Causal inference with observational data

Austin Nichols

December 12, 2007

1 Introduction

Identifying the causal impact of some variables X^T on y is difficult in the best of circumstances, but faces seemingly insurmountable problems in observational data, where X^T is not manipulable by the researcher and cannot be randomly assigned. Nevertheless, estimating such an impact or "treatment effect" is the goal of much research, even much research that carefully states all findings in terms of associations rather than causal effects. I will call the variables X^T the "treatment" or treatment variables, and the term simply denotes variables of interest—they need not be binary (0/1) nor have any medical or agricultural application.

Experimental research designs offer the most plausibly unbiased estimates, but experiments are frequently infeasible due to cost or moral objections—no one proposes to randomly assign smoking to individuals to assess health risks, or to randomly assign marital status to parents so as to measure the impacts on their children. Four types of quasi-experimental research designs offering approaches to causal inference using observational data are discussed below:

- ordinary regression and panel methods,
- matching and reweighting estimators,
- instrumental variables (IV) and related methods,
- and regression discontinuity (RD) designs,

in rough order of increasing internal validity (Shadish, Cook, and Campbell 2002). Each has strengths and weaknesses discussed below. In practice, the data often dictate the method, but it is incumbent upon the researcher to discuss and check (insofar as possible) the assumptions that allow causal inference with these models, and to qualify conclusions appropriately. Checking those assumptions is the focus of this paper.

A short summary of these methods and their properties is in order before we proceed. To eliminate bias, the regression and panel methods typically require confounding variables to be either measured directly or to be invariant along at least one dimension in the data, e.g. invariant over time. The matching and reweighting estimators require that selection of treatment X^T depend only on observable variables, both a stronger and weaker condition. Instrumental variables (IV) methods require extra variables that affect X^T but not outcomes directly, and throw away some information in X^T to get less efficient and biased estimates that are, however, consistent (i.e. approximately unbiased in sufficiently large samples). Regression discontinuity (RD) methods require that treatment X^T exhibit a discontinuous jump at a particular value (the "cutoff") of an observed assignment variable, and provide estimates of the effect of X^T for individuals with exactly that value of the assignment variable. To get plausibly unbiased estimates, then, one needs to either give up some efficiency or generalizability (or both, especially for IV and RD) or to make strong assumptions about the process determining X^T.

1.1 Identifying a causal effect

Consider an example to fix ideas. Suppose that for people suffering from depression, the impact of mental health treatment on work is positive. However, those who seek mental health treatment (or seek

more of it) are less likely to work, even conditional on all other observable characteristics, because their depression is more severe (in ways not measured by any data we can see). As a result, we estimate the impact of treatment on work, incorrectly, as being negative.

A classic example of an identification problem is the effect of college on earnings (Card 1999, 2001). College is surely nonrandomly assigned, and there are various important unobserved factors, including the alternatives available to individuals, their time preferences, the prices and quality of college options, academic achievement (often "ability" in economics parlance), and access to credit. Suppose college graduates earn 60 and others earn 40 on average. One simple (implausible but instructive) story might be that college has no real effect on productivity or earnings, but those who pass a test S that grants entry to college have productivity of 60 on average and go to college. Even in the absence of college, they would earn 60 if they could signal (see Spence 1973) productivity to employers via another means (e.g. by merely reporting the result of test S). In this case, extending college to a few people who failed test S would not improve their productivity at all and might not affect their earnings (if employers observed the result of test S).

If we could see the outcome for each individual case when treated and not treated (assuming a single binary treatment X^T), or an outcome y for each of the possible levels of X^T, we could calculate the treatment effect for each individual i and compute an average. Of course, this is not possible: each individual gets some level of X^T, or some history of X^T in a panel setting. So we must compare individuals i and j with different X^T to estimate an average treatment effect, but when X^T is nonrandomly assigned we have no guarantee that individuals i and j are comparable in their response to treatment, or what their outcome would have been given another X^T, even on average. The notion of "potential outcomes," due to Rubin (1974), is known as the Rubin causal model. Holland (1986) provided the classic exposition of this now dominant theoretical framework for causal inference, and Rubin (1990) clarified the debt that the Rubin causal model owes to Neyman (1923) and Fisher (1918, 1925).

In all the models discussed in this paper, we assume that the effect of treatment is on individual observations, and does not spill over onto other units. This is called the stable-unit-treatment-value assumption (SUTVA) by Rubin (1986). Often, this may be only approximately true, e.g., the effect of a college education is not only on the earnings of the recipient, since each worker participates in a labor market with other graduates and non-graduates.

What is the most common concern about observational data? If X^T is correlated with some other variable X^U that also has a causal impact on y, but we do not measure X^U, we might assess the impact of X^T as negative even though its true impact is positive. Sign reversal is an extreme case, sometimes called Simpson's paradox, though it is not a paradox and Simpson (1951) pointed out the possibility long after Yule (1903). More generally, the estimate of the impact of X^T may be biased and inconsistent when X^T is nonrandomly assigned. I.e. even if the sign of the estimated impact is not the opposite of the true impact, our estimate need not be near the true causal impact on average, nor approach it asymptotically. This central problem is usually called **omitted variable bias** or **selection bias** (here **selection** refers to the non-random selection of X^T, not selection on the dependent variable as in `heckman` and related models).

1.2 Sources of Bias and Inconsistency

The selection bias (or omitted variable bias) in an ordinary regression arises from endogeneity (a regressor is said to be **endogenous** if it is correlated with the error), a condition that also occurs if the explanatory variable is measured with error, or in a system of "simultaneous equations" (e.g. suppose work also has a causal impact on mental health, or higher earnings cause increases in education; in this case, it is not clear what impact, if any, our single-equation regressions identify).

Often a suspected type of endogeneity can be reformulated as a case of omitted variables, perhaps with a unobservable (as opposed to merely unobserved) omitted variable, about which we can nonetheless make some predictions from theory to sign the likely bias.

The formula for omitted variable bias in linear regression is instructive. With a true model

$$y = \beta_0 + X^T\beta_T + X^U\beta_U + \varepsilon$$

where we regress y on X^T but leave out X^U (for example, because we cannot observe it), the estimate

of β_T has bias

$$E(\hat{\beta}_T) - \beta_T = \delta\beta_U$$

where δ is the coefficient of an auxiliary regression of X^U on X^T (or the matrix of coefficients of stacked regressions when X^U is a matrix containing multiple variables) so the bias is proportional to the correlation of X^U and X^T and to the effect of X^U (the omitted variables) on y.

In nonlinear models, such as a `probit` or `logit` regression, the estimate will be biased and inconsistent even when X^T and X^U are uncorrelated, though Wooldridge (2002, p.471) demonstrates that some quantities of interest may still be identified under additional assumptions.

1.3 Sensitivity Testing

The wonderfully plain-English exposition by Manski (1995; 2007) 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 quasi-experimental 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, particularly for matching methods.

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

1.4 Systems of Equations

Some of the techniques discussed here to address selection bias are also used in the simultaneous equations setting. The literature on structural equations models is extensive, and a system of equations may encode a very complicated conceptual causal model, with many "causal arrows" drawn to and from many variables. The present exercise of identifying the causal impact of some limited set of variables X^T on a single outcome y can be seen as restricting our attention in such a complicated system to just one equation, and identifying just some subset of causal effects.

For example, in a simplified supply and demand system:

$$\ln Q_{supply} = e_s \ln P + a\text{TransportCost} + \varepsilon_s$$

$$\ln Q_{demand} = e_d \ln P + b\text{Income} + \varepsilon_d$$

where price (lnP) is endogenously determined by a market-clearing condition $\ln Q_{supply} = \ln Q_{demand}$, our present enterprise limits us to identifying only the demand elasticity e_d using factors that shift supply to identify exogenous shifts in price faced by consumers (exogenous relative to the second equation's error ε_s), or identifying only the supply elasticity e_s using factors that shift demand to identify exogenous shifts in price faced by firms (exogenous relative to the first equation's error ε_s).

See [R] `reg3` for alternative approaches that can simultaneously identify parameters in multiple equations, and Heckman and Vytlacil (2004) and Goldberger and Duncan (1973) for more detail.

1.5 Average Treatment Effects

In an experimental setting, typically the only two quantities to be estimated are the sample average treatment effect (SATE) or the population average treatment effect (PATE), both estimated via a difference in averages across treatment groups (equal in expectation to the mean of individual treatment effects over the full sample). In a quasi-experimental setting, several other average treatment effects are commonly estimated: the average treatment effect on the treated (ATT), the average treatment effect on the untreated (ATU) or control group (ATC), and a variety of local average treatment effects (LATE), local to some range of values or some subpopulation. In fact, one can imagine constructing at least 2^N different average treatment effect (ATE) estimates in a sample of N observations, restricting attention to two possible weights for each observation. Allowing a variety of weights and specifications leads to infinitely many LATE estimators, not all of which would be sensible.

For many decision problems, a highly relevant effect estimate is the marginal treatment effect (MTE), either the ATE for the marginal treated case—the expected treatment effect for the case that would get treatment with a small expansion of the availability of treatment—or the average effect of a small increase in a continuous treatment variable. Measures of comparable marginal treatment effects for several options can be used to decide where a marginal dollar (or metaphorical marginal dollar, including any opportunity costs and currency translations) should be spent. In other words, with finite resources, we care more about budget-neutral improvements in effectiveness than the effect of a unit increase in treatment, so we can choose among treatment options with equal cost. Quasi-experimental methods, especially IV and RD, often estimate such marginal treatment effects directly.

If the effect of a treatment X^T varies across individuals, i.e. it is not the case that $\beta_i = \beta$ for all i, then the ATE for different subpopulations will differ and we should expect different consistent estimators to converge to different quantities. This problem is larger than the selection bias issue. Even in the absence of endogenous selection of X^T (but possibly with some correlation between X_i^T and β_i, itself now properly regarded as a random variable) in a linear model, ordinary least squares (OLS) will not in general be consistent for the average over all i of individual effects β_i. Only with strong distributional assumptions can we proceed, e.g. if we assume β_i is normally distributed then the ATE may be consistently estimated by `xtmixed` or `xtrc`, or if we assume X^T is normally distributed then the ATE may be consistently estimated by OLS.

2 Regression and Panel Methods

If an omitted variable can be measured, or proxied by another variable, then an ordinary regression may yield an unbiased estimate. In the case where ordinary least squares is unbiased, or at least consistent, it would nearly always be preferred, since it produces the most efficient estimates (ignoring issues around weights or non-independent errors). The measurement error entailed in a proxy for an unobservable could actually exacerbate bias, rather than reducing it, however. In addition, one is usually concerned that cases with differing X^T may differ in other ways even conditional on all other observables X^C ("control" variables). Nonetheless, a sequence of ordinary regressions that add or drop variables can be instructive as to the nature of various forms of omitted variable bias in the data at hand.

A complete discussion of panel methods would not fit in any one book, much less this article. The idea can be illuminated with one short example using linear regression, however.

Suppose a model of the form

$$y = \beta_0 + X^T\beta_T + X^U\beta_U + \varepsilon$$

where we do not observe X^U. Suppose the omitted variables X^U vary only across groups, where group membership is indexed by i, so a representative observation can be written:

$$y_{it} = \beta_0 + X_{it}^T\beta_T + u_i + \varepsilon_{it}$$

where $u_i = X_i^U\beta_U$. Then we can eliminate the bias arising from omission of X^U by differencing:

$$y_{it} - y_{is} = (X_{it}^T - X_{is}^T)\beta_T + (\varepsilon_{it} - \varepsilon_{is})$$

using various definitions of s.

The idea of using panel methods to identify a causal impact is to use an individual i as its own control group, by including information from multiple points in time. In fact, the second dimension of the data indexed by t need not be time, but it is a convenient viewpoint.

A fixed-effect (FE) model such as `xtreg, fe i(id)` or `areg, a(id)` effectively subtracts off the within-i mean values of each variable, so for example $X_{is}^T = \bar{X^T}_i = \frac{1}{N_i}\sum_{s=1}^{N_i} X_{is}^T$, and the model

$$y_{it} - \bar{y}_i = (X_{it}^T - \bar{X^T}_i)\beta_T + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

can be estimated via OLS. This is also called the "within estimator" and is equivalent to a regression that includes an indicator variable for each group i, allowing for a different intercept term for each group.

An alternative to the FE model is to use the first difference (FD), i.e. $s = (t-1)$ or

$$y_{it} - y_{i(t-1)} = (X_{it}^T - X_{i(t-1)}^T)\beta_T + (\varepsilon_{it} - \varepsilon_{i(t-1)})$$

which is just `reg d.y d.x` in `tsset` data, or `xtivreg2 y x, fd` (Schaffer and Stillman 2007), which offers additional standard error (SE) corrections beyond `cluster` and `robust`.

A third option is to use the long difference (LD), keeping only two observations per group. For a balanced panel, if $t = b$ is the last observation and $t = a$ is the first, the model is:

$$y_{ib} - y_{ia} = (X_{ib}^T - X_{ia}^T)\beta_T + (\varepsilon_{ib} - \varepsilon_{ia})$$

producing only one observation per group (the difference of the first and last observations).

Figure 1 shows the interpretation of these three types of estimates by showing a single panel's contribution to the estimated effect of an indicator variable that is one for all $t > 3$ (t in $0, ..., 10$) and zero elsewhere, e.g. a policy that comes into effect at some point in time. The FE estimate compares the mean outcomes before and after; the FD estimate compares the outcome just prior to and just after the change in policy; and the LD estimate compares outcomes well before and well after the change in policy.

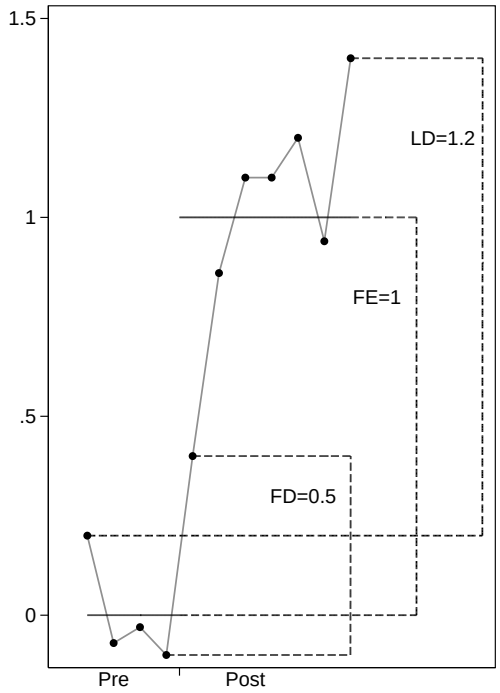

Figure 1: A Single Panel's Contributions to Various Estimates

Clearly, one must impose some assumptions on the speed with which X^T 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.

In panel models, one must usually think carefully about within-panel trends (note that we cannot usually obtain consistent estimates of within-panel trends for the same reason we cannot usually obtain consistent estimates of fixed effects; the number of parameters increases linearly in the number of panels

N) and the frequency of measurement. Baum (2006) discussed some filtering techniques to get different frequency "signals" from noisy data. A very simple method used in Baker, Benjamin, and Stanger (1999) is often attractive, as it offers a easy way to decompose any variable X_t into two orthogonal components: a high-frequency component $(X_t - X_{t-1})/2$ and a low-frequency component $(X_t + X_{t-1})/2$ that together sum to X_t.

A facile example of all three (FE, FD, and LD) is

```
webuse grunfeld, clear
xtreg inv ks, fe cl(com)
reg d.inv d.ks, cl(com)
su time, meanonly
gen t=time if time==r(min) | time==r(max)
tsset com t
reg d.inv d.ks, cl(com)
```

Clearly, different assumptions about the error process apply in each case, in addition to assumptions about the speed with which X^T affects y, and the FD and LD models require an ordered t index (such as time). The `cluster` option used above should be considered nearly *de rigeur* in panel models, to allow for errors that may be correlated within group, and not identically distributed across groups. The performance of the cluster-robust estimator is quite good with 50 or more clusters, or fewer if the clusters are large and balanced (Nichols and Schaffer 2007). In the LD case, the `cluster` option is equivalent to the `robust` option, since each group is represented by one observation.

Having eliminated bias due to unobservable heterogeneity across i units, it is often tempting to difference or demean again. For example, it is common to include indicator variables for t in fixed-effects models, for example:

```
webuse grunfeld, clear
qui tab year, gen(d)
xtreg inv ks d*, fe cl(com)
```

thus creating a two-way fixed-effects model. If individuals i are observed in different settings j, for example students who attend various schools or workers who reside in various locales over time, we can also include indicator variables for j in a fixed-effects model. Thus we can consider various n-way fixed-effects models, though models with large numbers of dimensions for fixed effects may rapidly become unstable or computationally challenging to estimate.

The LD, FD, and FE estimators use none of the cross-sectional differences across groups (individuals) i, which can lead to lower efficiency (relative to an estimator that exploits cross-sectional variation), and they also drop any variables that do not vary over t within i, so the coefficients on some variables of interest may not be estimated with these methods.

The random-effects estimator (RE) available from `xtreg` exploits cross-sectional variation and reports coefficients on variables that do not vary over t within i, but requires strong assumptions about error terms that are often violated in practice. In particular, for RE to be unbiased in situations where FE is unbiased, we must assume that u_{it} is uncorrelated with ${X^T}_{it}$ (which contradicts our starting point above, where we worried about a X^U correlated with X^T). There is no direct test of this assumption about an unobservable disturbance term, but `hausman` and `xtoverid` (Schaffer and Stillman 2006) offer a test that the coefficients estimated in both the RE and FE models are the same:

```
webuse grunfeld, clear
egen ik=max(ks*(year==1935)), by(com)
xtreg inv ks ik, re cl(com)
xtoverid
```

where a rejection casts doubt on whether RE is unbiased when FE is.

Various other [XT] entries, such as [XT] `xtmixed` and [XT] `xthtaylor`, offer a variety of other panel methods that generally make further assumptions about the distribution of disturbances and sources of endogeneity. Typically, there is a tradeoff between improved efficiency bought by making

assumptions about the data-generating process versus robustness to various violations of assumptions. See also Griliches and Hausman (1986) for additional considerations related to all of the above panel methods. Rothstein (2007) offers a useful applied examination of identifying assumptions in FE models and correlated random effects models.

Generally, panel methods eliminate the bias due to some unobserved factors and not others. Considering the FE, FD, and LD models, it is often hard to believe that all of the selection on unobservables is due to time-invariant factors.

3 Matching Estimators

In the case of a single discrete set of treatments X^T we wish to compare means or proportions much as we would in an experimental setting. We may be able to include indicators and interactions for factors (in X^C) that affect selection into the treatment group (defined by $X^T = 1$, say), to estimate the impact of treatment within groups of identical X^C using a fully saturated regression. There are also matching estimators (Cochran and Rubin 1973; Stuart and Rubin 2007) which compare observations with like X^C by pairing observations that are close by some metric (see also Imai and van Dyk, 2004, for generalizations of matching estimators). A set of alternative approaches involve reweighting so the joint or marginal distributions of X^C are identical for different groups.

Matching or reweighting approaches can give consistent estimates of a huge variety of average treatment effects, but only under the assumptions that the selection process depends on observables, and that the model used to match or reweight is a good one. In some sense, we push the problems associated with observational data from estimating the effect of X^T on y down onto estimating the effect of X^C on X^T. For this reason, 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.

3.1 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. The Stata implementation `nnmatch` was described by Abadie et al. (2004). Imbens (2004) covered details of Nearest Neighbor Matching methods. The downside to Nearest Neighbor Matching is that it can be computationally intensive, and bootstrapped standard errors are infeasible owing to the discontinuous nature of matching (Abadie and Imbens, 2006).

3.2 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. Stuart and Rubin (2007) offer a more formal but equally helpful discussion of best practices in matching.

Typically, one treatment case is matched to several control cases, but one-to-one matching is also common, and may be preferred (Glazerman, Levy, and Myers 2003). One Stata implementation `psmatch2` (Leuven and Sianesi, 2003) 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.

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. In practice, it is rarely the case that the ranges of estimated propensity scores are the same, but they do nearly always overlap, and generalizations about treatment effects should probably be 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. Generally, we need to use a kernel density estimator like `kdensity` to obtain smooth estimated densities of the propensity score for both treatment and control groups, but then areas of zero density will have positive density estimates. Thus some small value f_0 is redefined to be effectively zero, and the smallest connected range of estimated propensity scores $\widehat{\lambda}$ with $\widehat{f}(\widehat{\lambda}) \geq f_0$ for both treatment and control groups is used in the analysis, and observations outside this range are discarded.

Regardless of whether the estimation or extrapolation of estimates is limited to a range of propensities or ranges of X^C 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`.

3.3 Sensitivity Testing

Matching estimators have perhaps the most detailed literature on formal sensitivity testing. Rosenbaum (2002) provides a comprehensive treatment of formal sensitivity testing. Rosenbaum bounds on treatment effects may be constructed using `psmatch2` and `rbounds`, a user-written command 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 for sensitivity testing in matching models.

3.4 Reweighting

The propensity score can also be used to reweight treatment and control groups so the distribution of X^C looks the same in both groups. The basic idea is to use a `probit` or `logit` regression of treatment on X^C to estimate the conditional probability $\hat{\lambda}$ of being in the treatment group and to use the odds $\hat{\lambda}/(1-\hat{\lambda})$ as a weight. This is like inverting the test of randomization used in experimental designs to make the group status look as if it were randomly assigned.

As Morgan and Harding (2006) point out, all the matching estimators can also be thought of various reweighting schemes 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.

The reweighting approach leads to a whole class of weighted least squares estimators, and is connected to techniques described by DiNardo, Fortin, and Lemieux (1996), Autor, Katz, and Kearney (2005), Liebbrandt, Levinsohn, and McCrary (2005), and Machado and Mata (2005), and these techniques are related to various decomposition techniques in Blinder (1973), Oaxaca (1973), Yun (2004, 2005ab), Gomulka and Stern (1990), and Juhn, Murphy, and Pierce (1991, 1993). DiNardo (2002) usefully outlines some connections between propensity score methods and the decomposition techniques.

The `dfl` (Azevedo 2005), `oaxaca` (Jann 2005b), and `jmpierce` (Jann 2005a) programs on SSC offer Stata implementations for some of the latter. The decomposition techniques seek to attribute observed differences in an outcome y both to differences in X^C variables and differences in the associations between X^C variables and y, and they are most useful for comparing two distributions where the binary variable defining which group an observation belongs to is properly considered exogenous, e.g. sex or calendar year. See also Rubin (1986).

The reweighting approach is particularly useful in combining matching-type estimators with other methods, e.g. fixed-effects regression. After constructing weights $w = \hat{\lambda}/(1-\hat{\lambda})$ (or the product of

weights $w = w_0\hat{\lambda}/(1-\hat{\lambda})$ where w_0 is an existing weight on the data used in the construction of $\hat{\lambda}$) that equalize the distributions of X^C, other commands can be run on the reweighted data, e.g. `areg` for a fixed-effect estimator.

3.5 Examples

Imagine the outcome is wage and the treatment variable is union membership. One can imagine reweighting union members to have distributions of education, age, race/ethnicity, and other job and demographic characteristics equivalent to nonunion workers (or a subset of nonunion workers). One could compare otherwise identical persons within occupation and industry cells using a regression approach or `nnmatch` with exact matching on some characteristics. An example comparing several regressions with propensity score matching is:

```
webuse nlswork, clear
xi i.race i.ind i.occ
loc x "union coll age ten not_s c_city south nev_m _I*"
reg ln_w union
reg ln_w `x'
g u=uniform()
sort u
psmatch2 `x', out(ln_w) ate
tw kdensity _ps if _tr || kdensity _ps if !_tr
g w=_ps/(1-_ps)
reg ln_w `x' [pw=w] if _ps<.3
reg ln_w `x' [pw=w]
```

where the estimated union wage premium is about 13% in a regression, but about 15% in the matching estimate of the average benefit to union workers (the ATT) and about 10% on average for everyone (the ATE). The reweighted regressions give different estimates: for the more than 70% of individuals who are unlikely to be unionized (propensity under 30%), the wage premium is about 9%, and for the full sample, it is about 18%.

Arguably none of these estimates of wage premia correspond to a readily specified thought experiment, such as "what is the estimated effect on wages of being in a union for a randomly chosen individual?" (the ATE) or "what is the estimated effect on wages of being in a union for an individual just on the margin of being in a union or not?" (the LATE). DiNardo and Lee (2002) offer a much more convincing set of causal estimates of the LATE using an regression discontinuity (RD) design (see below).

We could also have estimated the wage premium of a college education by simply leaving switching `coll` and `union` in the above syntax (to find a wage premium of 25% in a regression or 27% using `psmatch2`). We could use data from Card (1995ab) on education and wages:

```
use http://fmwww.bc.edu/ec-p/data/wooldridge/card, clear
g byte coll=educ>15
loc x "coll age exper* smsa* south mar black reg662-reg669"
reg lw `x'
psmatch2 `x', out(lw) ate
```

to find a college wage premium of 29% using a regression or 30% using `psmatch2`. We return to this example in the next section using an Instrumental Variables method.

4 Instrumental Variables

An alternative to panel methods and matching estimators is to find another set of variables Z correlated with X^T but not correlated with the error term, e.g. e in

$$y = X^T\beta_T + X^C\beta_C + e$$

so Z must satisfy $E(Z'e) = 0$ and $E(Z'X^T) \neq 0$. The variables Z are called **excluded instruments**, and a class of Instrumental Variables (IV) methods can then be used to consistently estimate an impact of X^T on y.

Various interpretations of the IV estimate have been advanced, typically as the Local Average Treatment Effect, or **LATE** (Angrist, Imbens and Rubin; 1996), meaning the effect of X^T on y for those who are induced by their level of Z to have higher X^T. For the college graduate example, this might be the average gain $E_i[y_i(t) - y_i(0)]$ over all those i in the treatment group with $Z = 1$ (where Z might be "lived close to a college" or "received a Pell grant"), arising from an increase from $X^T = 0$ to $X^T = t$ in treatment, i.e. the wage premium due to college averaged over those who were induced to go to college by Z.

The instrumental variables estimators are in general only as good as the excluded instruments used, so naturally criticisms of the predictors in a standard regression model become criticisms of the excluded instruments in an IV model.

Also, the IV estimators are still biased, but consistent, and are much less efficient than OLS, so failure to reject the null should not be taken as acceptance of the alternative. That is, one should never compare the IV estimate only to a zero effect, but also to other plausible values, including the OLS estimate. Some other common pitfalls discussed below include improper exclusion restrictions (addressed via overidentification tests) and weak identification (addressed via diagnostics and robust inference).

It is worth repeating that IV estimators are biased in finite samples, and are only justified for very large samples. Nelson and Startz (1990) showed how strange the finite sample behavior of an IV estimator can be, and Bound, Jaeger, and Baker (1995) showed that even very large samples of millions of observations are insufficient for asymptotic justifications to apply in the presence of weak instruments (see also Stock and Yogo 2005).

4.1 Key assumptions

Because IV can lead one astray if any of the assumptions is violated, anyone using an IV estimator should conduct and report tests of:

- instrument validity (overidentification or **overid** tests),
- endogeneity,
- identification,
- presence of weak instruments,
- and misspecification of functional form (RESET).

Further discussion and suggestions on what to do when a test is failed appear in the relevant sections below.

4.2 Forms of IV

The standard IV estimator in a model

$$y = X^T\beta_T + X^C\beta_C + e$$

where we have Z satisfying $E(Z'e) = 0$ and $E(Z'X^T) \neq 0$ is

$$\hat{\beta}^{IV} = \begin{pmatrix} \hat{\beta}_T^{IV} \\ \\ \hat{\beta}_C^{IV} \end{pmatrix} = (X'P_ZX)^{-1}X'P_Zy$$

(ignoring weights), where $X = (X^TX^C)$ and P_Z is the projection matrix $Z_a(Z_a'Z_a)^{-1}Z_a'$ with $Z_a = (ZX^C)$. The idea is, we use the component of X^T along Z, which is exogenous, as the only source of variation in X^T that we use to estimate the effect on y.

These estimates are easily obtained in Stata 6 through 9 with the syntax `ivreg y xc* (xt* = z*)`, where `xc*` are all exogenous "included instruments" X^C and `xt*` are endogenous variables X^T. In Stata 10, the syntax is `ivregress 2sls y xc* (xt* = z*)`. For Stata 9 and above, the `ivreg2` command of Baum, Schaffer, and Stillman (2007) would be written

```
ivreg2 y xc* (xt* = z*)
```

Example data for trying out these commands can be easily generated, e.g.

```
use http://fmwww.bc.edu/ec-p/data/wooldridge/card, clear
ren lw y
ren nearc4 z
ren educ xt
ren exper xc
```

The standard IV estimator is equivalent to 2 or 3 forms of two-stage estimators. The first, which gave rise to the moniker **Two-Stage Least Squares** (2SLS), has you regress X^T on X^C and Z and predict $\hat{X}^T$, then regress y on $\hat{X}^T$ and X^C, and the coefficient on $\hat{X}^T$ is $\hat{\beta}_T^{IV}$, so:

```
foreach xt of varlist xt* {
 reg `xt' xc* z*
 predict `xt'_hat
 }
reg y xt*_hat xc*
```

will give the same estimates as the above IV commands, but the reported standard errors will be wrong as Stata will use $\hat{X}^T$ rather than X^T to compute them. Even though IV is not implemented in these two stages, the conceptual model of these first-stage and second-stage regressions is pervasive, and the properties of said first-stage regressions are central to the section on identification and weak instruments below.

The second two-stage estimator that generates identical estimates is a **control-function approach**. Regress each variable in X^T on the other variables in X^T and X^C and Z to predict the errors $\hat{v}_T = X^T - \hat{X}^T$, then regress y on X^T, $\hat{v}_T$, and X^C, and you will find that the coefficient on X^T is $\hat{\beta}_T^{IV}$, and tests of significance on each $\hat{v}_T$ are tests of endogeneity of each X^T. Thus:

```
cap drop *_hat
unab xt : xt*
foreach v of loc xt {
 loc otht: list xt-v
 reg `v' xc* z* `otht'
 predict v_`xt', resid
 }
reg y xt* xc* v_*
```

will give the IV estimates, though again the standard errors will be wrong. However, the tests of endogeneity (given by the reported p-values on variables `v_*` above) will be correct. A similar approach works for nonlinear models such as `probit` or `poisson` (`help ivprobit` and `findit ivpois` for relevant commands). The tests of endogeneity in nonlinear models given by the control-function approach are also quite robust (see e.g. Wooldridge, 2002, p.474 or p.665).

The third two-stage version of the IV strategy, which applies in the case of one endogenous variable and one excluded instrument, is sometimes called the **Wald estimator**. First regress X^T on X^C and Z (let $\hat{\pi}$ be the estimated coefficient on Z), then regress y on Z and X^C, (let $\hat{\gamma}$ be the estimated coefficient on Z), and the ratio of coefficients on Z ($\hat{\gamma}/\hat{\pi}$) is $\hat{\beta}_{IV}$, so:

```
reg xt z xc*
loc p=_b[z]
reg y z xc*
loc g=_b[z]
di `g'/`p'
```

will give the same estimate as the IV command `ivreg2 y xc* (xt=z)`. The regression of y on Z and X^C is sometimes called the **reduced form regression**, though this name is often applied to other regressions, so I will avoid using the term.

The generalized method of moments (GMM), limited-information maximum likelihood (LIML), and continuously-updated GMM estimation (CUE) forms of IV are discussed at length in Baum, Schaffer, and Stillman (2007) and various implementations are available from the `ivregress` and `ivreg2` commands. Some forms of IV may be expressed as k-class estimation, available from `ivreg2`, and there are many other forms of IV models, including official Stata commands such as `ivprobit`, `treatreg`, and `ivtobit`, and user-written additions such as `qvf` (Hardin, Schmiediche, and Carroll, 2003), `jive` (Poi, 2006), and `ivpois` (on SSC).

4.3 Finding Excluded Instruments

The hard part of IV is finding a suitable Z matrix. The excluded instruments in Z have to be strongly correlated with the endogenous X^T and uncorrelated with the unobservable error e. But note that the problem we wish to solve is that the endogenous X^T is correlated with the unobservable error e. A good story is the crucial element in any plausible IV specification. We the readers need to believe that Z is strongly correlated with the endogenous X^T but has no direct impact on y (is uncorrelated with the unobservable error e), since the assumptions are not directly testable. However, the tests discussed in the following sections can help buttress a convincing story, and should be reported in any case.

In general, specification search in the first stage regressions of X^T on some Z does not bias estimates or inference, nor does using generated regressors. However, it is easy to produce counterexamples to this general rule. For example, taking $Z = X^T + \nu$ where ν is a small random error will produce strong identification diagnostics, and might pass overidentification tests described in the next section, but will not improve estimates (and could lead to substantially less accurate inference).

If some Z are weak instruments, then regressing X^T on Z to get $\hat{X}^T$ and using $\hat{X}^T$ as the excluded instruments in an IV regression of y on X^T and X^C will likewise produce strong identification diagnostics, but will not improve estimates or inference. Hall, Rudebusch, and Wilcox (1996) reported that choosing instruments based on measures of the strength of identification could actually increase bias and size distortions.

4.4 Exclusion Restrictions in IV

The exclusion restrictions $E(Z'e) = 0$ cannot be directly tested, but if there are more excluded instruments than endogenous regressors, an overidentification (**overid**) test is feasible and the result should be reported. If there as exactly as many excluded instruments as endogenous regressors, the equation is **exactly identified**, and no overid test is feasible.

However, if Z is truly exogenous, it is likely also true that $E(W'e) = 0$, where W contains Z and squares and cross products of Z. Thus there is always a feasible overid test using an augmented set of excluded instruments (though $E(W'e) = 0$ is a stronger condition than $E(Z'e) = 0$, only in the case where we reject the overid null for each element of W not in Z do we not gain anything by augmenting the set of excluded instruments). For example, if you have two good excluded instruments, you might multiply them together, and square each, to produce five excluded instruments. Testing the three extra overidentification restrictions is like a RESET test of excluded instruments. In addition, interactions of Z and X^C may be good candidates for excluded instruments. For reasons discussed below, adding excluded instruments willy-nilly is a bad idea, and with many weak instruments, LIML or CUE is preferred to standard IV/2SLS.

Baum, Schaffer, and Stillman (2007) discuss the implementation of overid tests in `ivreg2` (see also `overid` from Baum, Schaffer, Stillman, and Wiggins, 2006). Passing the overid test (i.e. failing to reject the null of zero correlation) is neither necessary nor sufficient for instrument validity ($E(Z'e) = 0$), but rejecting the null in an overid test should lead you to reconsider your IV strategy, and perhaps to look for different excluded instruments.

4.5 Tests of Endogeneity

Even if we have an excluded instrument that satisfies $E(Z'e) = 0$ there is no guarantee that $E(X^{T'}\varepsilon) \neq 0$ as we have been assuming, and if $E(X^{T'}\varepsilon) = 0$ we prefer ordinary regression to IV. So we should test the null that $E(X^{T'}\varepsilon) = 0$ (a test of endogeneity), though this test requires instrument validity ($E(Z'e) = 0$), so it should follow any feasible overid tests.

Baum, Schaffer, and Stillman (2007) describe several methods to test the endogeneity of a variable in X^T, including the `endog` option of `ivreg2` and the standalone `ivendog` command (both available from SSC, with excellent help files). Section 4.2 above also shows how the control function form of IV can be used to test endogeneity of a variable in X^T.

4.6 Identification and Weak Instruments

Even if we have an excluded instrument that satisfies $E(Z'e) = 0$ there is no guarantee that $E(Z'X^T) \neq 0$. This is the second of the two crucial assumptions, and presents problems of various sizes in almost all IV specifications. The extent to which $E(Z'X^T) \neq 0$ determined the strength of identification. Baum, Schaffer, and Stillman (2007) describe tests of identification, which amount to tests of the rank of $E(Z'X^T)$. These rank tests address the concern that a number of excluded instruments may generate exogenous variation in one endogenous variable and be uncorrelated with another endogenous variable, so the equation is not identified even though it satisfies the order condition (the number of excluded instruments is at least as great as the number of endogenous variables). For example, if we have two endogenous variables X_1 and X_2 and three excluded instruments, all three excluded instruments may be correlated with X_1 and not with X_2. The identification tests look at the least partial correlation, or the minimum eigenvalue of the Cragg-Donald statistic (Stock and Yogo 2005), for example, and measures of whether at least one endogenous variable has no correlation with the excluded instruments.

Even if we reject the null of underidentification and conclude $E(Z'X^T) \neq 0$, we can still face a "weak instruments" problem if some elements of $E(Z'X^T)$ are close to zero. The IV estimate is always biased, but is less biased than OLS to the extent that identification is strong. In the limit of weak instruments, there would be no improvement over OLS in terms of bias and the bias would be 100% of OLS, and in the other limit, the bias would be zero percent of the OLS bias (though this would require that the correlation between X^T and Z be perfect, which is impossible since X^T is endogenous and Z is exogenous). In applications, you'd like to know where you are on that spectrum, even if only approximately.

There is also a distortion in the size of hypothesis tests. If you believe you are incorrectly rejecting a null hypothesis about 5% of the time (i.e. you have chosen a size $\alpha = 0.05$), you may actually face a size of 10% or 20% or more.

Stock and Yogo (2005) reported rule-of-thumb critical values to measure the extent of both of these problems. Their Table 1 shows the value of a statistic measuring the predictive power of the excluded instruments that will imply a limit of the bias to some percentage of OLS. For two endogenous variables and three excluded instruments (n=2, $K_2 = 5$) the minimum value to limit the bias to 20% of OLS is 5.91. `ivreg2` reports these values as **Stock-Yogo weak ID test critical values**: one set for various percentages of "maximal IV relative bias" (largest bias relative to OLS) and one set for "maximal IV size" (the largest size of a nominal 5% test).

The key point is that all IV and IV-type specifications can suffer from bias and size distortions, not to mention inefficiency and sometimes failures of exclusion restrictions. The Stock and Yogo (2005) approach measures how strong identification is in your sample, and `ranktest` (Kleibergen and Schaffer 2007) offers a way forward for cases where errors are not assumed to be independently and identically distributed, but neither provides solutions in the event that weak instruments appear to be a problem. A further limitation is that these identification statistics only apply to the linear case, not the nonlinear analogs, including those estimated via generalized linear models (GLM). In practice, researchers should report the identification statistics for the closest linear analog, i.e. run `ivreg2` and report the output alongside the output from `treatreg` or `qvf` or the like.

If you suspect weak instruments may be producing large bias or size distortions, you have several options. You can find better excluded instruments, possibly simply by transforming your existing in-

struments. You can use LIML or CUE which are more robust to many weak instruments than standard IV. Perhaps best of all, you can conduct inference that is robust to weak instruments; with one endogenous variable, use `condivreg` (Mikusheva and Poi 2006), or with more than one, use tests described by Anderson and Rubin (1949) and Baum, Schaffer, and Stillman (2007, sections 7.4 and 8).

4.7 Functional Form Tests in IV

As Baum, Schaffer, and Stillman (2007, section 9) and Wooldridge (2002, p.125) discuss, the RESET test regressing residuals on predicted y and powers thereof is properly a test of a linearity assumption, or a test of functional form restrictions. `ivreset` performs the IV version of the test in Stata. A more informative specification check is the graphical version of RESET: predict $\hat{X}^T$ after the first stage regressions, then compute forecasts $\hat{y} = X^T\hat{\beta}_T^{IV} + X^C\hat{\beta}_C$ and $\hat{y}_f = \hat{X}^T\hat{\beta}_T^{IV} + X^C\hat{\beta}_C$ and graph a `scatter` of the residuals $\hat{\varepsilon} = y - \hat{y}$ against $\hat{y}_f$.

```
ivreg2 y xc* (xt = z*), first
quietly regress xt xc* z*
predict xhat
quietly regress y xhat xc*
predict ehat, resid
predict yhatf
drop xhat
ren xt xhat
predict yhat
scatter ehat yhatf, name(f)
scatter ehat yhat, name(y)
scatter ehat xhat, name(x)
```

Any unmodeled nonlinearities may be apparent as a pattern in the scatterplot.

4.8 Standard Errors in IV

The largest issue in IV estimation is often that the variance of the estimator is much larger than ordinary regression. Just as with ordinary regression, the standard errors are asymptotically valid for inference under the restrictive assumptions that the disturbances are independently and identically distributed. Getting standard errors robust to various violations of these assumptions is easily accomplished using the `ivreg2` command described in Baum, Schaffer, and Stillman (2007). Many other commands estimating IV models offer no equivalent robust SE estimates, but it may be possible to assess the size and direction of SE corrections using the nearest linear analog, in the spirit of using estimated design effects in the survey regression context.

4.9 Inference in IV

Assuming we have computed consistent standard errors, and the best IV estimate we can, using a good set of Z and X^C variables, there remains the question of how we interpret the estimates and tests. Typically, IV identifies a particular LATE, namely the effect of an increase in X^T due to an increase in Z. If X^T were college and Z an exogenous source of financial aid, then the IV estimate of the effect of X^T on wages would be the college wage premium for those who were induced to attend college by being eligible for the marginally more generous aid package.

Angrist and Krueger (1991) estimated the effect of education on earnings using compulsory schooling laws as a justification for quarter of birth as Z. Even if the critiques of Bound, Jaeger, and Baker (1995) did not apply, the identified effect would be for an increase in education due to being forced to remain in school a few months more. That is, the measured wage effect of an additional year of education is for the eleventh grade, roughly, and only for those who would have dropped out if not for compulsory schooling laws.

Sometimes, a LATE of this form is exactly the estimate desired, but if we cannot reject that the IV estimate differs from the OLS estimate, or the IV confidence region includes the OLS confidence region, we may not have improved estimates, but merely produced noisier ones. Only in the case where the IV estimate differs can we hope to ascertain the nature of selection bias.

4.10 Examples

We can use the data from Card (1995ab) to estimate the impact of education on wages, where nearness to a college is posited as a source of exogenous variation in educational attainment:

```
use http://fmwww.bc.edu/ec-p/data/wooldridge/card, clear
loc x "exper* smsa* south mar black reg662-reg669"
reg lw educ `x'
ivreg2 lw `x' (educ=nearc2 nearc4), first endog(educ)
ivreg2 lw `x' (educ=nearc2 nearc4), gmm
ivreg2 lw `x' (educ=nearc2 nearc4), liml
```

to find the return to an additional year of education is about 7% using ordinary regression or 16% or 17% using IV methods. The Sargan statistic fails to reject that excluded instruments are valid, the test of endogeneity is marginally significant (giving different results at the 95% and 90% levels), and the Anderson-Rubin and Stock-Wright tests of identification strongly reject that the model is underidentified.

The test for weak instruments is the F-test on the excluded instruments in the first-stage regression, which at 7.49 with a p-value of 0.0006 seems to indicate that the excluded instruments influence educational attainment, but the size of Wald tests on `educ` that we specify as 5% might be roughly 25%. To construct an Anderson-Rubin confidence interval, we can

```
g y=.
foreach beta in .069 .0695 .07 .36 .365 .37 {
 qui replace y=lw-`beta'*educ
 qui reg y `x' nearc2 nearc4
 di as res "Test of beta=" `beta'
 test  nearc2 nearc4
 }
```

giving a confidence interval of (.07,.37); see also Nichols (2006, p.18) and Baum, Schaffer, and Stillman (2007, p.30). Thus, the IV confidence region includes the OLS estimate, and nearly includes the OLS confidence interval, so the evidence on selection bias is very weak. Still, if we accept the exclusion restrictions as valid, the evidence does not support a story where omitting ability (causing both increased wages and increased education) leads to positive bias. If anything, the bias seems likely to be negative, perhaps due to unobserved heterogeneity in discount rates or credit market failures. In the latter case, the omitted factor may be a social or economic disadvantage observable by lenders.

A similar set of conclusions apply if we model the education response as a binary treatment **college**:

```
g byte coll=educ>15
reg lw coll `x'
treatreg lw `x', treat(coll=nearc2 nearc4)
ivreg2 lw `x' (coll=nearc2 nearc4), first endog(coll)
ivreg2 lw `x' (coll=nearc2 nearc4), gmm
ivreg2 lw `x' (coll=nearc2 nearc4), liml
```

These regressions also indicate that the OLS estimate may be biased downward, though the OLS confidence interval is contained in the `treatreg` confidence interval and each of the IV intervals, so we cannot conclude much with confidence.

5 Regression Discontinuity Designs

The idea of the regression discontinuity (RD) design is to exploit an observable discontinuity in the level of treatment related to an assignment variable Z, so the level of treatment X^T jumps discontinuously at some value of Z, the "cutoff." Let Z_0 denote the cutoff. In the neighborhood of Z_0, under some often plausible assumptions, a discontinuous jump in the outcome y can be attributed to the change in the level of treatment. Near Z_0, the level of treatment can be treated *as if* it is randomly assigned. For this reason, the RD design is generally regarded as having the greatest internal validity of the quasi-experimental estimators.

Examples include share of votes received in a US Congressional election by the Democratic candidate as Z, which induces a clear discontinuity in X^T, the probability of a Democrat occupying office the following term, and X^T may affect various outcomes y, if Democratic and Republican candidates actually differ in close races (Lee 2001). DiNardo and Lee (2002) use the share of votes received for a union as Z, and unions may may affect the survival of a firm (but do not seem to). They point out that the union wage premium y can be consistently estimated only if survival is not affected (no differential attrition around Z_0), and they find negligibly small effects of unions on wages.

The standard treatment of RD is Hahn, Todd, and van der Klaauw (2001), who clarified the link to IV methods. Recent working papers by Imbens and Lemieux (2007) and McCrary (2007) focused on some important practical issues related to RD designs. Both papers are forthcoming in the *Journal of Econometrics* along with other useful papers on RD.

Many authors stress a distinction between "sharp" and "fuzzy" RD. In sharp RD designs, the level of treatment rises from zero to one at Z_0, as in the case where treatment is having a Democratic representative in the US Congress, or establishing a union, and a winning vote share defines Z_0. In fuzzy RD designs, the level of treatment increases discontinuously, or the probability of treatment increases discontinuously, but not from zero to one, so we may want to deflate by the expected increase in X^T at Z_0 in constructing our estimate of the causal impact of a one-unit change in X^T.

In sharp RD designs, the jump in y at Z_0 is the estimate of the causal impact of X^T. In a fuzzy RD design, the jump in y divided by the jump in X^T at Z_0 is the local Wald estimate (equivalent to a local IV estimate) of the causal impact. Note that the local Wald estimate reduces to the jump in y at Z_0 in a sharp RD design as the jump in X^T is one, so the distinction between fuzzy and sharp RD is not that sharp.

Some authors, e.g. Shadish, Cook, and Campbell (2002, 229), seem to characterize as fuzzy RD a wider class of problems where the cutoff itself may not be sharply defined, but without a true discontinuity, there can be no RD. The fuzziness in fuzzy RD arises only from probabilistic assignment of X^T in the neighborhood of Z_0.

5.1 Key assumptions and tests

The assumptions that allow us to infer a causal effect on y due to an abrupt change in X^T at Z_0 are that the change in X^T at Z_0 is truly discontinuous, Z is observed without error (Lee and Card 2007), y is a continuous function of Z at Z_0 in the absence of treatment (for individuals), and that individuals are not sorted across Z_0 in their responsiveness to treatment. None of these assumptions can be directly tested, but there are diagnostic tests that should always be employed.

The first is to test the null that no discontinuity in treatment occurs at Z_0, since without identifying a jump in X^T we will be unable to claim to identify the causal impact of said jump. The second is to test that there are no other extraneous discontinuities in X^T or y away from Z_0, since this would call into question whether the functions would be smooth through Z_0 in the absence of treatment. The third and fourth test that predetermined characteristics and the density of Z exhibit no jump at Z_0, since these call into question the exchangeability of observations on either side of Z_0. Then the estimate itself usually supplies a test that the treatment effect is nonzero (y jumps at Z_0 because X^T jumps at Z_0).

Abusing notation somewhat so that Δ is an estimate of the discontinuous jump in a variable, we can enumerate these tests as:

- (T1) $\Delta X^T(Z_0) \neq 0$

- (T2) $\Delta X^T(Z \neq Z_0) = 0$ and $\Delta y(Z \neq Z_0) = 0$
- (T3) $\Delta X^C(Z_0) = 0$
- (T4) $\Delta f(Z_0) = 0$
- (T5) $\Delta y(Z_0) \neq 0$ or $\left(\frac{\Delta y(Z_0)}{\Delta X^T(Z_0)}\right) \neq 0$

5.2 Methodological choices

Estimating the size of a discontinuous jump can be accomplished by comparing means in small bins of Z to the left and right of Z_0, or via a regression with various powers of Z, an indicator D for $Z > Z_0$, and interactions of all Z terms with D (estimating a polynomial in Z on both sides of Z_0, and comparing the intercepts at Z_0). However, since the goal is to compute an effect at precisely one point (Z_0) using only the closest observations, the standard approach is to use local linear regression, which minimizes bias (Fan and Gijbels 1996). In Stata 10, this is done with the `lpoly` command; users of previous versions can use `locpoly` (Gutierrez, Linhart, and Pitblado 2003).

Having chosen to use local linear regression, other key issues are the choice of bandwidth and kernel. Various techniques are available for choosing bandwidths (see e.g. Fan and Gijbels 1996, Stone 1974, 1977), and the triangle kernel has good properties in the RD context, due to being boundary optimal (Cheng et al. 1997).

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 `lpoly` smooth with the pattern in a `scatter` graph. Because different bandwidth choices can produce different estimates, the researcher should report at least three estimates as an informal sensitivity test: one using the preferred bandwidth, and estimates using twice and half the preferred bandwidth.

5.3 (T1) X^T jumps at Z_0

The identifying assumption is that X^T jumps at Z_0 due to some known legal or program design rules, but we can test that assumption easily enough. The standard approach to computing standard errors is to `bootstrap` the local linear regression, which requires wrapping the estimation in a program, for example,

```
program discont, rclass
 version 10
 syntax [varlist(min=2 max=2)] [, *]
 tokenize `varlist'
 tempvar z f0 f1
 qui g `z'=0 in 1
 local opt "at(`z') nogr k(tri) deg(1) `options'"
 lpoly `1' `2' if `2'<0, gen(`f0') `opt'
 lpoly `1' `2' if `2'>=0, gen(`f1') `opt'
 return scalar d=`=`f1'[1]-`f0'[1]'
 di as txt "Estimate: " as res `f1'[1]-`f0'[1]
 eret clear
end
```

In the program, the assignment variable Z is assumed to be defined so that the cutoff $Z_0 = 0$ (easily done with a single `replace` or `generate` command subtracting Z_0 from Z). The triangle kernel is used, and the default bandwidth is chosen by `lpoly`, which is probably suboptimal for this application. The local linear regressions are computed twice, once using observations on one side of the cutoff, for $Z < 0$, and once for $Z \geq 0$. The estimate of the jump uses only the predictions at the cutoff $Z_0 = 0$, so these are the only values computed by `lpoly`.

We can easily generate data to try this example program out:

```
ssc inst rd, replace
net get rd
use votex if i==1
ren lne y
ren win xt
ren d z
foreach v of varlist pop-vet {
 ren ‘v’ xc_‘v’
 }
bs r(d): discont y z
```

In a more elaborate version of this program called `rd` (which also supports earlier versions of Stata), available by typing `ssc inst rd, replace` in Stata, the default bandwidth is selected to include at least 30 observations in estimates at both sides of the boundary; other options are also available. Try `findit bandwidth` to find more sophisticated bandwidth choosers for Stata. The key point is to use the `at()` option so that the difference in local regression predictions can be computed at Z_0.

A slightly more elaborate version of this program would save local linear regression estimates at a number of points, and offer a graph:

```
program discont2, rclass
 version 10
 syntax [varlist(min=2 max=2)] [, s(str) Graph *]
 tokenize ‘varlist’
 tempvar z f0 f1 se0 se1 ub0 ub1 lb0 lb1
 su ‘2’, meanonly
 local N=round(100*(r(max)-r(min)))
 cap set obs ‘N’
 qui g ‘z’=(_n-1)/100 in 1/50
 qui replace ‘z’=-(_n-50)/100 in 51/‘N’
 local opt "at(‘z’) nogr k(tri) deg(1) ‘options’"
 lpoly ‘1’ ‘2’ if ‘2’<0, gen(‘f0’) se(‘se0’) ‘opt’
 qui replace ‘f0’=. if ‘z’>0
 qui g ‘ub0’=‘f0’+1.96*‘se0’
 qui g ‘lb0’=‘f0’-1.96*‘se0’
 lpoly ‘1’ ‘2’ if ‘2’>=0, gen(‘f1’) se(‘se1’) ‘opt’
 qui replace ‘f1’=. if ‘z’<0
 qui g ‘ub1’=‘f1’+1.96*‘se1’
 qui g ‘lb1’=‘f1’-1.96*‘se1’
 return scalar d=‘=‘f1’[1]-‘f0’[1]’
 return scalar f1=‘=‘f1’[1]’
 return scalar f0=‘=‘f0’[1]’
 forv i=1/50 {
  return scalar p‘i’=‘=‘f1’[‘i’]’
  }
 forv i=51/‘N’ {
  return scalar n‘=‘i’-50’=‘=‘f0’[‘i’]’
  }
 di as txt "Estimate: " as res ‘f1’[1]-‘f0’[1]
 if "‘graph’"!="" {
  la var ‘z’ "Assignment Variable"
  loc lines "|| line ‘f0’ ‘f1’ ‘z’"
  loc a "tw rarea ‘lb0’ ‘ub0’ ‘z’ || rarea ‘lb1’ ‘ub1’ ‘z’"
  ‘a’ || sc ‘1’ ‘2’, mc(gs14) leg(off) sort ‘lines’
  }
 if "‘s’"!="" {
```

```
  ren ‘z’ ‘s’‘2’
  ren ‘f0’ ‘s’‘1’0
  ren ‘lb0’ ‘s’‘1’lb0
  ren ‘ub0’ ‘s’‘1’ub0
  ren ‘f1’ ‘s’‘1’1
  ren ‘lb1’ ‘s’‘1’lb1
  ren ‘ub1’ ‘s’‘1’ub1
  }
 eret clear
end
```

In this version, the local linear regressions are computed at a number of points (in the example, the maximum of Z is assumed to be 0.5, so the program uses hundredths as a convenient unit for Z) on either side of the cutoff $Z_0 = 0$, but the estimate of a jump still uses only the two estimates at $Z_0 = 0$. Note that the `s()` option in the above program saves the local linear regression predictions (and `lpoly` confidence intervals) to new variables that can then be graphed. Graphs of all output are advisable, to assess the quality of the fit for each of several bandwidths. This program may also be bootstrapped, though recovering the standard errors around each point estimate from `bootstrap` for graphing the fit is much more work.

5.4 (T2) y and X^C continuous away from Z_0

While we need only assume continuity at Z_0, and need no assumption that the outcome and treatment variables are continuous at values of Z away from the cutoff Z_0 (i.e. $\Delta X^T(Z \neq Z_0) = 0$ and $\Delta y(Z \neq Z_0) = 0$), it is reassuring if we fail to reject the null of a zero jump at various values of Z away from the cutoff Z_0 (or reject the null only in 5% of cases or so). Having defined a program `discont`, we can easily randomly choose 100 placebo cutoff points $Z_p \neq Z_0$, without replacement in the example below, and test the continuity of X^T and y at each.

```
bys z: g f=_n>1 if z!=0
g u=uniform()
sort f u
replace u=(_n<=100)
levelsof z if u, loc(p)
foreach val of local p {
 cap drop newz
 g newz=z-‘val’
 bootstrap r(d), reps(100): discont y znew
 bootstrap r(d), reps(100): discont xt znew
}
```

5.5 (T3) X^C continuous around Z_0

If we can regard an increase in treatment X^T as randomly assigned in the neighborhood of the cutoff Z_0, then predetermined characteristics X^C such as race or sex of treated individuals should not exhibit a discontinuity at the cutoff Z_0. This is equivalent to the standard test of randomization in an experimental design, using a test of the equality of the mean of every variable in X^C across treatment and control groups (see `help hotelling` in Stata), or the logically equivalent test that all the coefficients on X^C in a regression of X^T on X^C are zero. As in the experimental setting, in practice the tests are usually done one at a time with no adjustment for multiple hypothesis testing (see `help _mtest` in Stata).

In the RD setting, this is simply a test that the measured jump in each predetermined X^C is zero at the cutoff Z_0, or $\Delta X^C(Z_0) = 0$ for all X^C. If we fail to reject that the measured jump in X^C is zero, for all X^C, we have some additional evidence that observations on both sides of the cutoff are exchangeable, at least in some neighborhood of the cutoff, and we can treat them as if they were randomly assigned treatment in that neighborhood.

Having defined the programs `discont` and `discont2`, we can simply type:

```
foreach v of varlist xc* {
 bootstrap r(d), reps(100): discont ‘v’ z
 discont2 ‘v’ z, s(h)
 sc ‘v’ z, mc(gs14) sort || line h‘v’0 h‘v’1 hz, name(‘v’)
 drop hz
 }
```

5.6 (T4) Density of Z continuous at cutoff

McCrary (2007) gives an excellent account of a violation of exchangability of observations around the cutoff. If individuals have preferences over treatment and can manipulate assignment, for instance by altering their Z or misreporting it, then individuals close to Z_0 may shift across the boundary. For example, some non-randomly selected subpopulation of those who are nearly eligible for food stamps may misreport income, while those who are eligible do not. This creates a discontinuity in the density of Z at Z_0. McCrary (2007) points out that the absence of a discontinuity in the density of Z at Z_0 is neither necessary nor sufficient for exchangability, but a failure to reject the null hypothesis that the jump in the density of Z at Z_0 is zero is reassuring nonetheless.

McCrary (2007) discussed a test in detail, and advocated a bandwidth chooser. We can also adapt our existing program to the purpose by using `kdensity` commands to estimate the density to the left and right of Z_0:

```
kdensity z if z<0, gen(f0) at(z) tri nogr
count f0 if z>=0
replace f0=f0/r(N)*‘=_N’/4
kdensity z if z>=0, gen(f1) at(z) tri nogr
count f1 if z<0
replace f1=f1/r(N)*‘=_N’/4
generate f=cond(z>=0,f1,f0)
bootstrap r(d), reps(100): discont f z
discont2 f z, s(h) g
```

Alternatively, we could wrap the `kdensity` estimation inside the program that estimates the jump, so that both are bootstrapped together; this approach is taken by `rd` available from `ssc inst rd, replace`.

5.7 (T5) Treatment Effect Estimator

Having defined the program `discont`, we can simply:

```
bootstrap r(d), reps(100): discont y z
```

to get an estimate of the treatment effect in a “sharp” RD setting where X^T jumps from zero to one at Z_0. For a “fuzzy” RD design, we wish to compute the jump in y scaled by the jump in X^T at Z_0, or the **local Wald estimate**, for which we need to modify our program to estimate both discontinuities. The program `rd` available from `ssc inst rd, replace` does this, but the idea is illustrated in the program below, using the previously defined `discont` program twice.

```
prog lwald, rclass
 version 10
 syntax varlist [, w(real .06) ]
 tokenize ‘varlist’
 di as txt "Numerator"
 discont ‘1’ ‘3’, bw(‘w’)
 loc n=r(d)
```

```
  return scalar numerator=‘n’
  di as txt "Denominator"
  discont ‘2’ ‘3’, bw(‘w’)
  loc d=r(d)
  return scalar denominator=‘d’
  return scalar lwald=‘n’/‘d’
  di as txt "Local Wald Estimate:" as res ‘n’/‘d’
  eret clear
end
```

This program takes three arguments, the variables y, X^T, and Z, assumes $Z_0 = 0$, and uses a hardwired default bandwidth of 0.06. Note that the default bandwidth selected by `lpoly` is inappropriate for these models, since we do not use a Gaussian kernel, and are interested in boundary estimates. The `rd` program on SSC is similar in spirit to the above, though it offers more options, particularly with regard to bandwidth selection.

5.8 Examples

Voting examples abound. A novel estimate in Nichols and Rader (2007) measures the effect of electing as a Representative a Democratic incumbent versus a Republican incumbent on a district's receipt of federal grants:

```
  ssc inst rd, replace
  net get rd
  use votex if i==1
  rd lne d, gr
  bs: rd lne d, x(pop-vet)
```

The above estimates that the marginally victorious Democratic incumbent brings roughly 20% less to his home district than a marginally victorious Republican incumbent. However, we cannot reject the null of zero difference, and that is true for a variety of bandwidth choices (Figure 2 shows the small insignificant effect). Note that the above is a sharp RD design, but the Wald estimator can be used to estimate effect, since the jump in `win` at 50% of vote share is one, and dividing by one has no impact on estimates.

Many good examples of fuzzy RD designs concern educational policy or interventions (e.g. van der Klaauw 2002 or Ludwig and Miller 2005). Many educational grants are awarded using deterministic functions of predetermined characteristics, lending themselves to evaluation using RD. For example, some US Department of Education grants to states are awarded to districts with poverty (or near-poverty) rates above a threshold, as determined by data from a prior Census, which satisfies all of the requirements for RD, though the size of the discontinuity in funding may often be insufficient to identify an effect. In many cases, a power analysis is warranted to determine the minimum detectable effect.

Returning to the Card (1995ab) example of the effect of education on earnings, we can imagine exploiting a discontinuity in the availability of college to residents of certain US states at the state boundary. College applicants who live 4.8 miles and 5 miles from a college may look very similar in various observable characteristics, but if a state boundary separates them at 4.9 miles from the college, and the college is a state institution, they may face very different probabilities of admission or tuition costs. The data in Card (1995ab) do not support this strategy, of course, since we would need to know the exact locations of all individuals relative to state boundaries, but it helps to clarify the assumptions that justify the IV approach. We need to assume that location relative to colleges is randomly sprinkled over potential applicants, which seems questionable (Black 1999), especially when one considers including parental education in the model.

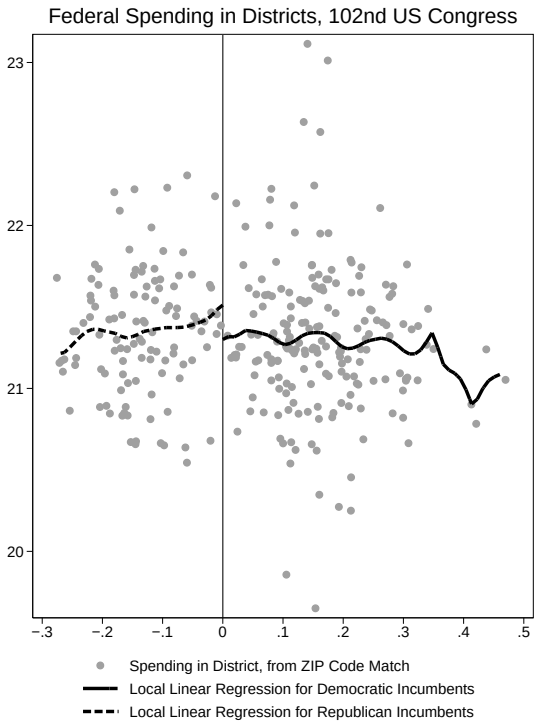

Figure 2: RD Example

6 Conclusions

In many circumstances, exploring data using quasi-experimental methods is the only option for estimating a causal effect, when experiments are infeasible, and may sometimes be preferred even when an experiment is feasible, particularly if a marginal treatment effect is of interest. However, the methods can suffer a number of severe problems when assumptions are violated, even weakly. For this reason, the details of implementation are frequently crucial, and a kind of cookbook or checklist for verifying essential assumptions are satisfied has been provided above for the interested researcher. As the topics discussed continue to be active research areas, this cookbook should be taken merely as a starting point for further explorations of the applied econometric literature on the relevant subjects.

7 References

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

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.

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

Angrist, Joshua D. and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics,* 106:4, pp. 9791014.

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

Azevedo, João Pedro. 2005. "dfl: Stata module to estimate DiNardo, Fortin and Lemieux Counterfactual Kernel Density." [SSC].

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.

Baum, Christopher. 2006. "Time series filtering techniques in Stata." Presentation at NASUG Meetings 2006 [http://ideas.repec.org/p/boc/asug06/2.html]

Baum, Christopher, Mark Schaffer, and Steven Stillman. 2007. "Enhanced routines for instrumental variables/GMM estimation and testing." Boston College Working Papers in Economics 667.

Baum, Christopher, Mark Schaffer, Steven Stillman, and Vince Wiggins. 2006. "overid: Stata module to calculate tests of overidentifying restrictions after ivreg, ivreg2, ivprobit, ivtobit, reg3." [SSC].

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.

Black, Sandy. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics,* 114: 577-599.

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. 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.

Card, David E. 1995a. "Earnings, Schooling and Ability Revisited." *Research in Labor Economics,* 14: 2348. Also NBER WP 4832.

Card, David E. 1995b. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling" in *Aspects of Labour Economics: Essays in Honour of John Vanderkamp,* edited by Louis Christofides, E. Kenneth Grant and Robert Swindinsky. University of Toronto Press. Also NBER WP 4483.

Card, David E. 1999. "The Causal Effect of Education on Earnings." *Handbook of Labor Economics,* 3(1): 180163.

Card, David E. 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica,* 69(5): 1127–1160.

Cheng, Ming-Yen, Jianqing Fan, and James S. Marron. 1997. "On Automatic Boundary Corrections." *Annals of Statistics*, 25(4), 1691-1708.

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

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.

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

Fan, Jianqing and Irene Gijbels. 1996. *Local Polynomial Modelling and Its Applications.* New York: Chapman and Hall.

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.

Glazerman, Steven, Dan M. Levy, and David Myers. (2003). "Nonexperimental versus experimental estimatesof earnings impacts." *The Annals of the American Academy of Political and Social Science,* 589: 63-93.

Goldberger, Arthur S., and Otis 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.

Griliches, Zvi and Jerry A. Hausman. 1986. "Errors in Variables in Panel Data." *Journal of Econometrics,* 31: 93-118. Originally published 1984 as NBER Technical Working Paper No. 37.

Gutierrez, Robert, J. M. Linhart, and J. S. Pitblado. 2003. "From the help desk: Local polynomial regression and Stata plugins." *Stata Journal*3(4):412–419. Also `findit locpoly` for updates (e.g. *Stata Journal* 6(4):597).

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.

Hall, Alastair R., Glenn D. Rudebusch, and David W. Wilcox. 1996. "Judging Instrument Relevance in Instrumental Variables Estimation." *International Economic Review,* 37(2): 283-298.

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. 2004. "Structural Equations, Treatment Effects and Econometric Policy Evaluation." *Econometrica,* 73(3): 669738. See also NBER technical working paper 306.

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

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

Imai, Kosuke and David A. van Dyk. (2004). "Causal inference with general treatment regimes: Generalizing the propensity score." *Journal of the American Statistical Association,* 99(467): 854-866.

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 and Thomas Lemieux. 2007. "Regression Discontinuity Designs: A Guide to Practice." NBER Working Paper 13039.

Jann, Ben. 2005a. "jmpierce: Stata module to perform Juhn-Murphy-Pierce decomposition." [SSC].

Jann, Ben. 2005b. "oaxaca: Stata module to compute decompositions of outcome differentials." [SSC].

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

Kleibergen, Frank and Mark Schaffer. 2007. "ranktest: module for testing the rank of a matrix using the Kleibergen-Paap rk statistic." [SSC].

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. 2006. "Regression Discontinuity Inference with Specification Error." NBER Technical Working Paper 322. Newer version forthcoming in *Journal of Econometrics*. Previous version: Center for Labor Economics Working Paper 74.

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". [SSC].

Ludwig, Jens and Douglas L. Miller. 2005. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." NBER WP 11702.

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 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. Related working paper.

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.

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.

Nichols, Austin. 2006. "Weak Instruments: An Overview and New Techniques." Presentation at NASUG Meetings 2006 [http://ideas.repec.org/p/boc/asug06/3.html]

Nichols, Austin and Kelly Rader. 2007. "Spending in the Districts of Marginal Incumbent Victors in the House of Representatives." Unpublished Working Paper.

Nichols, Austin and Mark E. Schaffer. 2007. "Cluster-robust and GLS Corrections." Unpublished Working Paper.

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

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

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.

Rothstein, Jesse. 2007. "Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference." Working Paper, Princeton University.

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

Rubin, Donald B. 1986. "Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers." *Journal of the American Statistical Association,* 81(396): 961–962.

Rubin, Donald B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science,* 5(4): 472–480.

Schaffer, Mark and Steven Stillman. 2007. "xtivreg2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models." [SSC]

Schaffer, Mark and Steven Stillman. 2006. "xtoverid: Stata module to calculate tests of overidentifying restrictions after xtreg, xtivreg, xtivreg2 and xthtaylor." [SSC]

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Simpson, Edward H. 1951. "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society,* Ser. B 13: 238–241.

Spence, Michael. 1973. "Job Market Signaling." *The Quarterly Journal of Economics,* 87(3): 355–374.

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.

Stone, Mervyn. 1974. "Cross-Validation and Multinomial Prediction." *Biometrika,* 61(3), 509-515.

Stone, Mervyn. 1977. "Asymptotics For and Against Cross-Validation." *Biometrika,* 64(1), 29-35.

Stuart, Elizabeth A. and Donald B. Rubin. 2007. "Best Practices in Quasi-Experimental Designs: Matching Methods for Causal Inference." in *Best Practices in Quantitative Social Science,* Edited by Jason Osborne. Thousand Oaks, CA: Sage Publications.

van der Klaauw, Wilbert. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression Discontinuity Approach." *International Economic Review,* 43: 1249-1287.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Yule, G. Udny. 1903. "Notes on the Theory of Association of Attributes in Statistics." *Biometrika,* 2: 121–134.

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.