Ridder (2003) show that using nonparametric estimates of the propensity score to construct weights is efficient relative to using true propensity scores or covariates, and achieves the theoretical bound on efficiency (but see Song 2009 for a case where this does not hold).

It is a problem that propensity scores are estimated, because that fact is not used in constructing standard errors, so most SEs are too small in some sense. Yet if we think that using estimated propensity scores and throwing away information on true propensity scores can improve efficiency, perhaps our standard errors are actually too large! This is an active research area, but most people will construct standard errors assuming no error in estimated propensity scores.

More reweighting

The large set of reweighting techniques lead to a whole class of estimators based on reweighting the treatment and control groups to have similar distributions of X in a regression. The reweighting techniques include DiNardo, Fortin, and Lemieux (1996), Autor, Katz, and Kearney (2005), Liebbrandt, Levinsohn, and McCrary (2005), and Machado and Mata(2005), and are related to decomposition techniques in Blinder (1973), Oaxaca (1973), Yun (2004, 2005ab), Gomulka and Stern (1990), and Juhn, Murphy, and Pierce (1991, 1993).

DiNardo (2002) draws some very useful connections between the decomposition and reweighting techniques, and propensity score methods, but a comprehensive review is needed.

Selection on unobservables etc.

Imagine the outcome is wage and the treatment variable is union membership—one can imagine reweighting union members to have equivalent education, age, race/ethnicity, and other job and demographic characteristics as nonunion workers. One could compare otherwise identical persons within occupation and industry cells using `nnmatch` with exact matching on some characteristics. The various propensity score methods offer various middle roads.

However, these estimates based on reweighting or matching are unlikely to convince someone unconvinced by OLS results. Selection on observables is not the type of selection most critics have in mind, and there are a variety of remaining problems unaddressed by reweighting or matching, such as selection into a pool eligible for assignment to treatment or control—e.g. in the union case, there may be differential labor market participation (so whether or not a particular person would be in a union is unknown for many cases). One hypothesized effect of unions is a reduction in the size of workforces—if unionized jobs produce different proportions working, the marginal worker is from a different part of the distribution in the two populations. DiNardo and Lee (2002) offer a much more convincing set of causal estimates using an RD design.

Outline

Outline

DD

The simplest panel method is identical to a design used in many RCT's, the difference in differences (DD) method.

	Pre	Post
Treatment	y_1	y_2
Control	y_3	y_4

ATE
$(y_2 - y_1) - (y_4 - y_3)$

The average treatment effect (ATE) estimate is the difference in differences. For example, the estimate might be the test score gain from 8th grade to 12th grade for those attending charter schools less the test score gain for those in regular public schools. This assumes that the kids in charters, who had to apply to get in, would not have had the same gains in a regular school, i.e. that there was no selection into treatment.

DD, DDD, D^n

Having differenced out the "time" effect or the "state" effect, it is natural to want to add dimensions and compute a difference in differences in differences, and so on. This is equivalent to adding indicator variables and interactions to a regression, and the usual concerns apply to the added variables.

Natural Experiments

The usual "good" diff-in-diff approach relies on a natural experiment, i.e. there was some change in policy or the environment expected to affect treatment for one group more than another, and the two groups should not otherwise have different experiences. For this to work well, the natural experiment should be exogenous itself (i.e. it should not be the case that the policy change is a reaction to behavior) and unlikely to induce people to "game the system" and change their behavior in unpredictable ways (e.g. the differentially treated group jealously overcompensates).

For example, in some US states in 1996, immigrants became ineligible for food stamps, but 17 states offered a substitute program for those in the country before 1996. As of July 2002, anyone in the country five years was eligible for food stamps and most of those in the country 4.9 years were not. One could compute a difference in mean outcomes (say, prevalence of obesity) across recent and less recent immigrants, across calendar years 1995 and 1996, across affected and unaffected states. Using 2002, you could compute a difference across the population of immigrants in the country 4 years or 5 years. See Kaushal (2007) for a related approach.

Good Natural Experiments

In most cases, these types of natural experiments call for one of the other methods below (the food stamp example cries out for an Regression Discontinuity approach using individual data). A hybrid of DD and another approach is often best.

In general, the more bizarre and byzantine the rules changes, and the more draconian the change, the more likely a natural experiment is likely to identify some effect of interest. A modest change in marginal tax rates may not provide sufficient power to identify any interesting behavioral parameters, but the top marginal estate tax rates falling from 45% in 2009 to zero in 2010 and then jumping to 55% in 2011 creates an interesting incentive for mercenary children to pull the plug on rich parents in the tax-free year.

Difference and Fixed Effects Models

The natural generalization of the diff-in-diff method is to compute a difference for each individual (person, firm, school, etc.), as in a first-difference model, or include an individual-specific intercept for the fixed effect (FE) model. This can be extended to two-way and n-way fixed effects just as the diff-in-diff can be extended to the diff-in-diff-in-diff etc.

Suppose ability A is fixed for each individual i and does not change as time t passes. A increases earnings Y and is correlated with higher schooling X, but we cannot observe A in the true model:

$$Y_{it} = X_{it}b + A_i + e_{it}$$

so we estimate a first-difference model to eliminate the unobservable A:

$$Y_{it} - Y_{i(t-1)} = (X_{it} - X_{i(t-1)})b + e_{it} - e_{i(t-1)}$$

Fixed Effects Models

To include individual-specific intercepts, we can demean the data:

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)b + v_{it}$$

or simply include an indicator variable for each individual i:

$$Y_{it} = (X_{it})b + a_i + v_{it}$$

In fact, assuming we have a individual ID variable `indiv`, we should use one of:

```
xtreg y x*, fe i(indiv) cluster(indiv)
areg y x*, abs(indiv) cluster(indiv)
```

instead of including indicators. The cluster option allows for errors to be serially correlated within panel (Arellano 1987; Kézdi 2004; Stock and Watson 2006).

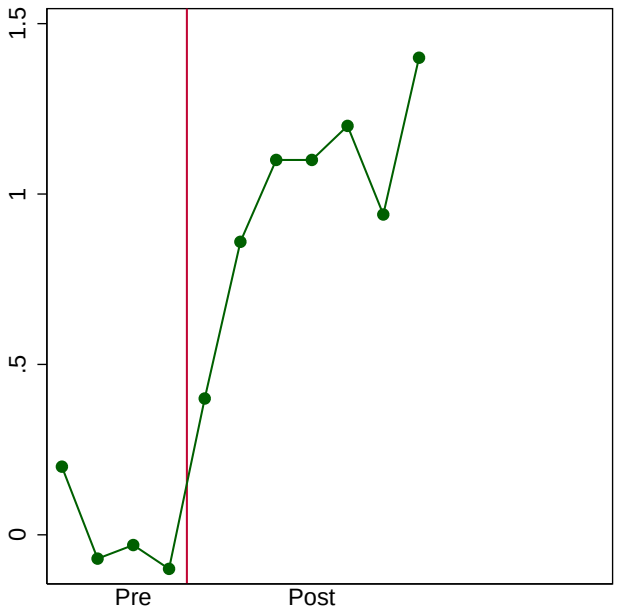
2-way Fixed Effects Models

Including additional sets of fixed effects, as for time periods, is easiest via indicator variables:

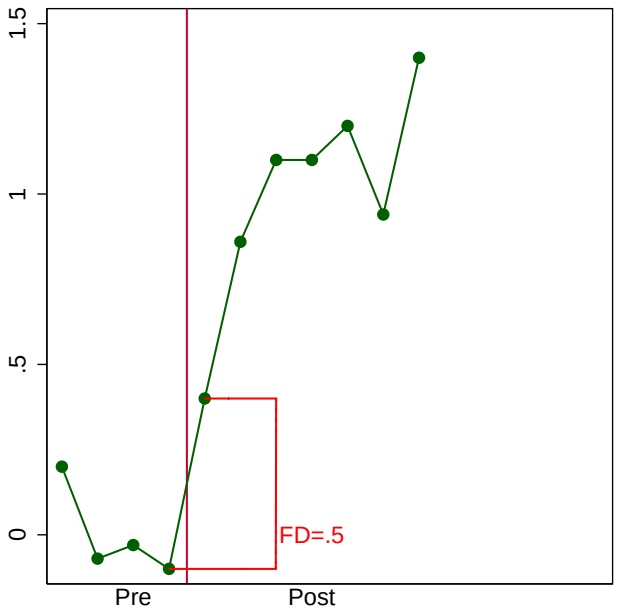
```
qui tab year, gen(iy)
drop iy1
areg y x* iy*, abs(indiv) cluster(indiv)
```

See Abowd, Creecy, and Kramarz (2002) and Andrews, Schank, and Upward (2005) for faster estimation of n-way fixed effects. See also Cameron, Gelbach, and Miller (2006) for two-way clustering of errors, and Cameron, Gelbach, and Miller (2007) for a bootstrap approach to estimating cluster-robust standard errors with fewer than 50 clusters.

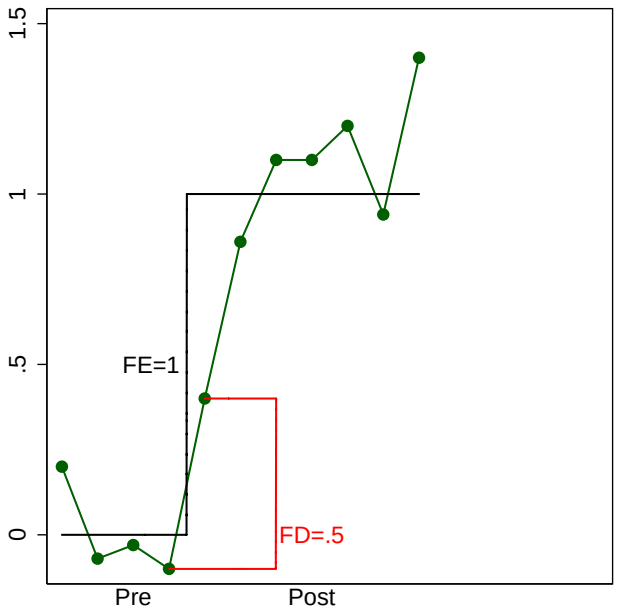
First Difference, Fixed Effects, and Long Difference



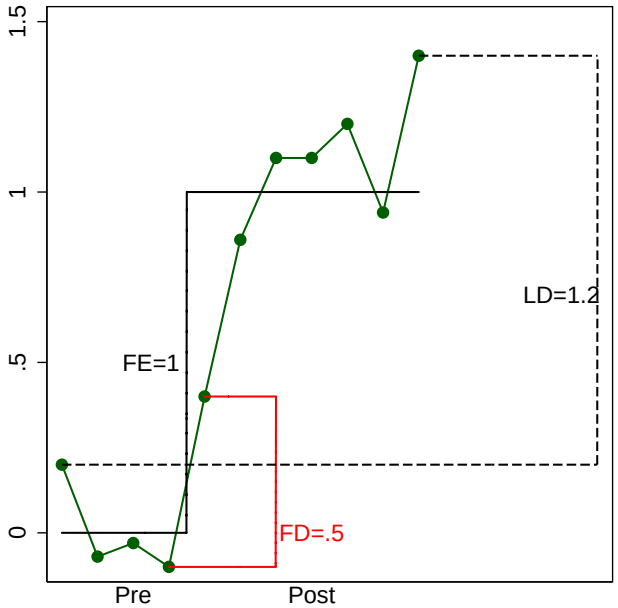
First Difference, Fixed Effects, and Long Difference



First Difference, Fixed Effects, and Long Difference



First Difference, Fixed Effects, and Long Difference



FD, FE, and LD

Clearly, one must impose some assumptions on the speed with which X affects Y, or have some evidence as to the right time frame for estimation. This type of choice comes up frequently when stock prices are supposed to have adjusted to some news, especially given the frequency of data available—economists believe the new information is capitalized in prices, but not instantaneously. Taking a difference in stock prices between 3:01pm and 3pm is inappropriate, but taking a long difference over a year is clearly inappropriate as well, since new information arrives continuously.

One should always think about within-panel trends and the frequency of measurement. Baum (2006) discussed some filtering techniques to get different frequency "signals" from noisy data. Personally, I like a simple method due to Baker, Benjamin, and Stanger (1999).

Growth Models

In the food stamps example, if you had daily observations on height and weight, you would not want to estimate the one-day change in obesity among the affected group. Similarly, for an educational intervention, the test scores on the second day are probably not the best measure, but are the test scores on the last day? In the Tennessee STAR experiment, some students were placed in smaller classes, and had higher test scores at the end of the year. If they don't have higher test scores after five years, should we care?

There is a large class of growth models, where the effect of X is assumed to be on the rate of change in Y. The natural way to specify these models is to include an elapsed time variable and interact it with X.

$$Y_{it} = X_{it}b + T_{it}g + TX_{it}f + A_i + e_{it}$$

For a binary X, the marginal effect of X for each observation is then $b + fT_{it}$ (or `_b[x]+_b[timex]*time` in Stata with T variable `time` and TX variable `timex`), and we can imagine taking the mean across relevant time periods or computing this quantity at some specified end point, i.e. we are back to choosing among first difference or long difference models.

In practice, growth models are usually estimated using hierarchical models, such as `xtmixed`, `xtmelogit`, `xtmepoisson` or some form of `gllamm` model (Rabe-Hesketh, Skrondal, and Pickles 2002).

Consistency

Note the assumption I started with: Suppose ability A is fixed for each individual i but does not change as time t passes. If ability doesn't change over time, what is the point of education? A facile observation, perhaps, but the point is that the assumed selection was of the most uncomplicated variety, and it is natural to think that people differ in unobservable ways over time as well.

If people differ in unobservable ways over time as well, or selection is more complicated, these panel methods will not provide consistent estimates of the effect b.

Other Panel Models

There are a variety of random effect and GLS methods that exploit distributional assumptions to estimate more complicated panel models, and there is the random coefficient case

$$Y_{it} = X_{it} b_i + e_{it}$$

which seems like a natural extension to the basic panel setting (see `xtrc` in Stata 10). With more assumptions come more violations of assumptions, but greater efficiency (and potentially less bias) if the assumptions hold.

In a fixed effects model, any time-invariant property drops out of the model, since there is no within-subject variation. For example, if education does not change over time for adults in the sample, it drops out of a fixed-effects model. A random-effects model does not have this property; it uses between-subject variation to identify the effect of the time-invariant characteristic. The important point is that random effects are assumed to be orthogonal to the X variables, in which case ignoring them does not introduce bias. If the individual-specific effects are not orthogonal to X, the random effects model may be badly biased. A robust Hausman test using `xtoverid` is in order after any random effects model.

Fixed Factors

All too often, researchers use a random effects model even if a Hausman test rejects the null that the random effects model provides consistent estimates, simply because they wish to estimate the impact of some factor that is fixed over time in addition to a time-varying treatment. See the help file for `xthtaylor` for an alternative that is allowed in that setting, which uses Instrumental Variables (IV).

Outline

Outline

IV methods

The idea of IV is to exploit another moment condition $E(Z'e) = 0$ when we think $E(X'e) \neq 0$. Put another way, Z moves X around in such a way that there is some exogenous variation in X we can use to estimate the causal effect of X on Y.

It is this way of characterizing IV that leads people to think they are getting an unbiased estimator, but it is worthwhile to remember the IV estimator is biased but consistent, and has substantially lower efficiency than OLS.

Thus, if a significant OLS estimate $\hat{b}$ becomes an insignificant $\widehat{b_{IV}}$ when using IV, one cannot immediately conclude that $b = 0$. Failure to reject the null should not lead you to accept it.

Assumptions and failures

There are a variety of things that can go wrong in IV, particularly if your chosen excluded instruments don't satisfy $E(Z'e) = 0$ or if Z is only weakly correlated with the endogenous X.

You should always test for endogeneity of your supposedly endogenous variable, since otherwise you would prefer OLS. Also, even if you reject exogeneity, you should test that your IV estimate differs from your OLS point estimate (not just from zero). You should also conduct overidentification tests, and identification tests, and tests for weak instruments.

Luckily, all of the tests are easily done in Stata, and some are part of official Stata as of release 10.

Forms of IV

- The IV Estimator: The instrumental variables estimator in `ivreg` from Stata 9 and before is a one-step estimator that can be thought of as equivalent to a variety of 2-step estimators; most think of it as a projection of y on the projection of X on Z.
- Two-stage Least Squares: (2SLS) is an instrumental variables estimation technique that is formally equivalent to the one-step estimator in the linear case. First, use OLS to regress X on Z and get $\hat{X} = Z(Z'Z)^{-1}Z'X$, thenS use OLS to regress y on $\hat{X}$ to get $\hat{\beta}_{IV}$.
- Ratio of Coefficients: If you have one endogenous variable X and one instrument Z, you can regress X on Z to get $\hat{\pi} = (Z'Z)^{-1}Z'X$ and regress y on Z to get $\hat{\gamma} = (Z'Z)^{-1}Z'y$, and the IV estimate $\hat{\beta}_{IV} = \hat{\gamma}/\hat{\pi}$. If X is a binary indicator variable, this ratio of coefficients method is known as the Wald estimator.
- The Control Function Approach: The most useful approach considers another set of two stages: use OLS to regress X on Z and get estimated errors $\hat{\nu} = X - Z(Z'Z)^{-1}Z'X$ then use OLS to regress y on X and $\hat{\nu}$ to get $\hat{\beta}_{IV}$.

Note that in every case the set of excluded instruments does not vary; if different instruments are to be used for different endogenous variables, you have a system estimator and should use `reg3` (and read Goldberger and Duncan 1973). Also, if you want to model nonlinearities in an endogenous variable X, e.g. by including X^2, you must treat added variables as new endogenous variables, so you may need additional excluded instruments.

One-step vs. two-step

The latter 3 two-step approaches will all give you the same answer as the one-step estimator, though you will have to adjust your standard errors to account for the two-step estimation as discussed in Wooldridge (2002, Section 12.5.2). The advantages of the control function approach are that it offers an immediate test of the endogeneity of X via a test of _b[v_x]=0:

```
reg x_endog x* z*
predict v_x, resid
reg y x_endog x* v_x
test v_x
```

(without adjusting SEs), and that generalizes to nonlinear second stage GLM estimation techniques such as probit or logit (for binary Y) and log (for nonnegative Y) links.

Other Flavors of IV: GMM

The GMM version of IV offers superior efficiency, and is implemented in Stata using `ivreg2` (see Baum, Schaffer, and Stillman (2003) and Baum, Schaffer and Stillman (2007)), or the Stata 10 command `ivregress`. GMM should be preferred in large samples if the null is rejected in a test of heteroskedasticity due to Pagan and Hall (1983) implemented in Stata as `ivhettest` by Schaffer (2004). Pesaran and Taylor (1999) discuss simulations of the Pagan and Hall statistic suggesting it performs poorly in small samples.

`ivreg2` can also estimate the "continuously updated GMM" of Hansen et al. (1996), which requires numerical optimization methods, in addition to offering numerous choices of standard error corrections.

Other Flavors of IV: LIML etc.

`ivreg2` and `ivregress` both can produce the Limited Information Maximum Likelihood (LIML) version of IV, though `ivreg2` also offers general k-class estimation, encompassing for example the UEVE estimator proposed by Devereaux (2007) to deal with measurement error. The Jackknife instrumental variables estimator (JIVE) can be estimated by `jive` (Poi 2006) though note too that Devereaux (2007) draws a link between JIVE and k-class estimators, and Davidson and MacKinnon (2006) deprecate JIVE. There are a variety of other IV methods of note, including for example the Gini IV (Schechtman and Yitzhaki 2001).

Tests for Endogeneity

If X is not endogenous, we would prefer OLS since the estimator has lower variance. So it is natural to report a test of the endogeneity of any "treatment" variables in X before presenting IV estimates. I already mentioned the control function approach; the `ivreg2` package offers a variety of other tests:

- orthog: The C statistic (also known as a "GMM distance" or "difference-in-Sargan" statistic), reported when using the `orthog(varlist)` option, is a test of the exogeneity of the excluded instruments in `varlist`.
- endog: Endogeneity tests of one or more potentially endogenous regressors can be implemented using the `endog(varlist)` option.
- ivendog: The endogeneity test statistic can also be calculated after `ivreg` or `ivreg2` by the command `ivendog`. Unlike the output of `ivendog`, the `endog(varlist)` option of `ivreg2` can report test statistics that are robust to various violations of conditional homoskedasticity.

OverID Tests

If more excluded instruments are available than there are endogenous variables, an overidentification (overID) test is feasible. Since excluded instruments can always be interacted with each other, with themselves (forming higher powers), or with other exogenous variables, it is easy to increase the number of excluded instruments.

Users of Stata 9.2 or earlier should use the `overid` command of Baum, Schaffer, Stillman, and Wiggins (2006) in the `ivreg2` package. As of Stata 10, the command `estat overid` may be used following `ivregress` to obtain the appropriate test. If the 2SLS estimator was used, Sargan's (1958) and Basmann's (1960) chi-squared tests are reported, as is Wooldridge's (1995) robust score test; if the LIML estimator was used, Anderson and Rubin's (1950) chi-squared test and Basmann's F test are reported; and if the GMM estimator was used, Hansen's (1982) J statistic chi-squared test is reported.

The null of an overID test is that the instruments Z are valid, i.e. that E(Z'e)=0, so a statistically significant test statistic indicates that the instruments may not be valid. In this case, an appeal to theorized connections between your variables may lead you to drop some excluded instruments and form others.

Identification Tests

For the parameters on k endogenous variables to be identified in an IV model, the matrix of excluded instruments must have rank at least k, i.e. they can't be collinear in the sample or in expectation. A number of tests of this rank condition for identification have been proposed, including the Anderson (1951) likelihood-ratio rank test statistic $-N\ln(1-e)$ where e is the minimum eigenvalue of the canonical correlations. The null hypothesis of the test is that the matrix of reduced form coefficients has rank $k-1$, i.e, that the equation is just underidentified. Under the null, the statistic is distributed chi-squared with degrees of freedom $L-k+1$ where L is the number of exogenous variables (included and excluded instruments). The Anderson (1951) statistic and the chi-squared version of the Cragg Donald (1993) test statistic $Ne/(1-e)$, are reported by `ivreg2` and discussed by Baum, Schaffer, and Stillman (2003, 2007).
Frank Kleibergen and Mark Schaffer produced the Stata program `ranktest` to implement the *rk* test for the rank of a matrix proposed by Kleibergen and Paap (2006). The *rk* test is a generalization of the Anderson (1951) test that allows for heteroscedasticity/autocorrelation consistent (HAC) variance estimates.
A rejection of the null for any of these rank tests indicates that the model is identified.

Tests for Weak Instruments

Weak instruments result in incorrect size of tests (typically resulting in overrejection of the null hypothesis of no effect) and increased bias. Bound, Jaeger, and Baker (1993, 1995) pointed out how badly wrong IV estimates might go if the excluded instrument is weakly correlated with the endogenous variable, and Staiger and Stock (1997) formalized the notion of weak instruments. Stock and Yogo (2005) provide tests of weak instruments (based on the Cragg and Donald (1993) statistic) for some models, which are reported by `ivreg2` when possible. Andrews, Moreira, and Stock (2006, 2007), Chao and Swanson (2005), Dufour (2003), Dufour and Taamouti (1999, 2007), Kleibergen (2007), and Stock, Wright, and Yogo (2002), among others, discuss various approaches to inference robust to weak instruments.

Anderson and Rubin (1949) propose a test of structural parameters (the AR test) that turns out to be robust to weak instruments (i.e. the test has correct size in cases where instruments are weak, and when they are not). Kleibergen (2002) proposed a Lagrange multiplier test, also called the score test, but this is now deprecated since Moreira (2001, 2003) proposed a Conditional Likelihood Ratio (CLR) test that dominates it, implemented in Stata by Mikusheva and Poi (2006).

Nichols (2006) reviews the literature on these issues. Briefly, if you have one endogenous variable and homoskedasticity and the first-stage F-stat is less than 15, use the CLR test `condtest` by Mikusheva and Poi (2006). If you have multiple endogenous variables, or H/AC/clustered errors, use the AR test, but note its confidence region need be neither bounded nor connected, and may not contain the point estimate. In theory, either the AR test or the CLR test can be inverted to produce a confidence region for the parameter or parameters of interest, but in practice this requires numerical methods and is computationally costly.

Generalizations using the CF approach

The class of "control function" approaches is simply enormous, but here I refer to the fourth "flavor" of IV above where the endogenous X variables are regressed on all exogenous variables (included and excluded instruments), error terms are predicted, and included in a regression of outcomes on X variables. The last stage, the regression of outcomes on X variables, need not be a linear regression, but might be `probit` or `logit` or `poisson` or any `glm` model. The standard errors have to be corrected for the two-stage estimation, as in Wooldridge (2002, Section 12.5.2), but in practice, the corrections seem to make little difference to estimated standard errors. The bootstrap is an easier and generally more robust correction in practice.

See Imbens and Wooldridge (2007; lecture 6) for much more detail on various control function approaches.

Generalizations to nonnegative outcomes

With a nonnegative outcome variable, such as a count, or earnings, or the like, a `glm`-style estimate seems preferable. `ivpois` on SSC is a GMM estimator of such a model, and Stata 11 now includes a generalized GMM estimator for other customized IV models.

The zeroth stage

Generated regressors normally require corrections to standard errors; this is not the case for excluded instruments. So various estimation strategies to estimate Z are allowed before running an IV model (like a stage zero before the stage 1 and 2 of IV). In fact, a wide variety of specification search is "allowed" in the first stage as well.

Wooldridge (2002) procedure 18.1 is a useful implementation of this: a binary endogenous variable X is no problem for IV, but estimation is typically woefully inefficient. Improved efficiency may be obtained by first regressing X on the included and excluded instruments via `probit` or `logit`, predicting the probability $\hat{X}$, and using $\hat{X}$ as the single excluded instrument.

A related approach is to predict a continuous endogenous X in some previous step and then use the prediction $\hat{X}$ as the instrument Z in the IV regression (e.g. Dahl and Lochner 2005).

Note that the weak instrument diagnostics will fail miserably using these approaches!

treatreg

For the case of a single binary endogenous variable, as an alternative to Wooldridge (2002) procedure 18.1 or straight IV, the `treatreg` command offers increased efficiency if distributional assumptions are met. The two-step estimator is another control function approach, where the inverse Mills ratio (a generalized residual) is the predicted component in the first stage regression that is then included in the second stage; MLE is the default.

These 3 consistent estimators can produce very different estimates and inference in small samples:

```
sysuse auto, clear
treatreg pri wei, treat(for=mpg)
treatreg pri wei, treat(for=mpg) two
ivreg pri wei (for=mpg)
qui probit for wei mpg
qui predict ghat if e(sample)
ivreg pri wei (for=ghat)
```

Other binary variable models

With a binary outcome and a continuous endogenous variable, `ivprobit` is the method of choice.

With a binary outcome and a binary endogenous variable, `biprobit` is one approach. `cmp` on SSC gives identical answers in this case, but can be adapted to numerous other types of models using the same basic strategy.

There are a variety of semiparametric methods not currently programmed in Stata, some of which use ideas in Klein and Spady (1993) and Klein and Vella (2009), exploiting assumptions about second moments, and using kernel density estimates (sometimes with higher-order kernels, or trimming).

Outline

The RD Design

The idea of the Regression Discontinuity (RD) design (due to Thistlewaite and Campbell 1960) is to use a discontinuity in the level of treatment related to some observable to get a consistent LATE estimate, by comparing those just eligible for the treatment to those just ineligible.

Hahn, Todd, and Van der Klaauw (2001) is the standard treatment, and a number of papers on RD appear in a special issue of the Journal of Econometrics, including notably a practical guide by Imbens and Lemieux (2008). Cook (2008) provides an entertaining history of the method's development. Lee and Card (2008) discuss specification error in RD.

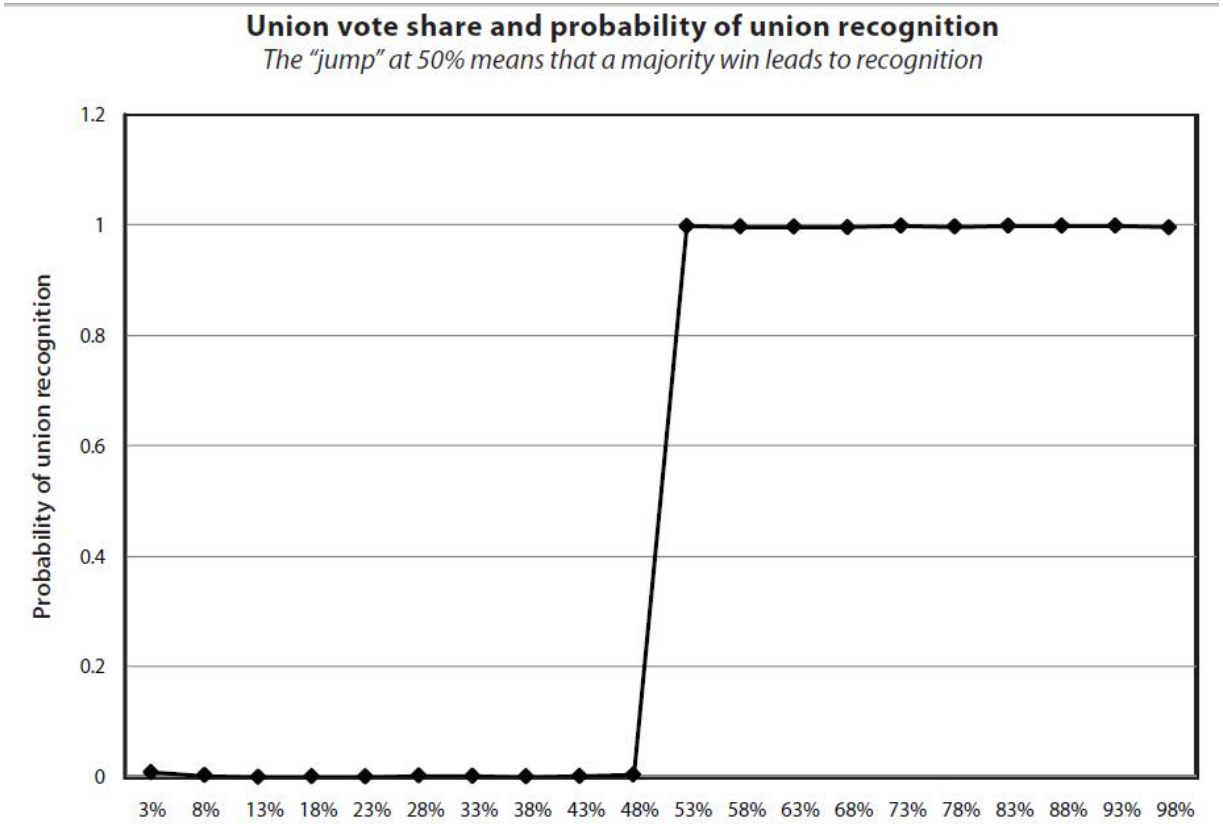
The RD design is generally regarded as having the greatest internal validity of all quasi-experimental methods. Its external validity is less impressive, since the estimated treatment effect is local to the discontinuity.

RD Design Validity

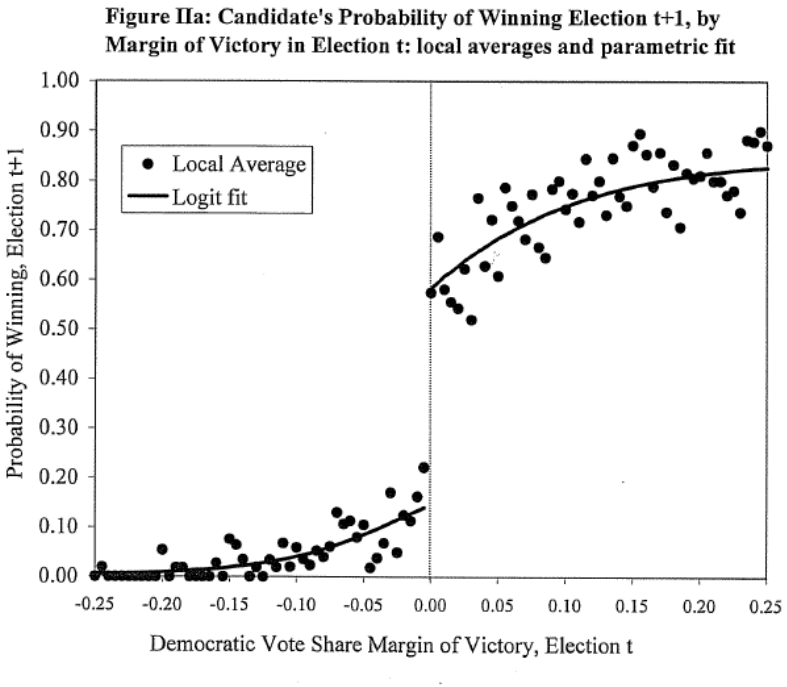
For example, the What Works Clearinghouse (established in 2002 by the U.S. Department of Education's Institute of Education Sciences) uses Evidence Standards to "identify studies that provide the strongest evidence of effects: primarily well conducted randomized controlled trials and regression discontinuity studies, and secondarily quasi-experimental studies of especially strong design."

- "Meets Evidence Standards"—randomized controlled trials (RCTs) that do not have problems with randomization, attrition, or disruption, and regression discontinuity designs that do not have problems with attrition or disruption.
- "Meets Evidence Standards with Reservations"—strong quasi-experimental studies that have comparison groups and meet other WWC Evidence Standards, as well as randomized trials with randomization, attrition, or disruption problems and regression discontinuity designs with attrition or disruption problems.
- "Does Not Meet Evidence Screens"—studies that provide insufficient evidence of causal validity or are not relevant to the topic being reviewed.

One Voting Example (first stage)



Another Voting Example (second stage)



RD Design Elements

What do we need for an RD design?

The first assumption is that treatment is not randomly assigned, but is assigned based at least in part on a variable we can observe. I call this variable the assignment variable, or Z, but it is often called the "running" or "forcing" variable.

The crucial second assumption is that there is a discontinuity at some cutoff value of the assignment variable in the level of treatment. For example, in the food stamps example, immigrants in the country five years are eligible, those in the country one day or one hour less than five years are not.

RD Design cont.

The third crucial assumption is that individuals cannot manipulate the assignment variable (e.g. by backdating paperwork) to affect whether or not they fall on one side of the cutoff or the other, or more strongly, observations on one side or the other are **exchangeable** or otherwise identical.

The fourth crucial assumption is that the other variables are smooth functions of the assignment variable conditional on treatment, i.e. the only reason the outcome variable should jump at the cutoff is due to the discontinuity in the level of treatment. (Actually, only continuity in Z of potential outcomes $Y(X)$ at the cutoff is required, but some global smoothness is an appealing and more testable assumption.)

Note this differs from IV, in that the assignment variable Z can have a direct impact on the outcome Y, not just on the treatment X, though not a discontinuous impact.

Deterministic or Probabilistic Assignment

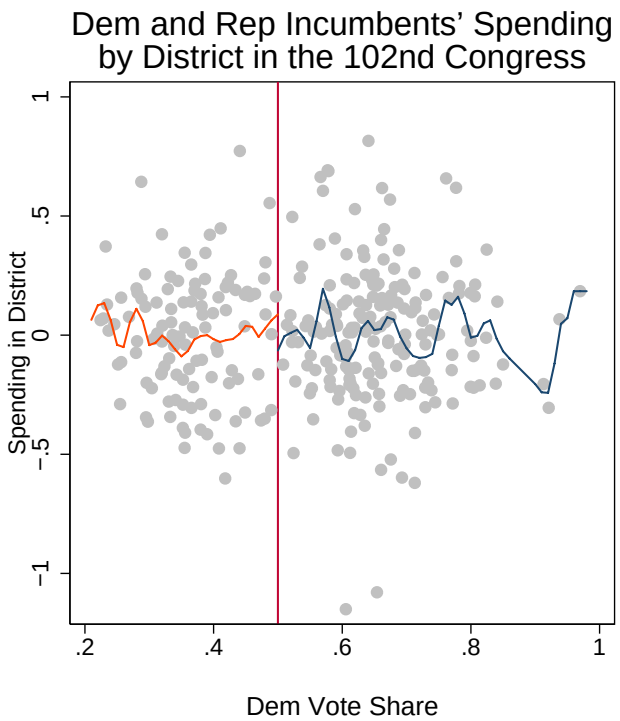
In one version of RD, everyone above some cutoff gets treatment (or a strictly higher level of treatment), and in another, the conditional mean of treatment jumps up at the cutoff. These two types are often called the Sharp RD Design and the Fuzzy RD Design, but I don't much like the Fuzzy versus Sharp terminology. The discontinuity must be "sharp" in either case. One issue is whether there is common support for the propensity of treatment—if not, then we can say for certain that folks to the left of the cutoff certainly don't get more than $\bar{x}$ amount of treatment and folks to the right get no less than $\underline{x}$, and $\bar{x} < \underline{x}$ gives us one kind of "sharp" design since there is no overlap.

The special case where we know the conditional mean of treatment above and below the cutoff, as with a binary treatment where $\bar{x} = 0$ and $\underline{x} = 1$, I call the deterministic RD design since there is "deterministic assignment" of treatment conditional on the observed assignment variable. In any other case, we have to estimate the jump in the conditional mean of treatment at the discontinuity.

Voting

One obvious "deterministic assignment" example of a regression discontinuity occurs in elections with two options, such as for/against (e.g. unionization) or Republican/Democrat. Across many firms, those which unionized with a pro-union vote of 80% are likely quite different along many dimensions from those that failed to unionize with a pro-union vote of 80%. Across many congressional districts, those with 80% voting for Republicans are likely quite different from those with 80% voting for Democrats. However, the firm that unionizes with 50% voting for the union is probably not appreciably different from the one that fails to unionize with 49.9% voting for the union. Whether or not those people casting the few pivotal votes showed up or not is often due to some entirely random factor (they were out sick, or their car didn't start, or they accidentally checked the wrong box on the ballot). DiNardo and Lee (2002) found little effect of unions using an RD design. Lee, Moretti, and Butler (2004) looked at the effect of party affiliation on voting records in the US Congress (testing the median voter theorem's real-world utility) and Lee (2001) looked at the effect of incumbency using similar methods.

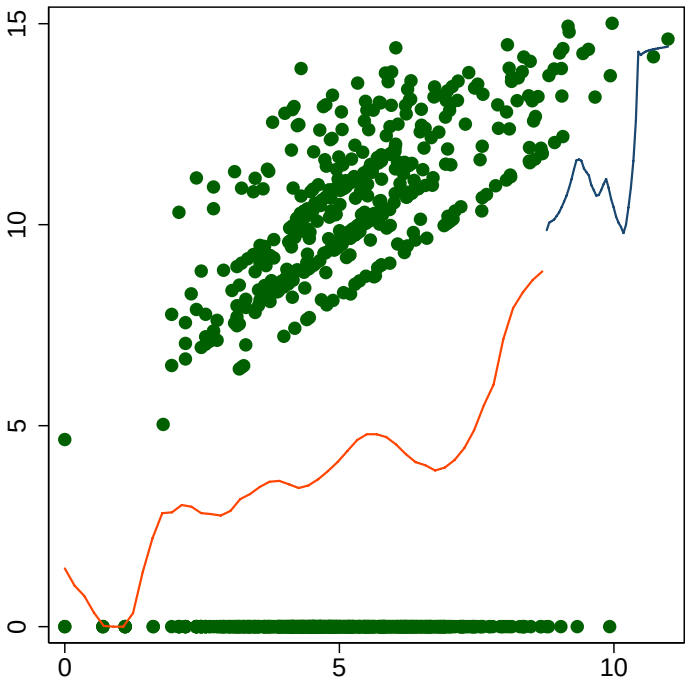
Another Voting Example



Educational grant example

An example of "probabilistic assignment" is a US Department of Education grant program that is available to high-poverty school districts within each state. High-poverty districts are clearly different from low-poverty districts, but districts on either side of the cutoff point are essentially exchangeable. The first difference here is that districts that qualify (are above the cutoff) do not automatically get the grant; they must apply. A second difference is that low-poverty districts can enter into consortia with high-poverty districts and get funds even though they are below the cutoff. However, a district cannot unilaterally apply, which creates a discontinuity in the costs of applying at the cutoff, and we can observe the assignment variable and the cutoff to see if there is a discontinuity in funds received:

Educational grants as a function of poor students



Local Wald Estimator

In the "sharp" or "deterministic assignment" version, the estimated treatment effect is just the jump in expected outcomes at the cutoff: in other words, the expected outcome for units just above the cutoff (who get treated), call this y^+, minus the expected outcome for units just below the cutoff (who don't get treated, but are supposed to be otherwise identical), call this y^-, or

$$\widehat{LATE} = (y^+ - y^-)$$

since the jump in the level of treatment is exactly one unit at the cutoff.

In the "fuzzy" or "probabilistic assignment" version, the jump in outcomes is "caused" by some jump in treatment that need not be one—but the ratio of coefficients method for IV, the Wald estimator, suggests how to estimate the effect of a unit change in treatment: just form the ratio of the jump in outcomes to the jump in treatment. The Local Wald Estimator of LATE is thus $(y^+ - y^-)/(x^+ - x^-)$, where $(x^+ - x^-)$ is the estimated discontinuous jump in expected treatment. Note that this second estimator reduces to the first given "deterministic assignment" since $(x^+ - x^-) = 1$ in this case, so the distinction between "sharp" and "fuzzy" RD is not too sharp.

LATE

The RD estimator assumes "as-if random" assignment of the level of treatment in the neighborhood of the cutoff, so that observations with levels of the assignment variable Z close to the cutoff Z_0 form the "experimental" group. Just as is often the case with true experiments, we cannot generalize the estimated treatment effect to the rest of the population; for RD, our estimated treatment effect technically applies only to individuals with Z exactly equal to Z_0, i.e. a set of measure zero in the population. RD is therefore a very localized sort of Local Average Treatment Effect (LATE) estimator with high internal validity and low external validity.

In this respect, RD is most like a RCT, in which subjects are often selected nonrandomly from the population and then randomly assigned treatment, so the unbiased estimate of average treatment effects applies only to the type of subpopulation selected into the subject pool. At least in RD, we can characterize exactly the population for whom we estimate the LATE: it is folks with Z exactly equal to Z_0.

Polynomial or Local Polynomial in Z

There is a great deal of art involved in the choice of some continuous function of the assignment variable Z for treatment and outcomes. The researcher chooses some high-order polynomial of Z to estimate separately on both sides of the discontinuity, or better, a local polynomial, local linear, or local mean smoother, where the art is in the choice of kernel and bandwidth. Stata 10 now offers `lpoly` which supports `aweights`; users of prior versions can `findit locpoly` and `expand` their data to get weighted local polynomial estimates. The default in both is local mean smoothing, but local linear regression is preferred in RD designs.

Choice of Bandwidth

There are several rule-of-thumb bandwidth choosers and cross-validation techniques for automating bandwidth choice, but none is foolproof. McCrary (2007) contains a useful discussion of bandwidth choice, and claims that there is no substitute for visual inspection comparing the local polynomial smoother with the pattern in the scatterplot.

Because different bandwidth choices can produce different estimates, the researcher should really report more than one estimate, or perhaps at least three: the preferred bandwidth estimate, and estimates using twice and half the preferred bandwidth. I believe a future method might incorporate uncertainty about the bias-minimizing bandwidth by re-estimating many times using different bandwidth choices.

As it is, though, local polynomial regression is estimating hundreds or thousands or even hundreds of thousands of regressions. Bootstrapping these estimates requires estimating millions of regressions, or more. Still, computing time is cheap now.

More Choices

Of somewhat less importance, but equally art over science, is the **choice of kernel**. Most researchers use the default epanechnikov kernel, but the triangle kernel typically has better properties at boundaries, and it is the estimates at the boundaries that matter in this case.

Show the Data

Given how much choice the researcher has over parameters in a supposedly nonparametric strategy, it is always wise to show a `scatter` or `dotplot` of the data with the local polynomial smooth superimposed, so readers may be reassured no shenanigans of picking parameters were involved.

Testing for Existence of a Discontinuity

The first test should be a test that the hypothesized cutoff in the assignment variable produces a jump in the level of treatment. In the voting example, this is easy: the probability of winning the election jumps from zero to one at 50%, but in other settings the effect is more subtle: in the education example, the discontinuity is far from obvious.

In any case, the test for a discontinuity in treatment X is the same as a test for a discontinuity in the outcome Y. Simply estimate a local linear regression of X on Z, both above and below the cutoff, perhaps using a triangle kernel with a bandwidth that guarantees 10-20 observations are given positive weight at the boundary, approaching the cutoff from either side. The local estimate at the cutoff for regressions constrained below the cutoff is x^- and the local estimate at the cutoff for regressions constrained above the cutoff is x^+. The computation of the difference $x^+ - x^-$ can be wrapped in a program and bootstrapped for a test of discontinuity.

Testing for a Discontinuity

Assume for the sake of the example that the assignment variable Z is called `share` and ranges from 0 to 100 (e.g. percent votes received) with an assignment cutoff at 50. Then we could write a simple program:

```
prog discont, rclass
version 10
cap drop z f0 f1
g z=_n in 1/99
lpoly ‘1’ share if share<50,  gen(f0) at(z) nogr k(tri) bw(2) deg(1)
lpoly ‘1’ share if share>=50, gen(f1) at(z) nogr k(tri) bw(2) deg(1)
return scalar d=‘=f1[50]-f0[50]’
end
bootstrap r(d), reps(1000): discont x
```

The computation of the difference in outcomes $y^+ - y^-$ is obtained by replacing `x` with `y` above. The Local Wald Estimator of LATE is then $(y^+ - y^-)/(x^+ - x^-)$, which quantity can also be computed in a program and bootstrapped. The SSC package `rd` automates many of these tasks. Imbens and Lemieux (2008) provide analytic standard error formulae. Imbens and Kalyanaraman (2009) discuss bandwidth choice.

Testing for Sorting at the Discontinuity

McCrary (2007) gives a very detailed exposition of how one should test this assumption by testing the continuity of the density of the assignment variable at the cutoff. As he points out, the continuity of the density of the assignment variable is neither necessary nor sufficient for exchangeability, but it is reassuring.

McCrary (2007) also provides tests of sorting around the discontinuity in voting in US Congressional elections, where there is no sorting, and in roll-call votes in Congress, where there is sorting.

Testing for Extraneous Discontinuities in Y and X

Another useful test is that there are no extra jumps in the levels of treatment or the outcome (both should be smooth functions of the assignment variable at other points) where no hypothesized cutoff exists. This is a test of the fourth crucial assumption. One can easily pick 100 random placebo cutoff points, and test the difference in X and the difference in Y (about 5 placebo cutoffs will show significant jumps, of course).

Testing for Extraneous Discontinuities in Other Potential Outcomes

An important though supererogatory piece of evidence is that there are no jumps in variables that are not suspected to be affected by the discontinuity. This amounts to estimating the difference $h^+ - h^-$ at the cutoff for every other variable h (e.g. demographic characteristics, etc.).

This makes for some long papers, with page after page of scatterplots showing no relationship where none was expected, but is very reassuring that there is not some major difference between treated and untreated observations around the discontinuity (i.e. these graphs are more evidence in favor of exchangeability).

Outline

Sensitivity Testing

The exposition by Manski (1995) demonstrates how a causal effect can be bounded under very unrestrictive assumptions, and then the bounds can be narrowed under more restrictive parametric assumptions. Given how sensitive the QE methods are to assumptions (selection on observables, exclusion restrictions, exchangeability, etc.), some kind of sensitivity testing is order no matter what method is used.

Rosenbaum (2002) provides a wealth of detail on formal sensitivity testing under various parametric assumptions. Eliason (2007) provides a short Stata example of calculating Rosenbaum bounds on treatment effects using `psmatch2` and `rbounds`, due to DiPrete and Gangl (2004), who compare Rosenbaum bounds in a matching model to IV estimates. `sensatt` by Nannicini (2006) and `mhbounds` by Becker and Caliendo (2007) are additional Stata programs to aid in the construction of bounds. Rosenbaum's "gamma" measure of sensitivity is a useful summary measure of potential sensitivity, often estimated via simulation.

Lee (2005) advocates another very useful method of bounding treatment effects, used in Liebbrandt, Levinsohn, and McCrary (2005).

Panel and IV methods

Estimators such as `xtivreg` (see also `xtivreg2` due to Schaffer 2007) or the Arellano and Bond (1991) estimator in `xtabond` (see also `xtabond2` and Roodman 2006) offer a combination of panel methods and IV estimation.

Panel matching/reweighting

Many of the reweighting papers e.g. DiNardo, Fortin, and Lemieux (1996), Autor, Katz, and Kearney (2005), etc., are marriages of matching/reweighting estimators and panel methods.

The preferred estimator in Heckman, Ichimura, and Todd (1997) is a marriage of matching and diff-in-diff estimation.

Reweighted IV or RD

It is also interesting to consider reweighting so compliers in IV (those induced to take a binary treatment by a single excluded binary instrument) look like the rest of the distribution in observable variables, or more generally to match or reweight to impute the LATE estimates to the rest of the sample, and get an ATE estimate.

Similarly, one can imagine reweighting/matching marginal cases in RD to get at the ATE. I have not seen this in the literature, though; probably the finite sample performance is poor.

RD meets IV

As mentioned above, the LATE estimate in the so-called "fuzzy RD" design $(y^+ - y^-)/(x^+ - x^-)$ is a Local Wald Estimator, or a type of local IV.

If one were willing to dispose of local polynomials and assume a form for X and Y as functions of the assignment variable Z, the RD approach can be recast as straight IV where the terms with Z are included instruments and an indicator for Z above the cutoff is the sole excluded instrument.

RD meets DD

One can also imagine estimating a diff-in-diff version of the RD estimator, given the advent of some policy with an eligibility cutoff, where the difference across times t and 0:

$$\left[(y_t^+ - y_t^-) - (y_0^+ - y_0^-)\right] / (x_t^+ - x_t^-)$$

would be the estimated program impact. One could also estimate the difference in local Wald estimates

$$\left[(y^+ - y^-)/(x^+ - x^-)\right]_t - \left[(y^+ - y^-)/(x^+ - x^-)\right]_0$$

if the difference in x in the "pre" period is nonzero (if $(x^+ - x^-)$ might be zero, there would be a lot of instability in the estimate).

An application where this might be useful is if we expect an underlying discontinuity at the cutoff in the absence of treatment but we can use the observed jump in x and y before treatment begins to difference that out. For example, a new treatment is applied only to those 65 or older, but there is already an effect at 65 due to a jump in eligibility for Medicare (a large public health insurance system). Or a new treatment is applied only to those whose children are 18 or older, but there is already an effect at 18 due to parents' ideas about when children should fend for themselves. If $y_0^+ - y_0^-$ (and/or $x_0^+ - x_0^-$) is nonzero, we can give up the internal validity of regression discontinuity, and downgrade to the internal validity of panel estimators, but get an unbiased estimate under stronger conditions. If the jumps at the cutoff are not changing over time in the absence of treatment, the differenced local Wald estimators will be unbiased for the local average treatment effect.

Conclusions

None of these methods is perfect. The gold standard, an RCT, has the best internal validity but may have poor external validity. Of methods using observational data, the RD design is closest to an RCT, and also has high internal validity but low external validity. IV methods can eliminate bias from selection on unobservables in the limit, but may have very poor performance in finite samples. The hypothetical internal validity of IV is high, but the practical internal validity of IV is often low, and the external validity not much greater than RD.

Panel methods can eliminate bias from selection on unobservables that do not change over time, or satisfy other strong distributional assumptions, but the required assumptions are often untenable in practice. Matching and reweighting methods can eliminate bias due to selection on observables, and give efficient estimates of many types of treatment effects in many settings, but it is rarely the case that selection depends only on observables, in which case matching can actually exacerbate bias. Regression or matching methods applied to population data often have very high external validity, but internal validity that is often questionable.

Conclusions cont.

In practice, the data often dictate the method. If one has access to experimental data, one worries less about selection (though IV is often used to correct for selection of treatment status contrary to assignment). Given observational data, if one can find a discontinuity in expected treatment with respect to an observable assignment variable, one uses RD; if one can conceive of plausible excluded instruments, one uses IV. In the absence of these features of the data, repeated measures may used to control for invariant unobservables, or observations may be matched on observables.

Checking that your model is not badly misspecified, and conducting various kinds of sensitivity tests, is perhaps the most valuable way to minimize bias in published estimates. Nichols (2007, 2008) offers a kind of "checklist" of things to look at in these models (in Stata) and there will be a monograph with more user-friendly text and examples later this year (forthcoming from Stata Press).

Abadie, Alberto and Guido W. Imbens. 2006. "On the Failure of the Bootstrap for Matching Estimators." NBER technical working paper 325.

Abadie, Alberto, David Drukker, Jane Leber Herr, and Guido W. Imbens, 2004. "Implementing matching estimators for average treatment effects in Stata," *Stata Journal* 4(3): 290-311

Abadie, Alberto, Joshua D. Angrist, and G. Imbens, (2002), "Instrumental Variables Estimation of Quantile Treatment Effects," *Econometrica* 70(1): 91-117.

Abadie, Alberto, and Guido W. Imbens. 2002. "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," NBER technical working paper 283.

Abowd, J., Creecy, R. and Kramarz, F. 2002. "Computing person and firm effects using linked longitudinal employer-employee data." Technical Paper 2002-06, U.S. Census Bureau.

Anderson, T. W. and H. Rubin (1949). "Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations." *Annals of Mathematical Statistics*, 21: 570-582.

Andrews, Donald W. K.; Marcelo J. Moreira; and James H. Stock. 2007. "Performance of Conditional Wald Tests in IV Regression with Weak Instruments." *Journal of Econometrics*, 139(1): 116-132. Working paper version online with supplements at Stock's website.

Andrews, Donald W. K.; Marcelo J. Moreira; and James H. Stock. 2006. "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression." *Econometrica* 74: 715-752. Earlier version published as NBER Technical Working Paper No. 299 with supplements at Stock's website.

Andrews, Martyn, Thorsten Schank, and Richard Upward. "Practical fixed effects estimation methods for the three-way error components model." University of Nottingham Working Paper.

Angrist, Joshua D. and Alan B. Krueger. 2000. "Empirical Strategies in Labor Economics," in A. Ashenfelter and D. Card eds. *Handbook of Labor Economics, vol. 3*. New York: Elsevier Science.

Angrist, Joshua D., Guido W. Imbens and D.B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91, 444-472.

Angrist, Joshua D., Guido W. Imbens, and Alan B.Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Applied Econometrics* 14(1): 57-67.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics* Princeton: Princeton University Press.

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics*, 49: 431-34.

Arellano, M., and S. Bond. 1991. "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations." *Review of Economic Studies* 58: 277-297.

Athey, Susan and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74 (2): 431-497.

Autor, David H., Lawrence F. Katz, Melissa S. Kearney. 2005. "Rising Wage Inequality: The Role of Composition and Prices." NBER Working Paper 11628.

Baker, Michael, Dwayne Benjamin, and Shuchita Stanger. 1999. "The Highs and Lows of the Minimum Wage Effect: A Time-Series Cross-Section Study of the Canadian Law." *Journal of Labor Economics*, 17(2): 318-350.

Basmann, R.L. 1960. "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics." *Journal of the American Statistical Association* 55(292): 650-59.

Baum, Christopher F. 2006. "Time-series filtering techniques in Stata." Presented at NASUG5.

Baum, Christopher F., Mark E. Schaffer, Steven Stillman, and Vince Wiggins. 2006. "overid: Stata module to calculate tests of overidentifying restrictions after ivreg, ivreg2, ivprobit, ivtobit, reg3." RePEc or `findit overid`.

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2007. "Enhanced routines for instrumental variables/GMM estimation and testing." Unpublished working paper.

Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2003. "Instrumental variables and GMM: Estimation and testing." *Stata Journal* 3(1), 1-31. Also Boston College Department of Economics Working Paper No 545

Becker, Sascha O. and Andrea Ichino. 2002. "Estimation of average treatment effects based on propensity scores", *The Stata Journal* 2(4): 358-377. Also `findit pscore` for updates (e.g. *Stata Journal* 5(3): 470).

Becker, Sascha O. and Marco Caliendo, 2007. "mhbounds - Sensitivity Analysis for Average Treatment Effects." IZA Discussion Paper 2542.

Blinder, Alan S. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *The Journal of Human Resources* 8(4): 436-455.

Bound, John, David A. Jaeger, and Regina Baker. 1993. "The Cure Can Be Worse than the Disease: A Cautionary Tale Regarding Instrumental Variables." NBER Technical Working Paper No. 137.

Bound, John, David A. Jaeger, and Regina Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation Between the Instruments and the Endogenous Explanatory Variables is Weak." *Journal of the American Statistical Association*, 90(430), 443-450.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2006. "Robust Inference with Multi-Way Clustering." NBER Technical Working Paper T0327.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2007. "Bootstrap-Based Improvements for Inference with Clustered Errors" FSU College of Law, Law and Economics Paper 07/002.

Chao, John C. and Norman R. Swanson. (2005). "Consistent Estimation with a Large Number of Weak Instruments." *Econometrica*, 73(5), 1673-1692. Working paper version available online.

Cochran, William G. 1965. "The Planning of Observational Studies of Human Populations" and discussion. *Journal of the Royal Statistical Society* A128(2): 234-266.

Cochran, William G., and Donald B. Rubin. 1973. "Controlling Bias in Observational Studies: A Review." *Sankhya* 35: 417-46.

Cook, Thomas D. 2008. "Waiting for Life to Arrive: A History of the Regression-Discontinuity Design in Psychology, Statistics and Economics." *Journal of Econometrics,* 142(2).

Cragg, J.G. and S.G. Donald. (1993). "Testing Identifiability and Specification in Instrumental Variable Models," *Econometric Theory*, 9, 222-240.

Dahl, Gordon and Lance Lochner. 2005. "The Impact of Family Income on Child Achievement." NBER Working Paper 11279.

Davidson, J. and MacKinnon. 2006. "The Case against JIVE." *Journal of Applied Econometrics* 21: 827–833.

Devereux, Paul J. 2007. "Improved Errors-in-Variables Estimators for Grouped Data." *Journal of Business and Economic Statistics*, 25(3): 278-287.

DiNardo, John. 2002. "Propensity Score Reweighting and Changes in Wage Distributions" University of Michigan Working Paper.

DiNardo, John and David Lee. 2002. "The Impact of Unionization on Establishment Closure: A Regression Discontinuity Analysis of Representation Elections." NBER Working Paper 8993.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica*, 64(5): 1001-1044.

DiNardo, John and Justin L. Tobias. 2001. "Nonparametric Density and Regression Estimation." *The Journal of Economic Perspectives*, 15(4): 11-28.

DiPrete, Thomas A. and Markus Gangl. 2004. "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments." *Sociological Methodology*, 34: 271-310. Stata code to estimate Rosenbaum bounds

Dufour, Jean-Marie. 2003. "Identification, Weak Instruments, and Statistical Inference in Econometrics." *Canadian Journal of Economics*, 36, 767-808.

Dufour, Jean-Marie and Mohamed Taamouti. 1999; revised 2003. "Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments." Manuscript, Department of Economics, University of Montreal.

Dufour, Jean-Marie and Mohamed Taamouti. 2007. "Further results on projection-based inference in IV regressions with weak, collinear or missing instruments." *Journal of Econometrics*, 139(1): 133-153.

Eliason, Scott R. 2007. "Calculating Rosenbaum Bounds in Stata: Average Causal Effects of College v. HS Degrees on Wages Example." University of Minnesota paper. See also this website.

Fisher, Ronald A. 1918. "The causes of human variability." *Eugenics Review* 10: 213-220.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, Ronald A. 1926. "The arrangement of field experiments." *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513.

Frölich, Markus. 2007a. "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates." *Journal of Econometrics* 139 (1), 35-75. Also IZA Discussion Paper 588.

Frölich, Markus. 2007b. "Propensity score matching without conditional independence assumptionwith an application to the gender wage gap in the United Kingdom." *The Econometrics Journal* 10(2), 359407.

Frölich, Markus. 2004. "What is the Value of Knowing the Propensity Score for Estimating Average Treatment Effects?" *Econometric Reviews* 23(2): 167-174. Also IZA Discussion Paper 548.

Goldberger, Arthur S. 1972. "Selection bias in evaluating treatment effects: Some formal illustrations." Discussion paper 123-72, Institute for Research on Poverty, University of Wisconsin, Madison.

Goldberger, Arthur S., and 0tis D. Duncan. 1973. *Structural Equation Models in the Social Sciences*. New York: Seminar Press.

Gomulka, Joanna, and Nicholas Stern. 1990. "The Employment of Married Women in the United Kingdom 1970-83." *Economica* 57: 171-199.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1): 201-209.

Hardin, James W., Henrik Schmiediche, and Raymond J. Carroll. 2003. "Instrumental variables, bootstrapping, and generalized linear models." *Stata Journal* 3(4): 351-360. See also http://www.stata.com/merror/.

Heckman, James J. and Edward Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences of the United States of America* 96:4730-34.

Heckman, James J. and Edward Vytlacil. 2000. "The Relationship between Treatment Parameters within a Latent Variable Framework." *Economics Letters* 66:33-39.

Heckman, James J. and Edward Vytlacil. 2004. "Structural Equations, Treatment Effects and Econometric Policy Evaluation." *Econometrica* 73(3): 669738.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64(4): 605-654.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71(4): 1161-1189. (See also NBER WP T0251.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association* 8(396): 945-960.

Imbens, Guido and Karthik Kalyanaraman. 2009. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator, NBER working paper 14726.

Imbens, Guido and Jeffrey Wooldridge. "What's New in Econometrics." NBER Summer Institute Course notes.

Imbens, Guido and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics,* 142(2). See also NBER Working Paper 13039.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86(1): 4-29, 06. Earlier version available as NBER Technical Working Paper 0294.

Imbens, Guido W. 2006. "Matching methods for estimating treatment effects using Stata." Presented at NASUG6.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467-75

Jann, Ben. 2007. "Univariate kernel density estimation." Working paper and Stata package.

Juhn, Chinhui, Kevin M. Murphy, Brooks Pierce. 1993. "Wage Inequality and the Rise in Returns to Skill." *Journal of Political Economy* 101(3): 410-442.

Juhn, Chinhui, Kevin M. Murphy, Brooks Pierce. 1991. "Accounting for the Slowdown in Black-White Wage Convergence." in *Workers and Their Wages*, ed. Marvin Kosters, Washington, DC: AEI Press.

Kaushal, Neeraj. 2007. "Do Food Stamps Cause Obesity? Evidence from Immigrant Experience." NBER Working Paper No. 12849

Kézdi, Gábor. 2004. "Robust Standard Error Estimation in Fixed-Effects Panel Models." Hungarian Statistical Review Special(9): 96-116.

Kinal, Terrence W. 1980. "The Existence of Moments of k-Class Estimators." *Econometrica*, 48(1), 241-250.

Kleibergen, Frank. 2002. "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression." *Econometrica*, 70(5), 1781-1803.

Kleibergen, Frank. 2007. "Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics." *Journal of Econometrics* 139(1): 181-216.

Kleibergen, Frank and Richard Paap. 2006. "Generalized reduced rank tests using the singular value decomposition." *Journal of Econometrics* 133(1): 97-126. Preprint.

Klein, Roger W. and Francis Vella. 2009. "A Semiparametric Model for Binary Response and Continuous Outcomes under Index Heteroscedasticity" *Journal of Applied Econometrics,* 24(5): 735–762.

Klein, Roger W. and Richard H. Spady. 1993. "An efficient semiparametric estimator for discrete choice models" *Econometrica,* 61: 387–421. http://www.jstor.org/stable/pdfplus/2951556.pdf

Lee, David S., Enrico Moretti, and Matthew J. Butler. 2004. "Do Voters Affect Or Elect Policies? Evidence From The U. S. House." *Quarterly Journal of Economics* 119(3): 807-859.

Lee, David S. 2001. "The Electoral Advantage to Incumbency and Voters' Valuation of Politicians' Experience: A Regression Discontinuity Analysis of Elections to the U.S. House." NBER Working Paper 8441. New version "Randomized Experiments from Non-random Selection in U.S. House Elections" forthcoming in *Journal of Econometrics* with a Supplemental Mathematical Appendix.

Lee, David S. 2005. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." NBER Working Paper 11721 with errata. Previous version: Trimming for Bounds on Treatment Effects with Missing Outcomes, NBER Technical Working Paper 277.

Lee, David S. and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics,* 142(2). See also NBER Technical Working Paper 322 and previous version: Center for Labor Economics Working Paper 74.

Lee, Lung-Fei. 1992. "Amemiya's Generalized Least Squares and Tests of Overidentification in Simultaneous Equation Models with Qualitative or Limited Dependent Variables." *Econometric Reviews* 11(3): 319-328.

Leibbrandt, Murray, James Levinsohn, and Justin McCrary. 2005. "Incomes in South Africa Since the Fall of Apartheid." NBER Working Paper 11384.

Leuven, Edwin and Barbara Sianesi. 2003. "psmatch2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing." RePEc or `findit psmatch2`.

Machado, Jos and Jos Mata. 2005. "Counterfactual Decompositions of Changes in Wage Distributions Using Quantile Regression." *Journal of Applied Econometrics* 20(4): 445-65.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.

McCrary, Justin. 2007. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." NBER Technical Working Paper 334.

Mikusheva, Anna. 2005. "Robust Confidence Sets in the Presence of Weak Instruments." Unpublished manuscript, Harvard University.

Mikusheva, Anna and Brian P. Poi. 2006. "Tests and confidence sets with correct size in the simultaneous equations model with potentially weak instruments." *The Stata Journal* 6(3): 335-347. Working paper.

Moreira, Marcelo J. 2001. "Tests With Correct Size When Instruments Can Be Arbitrarily Weak." Working paper available online.

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica*, 71 (4), 1027-1048. Working paper version available on Moreira's website.

Morgan, Stephen L. and David J. Harding. 2006. "Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice." *Sociological Methods and Research* 35(1)3-60.

Nannicini, Tommaso. 2006. "A simulation-based sensitivity analysis for matching estimators." presented at NASUG5, working paper online.

Newey, W.K. 1987. "Efficient Estimation of Limited Dependent Variable Models with Endogeneous Explanatory Variables". *Journal of Econometrics* 36: 231-250.

Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," translated with an introduction by D. M. Dabrowska and T. P. Speed. 1990. *Statistical Science* 5(4): 465-472.

Neyman, Jerzy, K. Iwaskiewicz, and St. Kolodziejczyk. 1935. "Statistical problems in agricultural experimentation" (with discussion). *Supplement to the Journal of the Royal Statistical Society* 2(2): 107-180.

Nichols, Austin. 2006. "Weak Instruments: An Overview and New Techniques." presented at NASUG5.

Nichols, Austin. 2007. "Causal inference with observational data." *Stata Journal* 7(4): 507-541.

Nichols, Austin. 2008. "Erratum and discussion of propensity score reweighting." *Stata Journal* 8(4): 532-539.

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14(3): 693-709.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave. 1996. *Does training for the disadvantaged work? Evidence from the national JTPA Study*. Washington, DC: The Urban Institute Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Poi, Brian P. 2006. "Jackknife instrumental variables estimation in Stata." *Stata Journal* 6(3): 364-376.

Rabe-Hesketh, Sophia, Anders Skrondal, and Andrew Pickles. 2002. "Reliable estimation of generalised linear mixed models using adaptive quadrature." *Stata Journal* 2: 1-21. See also [http://gllamm.org].

Roodman, David M. 2006. "How to Do `xtabond2`: An Introduction to Difference and System GMM in Stata." CGDev WP 103 and presentation at NASUG5.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41-55.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688-701.

Sargan, J.D. 1958. "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica* 26: 393-415.

Schaffer, Mark E., and Stillman, Steven. 2006. "xtoverid: Stata module to calculate tests of overidentifying restrictions after xtivreg, xtivreg2, xthtaylor." http://ideas.repec.org/c/boc/bocode/s456779.html

Schaffer, Mark E. 2007. "xtivreg2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models." http://ideas.repec.org/c/boc/bocode/s456501.html

Schechtman, Edna and Shlomo Yitzhaki. 2001. "The Gini Instrumental Variable, or 'The Double IV Estimator'." at SSRN.

Smith, Jeffrey A., and Petra E. Todd. 2005. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125: 305-353.

Smith, Jeffrey A., and Petra E. Todd. 2001. "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods." *The American Economic Review* 91(2): 112-118.

Song, Kyungchul. 2009. "Efficient Estimation of Average Treatment Effects under Treatment-Based Sampling." PIER Working Paper 09-011.

Staiger, Douglas and James H. Stock (1997). "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 65, 557-586.

Stock, James H. and Motohiro Yogo (2005), "Testing for Weak Instruments in Linear IV Regression." Ch. 5 in J.H. Stock and D.W.K. Andrews (eds), *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, Cambridge University Press. Originally published 2001 as NBER Technical Working Paper No. 284; newer version (2004) available at Stock's website.

Stock, James H.; Jonathan H. Wright; and Motohiro Yogo. (2002). "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business and Economic Statistics*, 20, 518-529. Available from Yogo's website.

Stock, James H. and Mark W. Watson. 2006. "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression." NBER Technical Working Paper 323.

Thistlewaite, D. L., and Campbell, Donald T. (1960). "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment." *Journal of Educational Psychology* 51: 309-317.

Tukey, J. W. 1954. "Causation, regression and path analysis," in *Statistics and Mathematics in Biology*. Ames: Iowa State College Press.

Wold, Herman. 1956. "Causal inference from observational data: A review of ends and means." *Journal of the Royal Statistical Society.* A119(1): 28-61.

Wooldridge, J.M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press. Available from Stata.com.

Yun, Myeong-Su. 2004. "Decomposing Differences in the First Moment." *Economics Letters* 82(2): 275-280. See also IZA Discussion Paper 877.

Yun, Myeong-Su. 2005a. "A Simple Solution to the Identification Problem in Detailed Wage Decompositions." *Economic Inquiry* 43(4): 766-772. See also IZA Discussion Paper 836.

Yun, Myeong-Su. 2005b. "Normalized Equation and Decomposition Analysis: Computation and Inference." IZA Discussion Paper 1822.