Econometric Methods for Policy Evaluation

By Joan Llull[*],[§]

CEMFI. Winter 2016

I. Motivation: Structural vs Treatment Effect Approaches

The evaluation of public (and private) policies is very important for efficiency, and ultimately to improve welfare. There is a vast literature in economics, mostly in public economics, but also in development economics and labor economics, devoted to the evaluation of different programs. Examples include training programs, welfare programs, wage subsidies, minimum wage laws, taxation, Medicaid and other health policies, school policies, feeding programs, microcredit, and a variety of other forms of development assistance. These analyses aim at quantifying the effects of these policies on different outcomes, and ultimately on welfare.

The classic approach to quantitative policy evaluation is the ***structural approach***. This approach specifies a class of theory-based models of individual choice, chooses the one within the class that best fits the data, and uses it to evaluate policies through simulation. This approach has the main advantage that it allows both *ex-ante* and *ex-post* policy evaluation, and that it permits evaluating different variations of a similar policy without need to change the structure of the model or reestimate it (out of sample simulation). The main critique to this approach, though, is that there is a host of untestable functional form assumptions that undermine the force of the structural evidence because they have unknown implications for the results, give researchers too much discretion, and its complexity often affects transparency and replicability. Some people has argued that this approach puts too much emphasis on external validity at the expense of a more basic internal validity.

During the last two decades, the ***treatment effect approach*** has established itself as an important competitor that has introduced a different language, different priorities, techniques, and practices in applied work. This approach has changed the perception of evidence-based economics among economists, public opinion, and policy makers. The main goal of this approach is to evaluate (*ex-*

[*] Departament d'Economia i Història Econòmica. Universitat Autònoma de Barcelona. Facultat d'Economia, Edifici B, Campus de Bellaterra, 08193, Cerdanyola del Vallès, Barcelona (Spain). E-mail: joan.llull[at]movebarcelona[dot]eu. URL: http://pareto.uab.cat/jllull.

[§] These materials are based on earlier materials from the course by Manuel Arellano, available at http://www.cemfi.es/∼arellano.

post) the impact of an existing policy by comparing the distribution of a chosen outcome variable for individuals affected by the policy (the treatment group), with the distribution of unaffected individuals (control group). The main challenge of this approach is to find a way to perform the comparison in such a way that the distribution of outcome for the control group serves as a good counterfactual for the distribution of the outcome for the treated group in the absence of treatment. The main focus of this approach is in the understanding of the sources of variation in data with the objective of identifying the policy parameters, even though these parameters are formally not valid representations of the outcomes of implementing the same policy in an alternative environment, or of implementing variations of the policy even to the same environment. Thus, this approach helps in the assessment of future policies in a more informal way.

The main advantage of this approach is that, given its focus on internal validity, the exercise gives transparent and credible identification. The main disadvantage is that estimated parameters are not useful for welfare analysis because they are not deep parameters (they are reduced-forms instead), and as a result, they are not policy-invariant (Lucas, 1976; Heckman and Vytlacil, 2005). In that respect, a treatment effect exercise is less ambitious.

The deep differences between the two approaches has split the economics profession into two camps whose research programs have evolved almost independently despite focusing on similar questions (Chetty, 2009). However, recent developments have changed this trend, as researchers realized about the important complementarity between the two. The survey articles by Chetty (2009) and Todd and Wolpin (2010) review the progress made along those lines and point to avenues for future developments.

In this part of the course we will review the main designs for policy evaluation under the treatment effect approach. In the second part (with Pedro) you will review structural approaches. And by the end of this part, if time permits, I will introduce some bridges between the two, which will serve as an introduction to the second part.

II. Potential Outcomes and Causality: Treatment Effects

A. Potential outcomes and treatment effects

Consider the population of individuals that are susceptible of a treatment. Let Y_{1i} denote the outcome for an individual i if exposed to the treatment ($D_i = 1$), and let Y_{0i} be the outcome for the same individual if not exposed ($D_i = 0$). The ***treatment effect*** for individual i is thus $Y_{1i} - Y_{0i}$. Note that Y_{1i} and Y_{0i} are

potential outcomes in the sense that we only observe $Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i)$. This poses the main challenge of this approach, as the treatment effect can not be computed for a given individual. Fortunately, our interest is not in treatment effects for specific individuals *per se*, but, instead, in some characteristics of their distribution.

For most of the time, we will focus on two main parameters of interest. The first one is the ***average treatment effect*** (ATE):

$$\alpha_{ATE} \equiv \mathbb{E}[Y_1 - Y_0], \tag{1}$$

and the second one is ***average treatment effect on the treated*** (TT):

$$\alpha_{TT} \equiv \mathbb{E}[Y_1 - Y_0|D = 1]. \tag{2}$$

As noted, the main challenge is that we only observe Y. The standard measure of association between Y and D (the regression coefficient) is:

$$\begin{aligned} \beta &\equiv \mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0] \\ &= \underbrace{\mathbb{E}[Y_1 - Y_0|D = 1]}_{\alpha_{TT}} + (\mathbb{E}[Y_0|D = 1] - \mathbb{E}[Y_0|D = 0]), \end{aligned} \tag{3}$$

which differs from α_{TT} unless the second term is equal to zero. The second term indicates the difference in potential outcomes when untreated for individuals that are actually treated and individuals that are not. A nonzero difference may result from a situation in which treatment status is the result of individual decisions where those with low Y_0 choose treatment more frequently than those with high Y_0.

From a structural model of D and Y, one could obtain the implied average treatment effects. Instead, here, they are defined with respect to the distribution of potential outcomes, so that, relative to the structure, they are ***reduced-form causal effects***. Econometrics has conventionally distinguished between reduced form effects, uninterpretable but useful for prediction, and structural effects, associated with rules of behavior. The treatment effects provide this intermediate category between predictive and structural effects, in the sense that recovered parameters are causal effects, but they are uninterpretable in the same sense as reduced form effects.

An important assumption of the potential outcome representation is that the effect of the treatment on one individual is independent of the treatment received by other individuals. This excludes equilibrium or feedback effects, as well as strategic interactions among agents. Hence, the framework is not well suited to the evaluation of system-wide reforms which are intended to have substantial equilibrium effects.

Sample analogs for α_{ATE} and α_{TT} are:

$$\alpha^S_{ATE} \equiv \frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}}(Y_{1i} - Y_{0i}) \tag{4}$$

$$\alpha^S_{TT} \equiv \frac{1}{\sum_{\mathrm{i=1}}^{\mathrm{N}} D_i}\sum_{\mathrm{i=1}}^{\mathrm{N}} D_i(Y_{1i} - Y_{0i}). \tag{5}$$

If factual and counterfactual potential outcomes were observed, these quantities could be estimated without error. However, since they are not, the distinction is not very useful on practical grounds. Importantly, though, depending on whether we estimate population (α) or sample (α^S) average treatment effects, standard errors will be different, so we should take this into account when computing confidence intervals. The sample average version of β is given by:

$$\begin{aligned}\beta^S &\equiv \bar{Y}_T - \bar{Y}_C \\ &\equiv \frac{1}{N_1}\sum_{\mathrm{i=1}}^{\mathrm{N}} Y_i D_i - \frac{1}{N_0}\sum_{\mathrm{i=1}}^{\mathrm{N}}(1 - D_i)Y_i, \end{aligned} \tag{6}$$

where $N_1 \equiv \sum_{\mathrm{i=1}}^{\mathrm{N}} D_i$ is the number of treated individuals, and $N_0 \equiv N - N_1$ is the number of untreated.

B. Identification of treatment effects under different assumptions

The identification of the treatment effects depends on the assumptions we can make on the relation between potential outcomes and the treatment. The easiest case is when the distribution of the potential outcomes is independent of the treatment:

$$(Y_1, Y_0) \perp D. \tag{7}$$

This situation is typical in randomized experiments, where individuals are assigned to treatment or control in a random manner. When this happens, $F(Y_1|D = 1) = F(Y_1)$, and $F(Y_0|D = 0) = F(Y_0)$, which implies that $\mathbb{E}[Y_1] = \mathbb{E}[Y_1|D = 1] = \mathbb{E}[Y|D = 1]$ and $\mathbb{E}[Y_0] = \mathbb{E}[Y_0|D = 0] = \mathbb{E}[Y|D = 0]$, and, as a result, $\alpha_{ATE} = \alpha_{TT} = \beta$. In this case, an unbiased estimate of α_{ATE} is given by the difference between the average outcomes for treatments and controls:

$$\widehat{\alpha}_{ATE} = \bar{Y}_T - \bar{Y}_C = \beta^S. \tag{8}$$

In this context, there is no need to "control" for other covariates, unless there is direct interest in their marginal effects, or in effects for specific groups.

A less restrictive assumption is conditional independence:

$$(Y_1, Y_0) \ \perp \ D|X, \tag{9}$$

where X is a vector of covariates. This situation is known as matching, as for each "type" of individual (i.e. each value of covariates) we can match treated and control individuals, so that the latter act as counterfactuals for the former. Conditional independence implies $\mathbb{E}[Y_1|X] = \mathbb{E}[Y_1|D = 1, X] = \mathbb{E}[Y|D = 1, X]$ and $\mathbb{E}[Y_0|X] = \mathbb{E}[Y_0|D = 0, X] = \mathbb{E}[Y|D = 0, X]$, and, as a result:

$$\begin{aligned}\alpha_{ATE} = \mathbb{E}[Y_1 - Y_0] &= \int \mathbb{E}[Y_1 - Y_0|X]dF(X) \\ &= \int (\mathbb{E}[Y|D = 1, X] - \mathbb{E}[Y|D = 0, X])dF(X),\end{aligned} \tag{10}$$

or, in words, compute the difference in average observed outcomes of treated and controls for each value of X, and integrate over the distribution of X. For the treatment effect on the treated:

$$\begin{aligned}\alpha_{TT} &= \int \mathbb{E}[Y_1 - Y_0|D = 1, X]dF(X|D = 1) \\ &= \int \mathbb{E}[Y - \mathbb{E}[Y_0|D = 1, X]|D = 1, X]dF(X|D = 1) \\ &= \int \mathbb{E}[Y - \mu_0(X)|D = 1, X]dF(X|D = 1),\end{aligned} \tag{11}$$

where $\mu_0(X) \equiv \mathbb{E}[Y|D = 0, X]$, and we use the fact that $\mathbb{E}[Y|D = 0, X] = \mathbb{E}[Y_0|X] = \mathbb{E}[Y_0|D = 1, X]$. The function $\mu_0(X)$ is used as an imputation for Y_0.

Finally, sometimes we cannot assume conditional independence:

$$(Y_1, Y_0) \ \not\perp \ D|X. \tag{12}$$

In this case, we will need some variable Z that constitutes an ***exogenous source of variation in*** D, in the sense that it satisfies the independence assumption:

$$(Y_1, Y_0) \ \perp \ Z|X, \tag{13}$$

and the relevance condition:

$$Z \ \not\perp \ D|X. \tag{14}$$

As we discuss below, in this context we are only going to be able to identify an average treatment effect for a subgroup of individuals, and we call the resulting parameter a ***local average treatment effect***.

III. Social Experiments

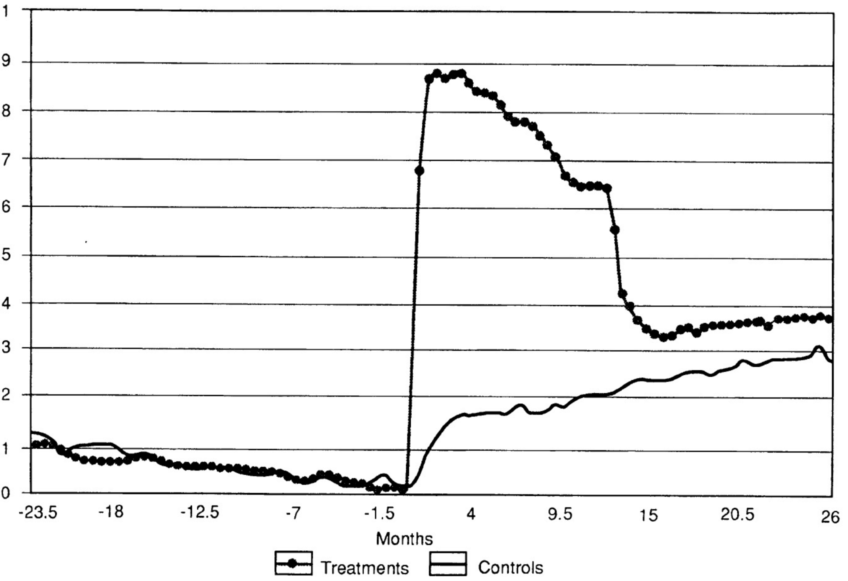
In the treatment effect approach, a randomized field trial is regarded as the ideal research design. Observational studies are seen as more speculative attempts to generate the force of evidence of experiments. In a controlled experiment, treatment status is randomly assigned by the researcher, which by construction, ensures independence. Thus, as noted above, $\alpha_{ATE} = \alpha_{TT} = \beta$.

There is a long history of randomized field trials in social welfare in the U.S., beginning in the 1960s (see Moffitt (2003) for a review). Early experiments had many flaws due to the lack of experience in designing them, and in data analysis. During the 1980s, the U.S. Federal Government started to encourage states to use experimentation, eventually becoming almost mandatory. The analysis of the 1980s experimental data consisted of simple treatment-control differences. The force of the results had a major influence on the 1988 legislation. In spite of these developments, randomization encountered resistance from many U.S. states on ethical grounds. Even more so in other countries, where treatment groups have often been formed by selecting areas for treatment instead of individuals.

One example of this is the National Supported Work program (NSW), which was designed in the U.S. in the mid 1970s to provide training and job opportunities to disadvantaged workers, as part of an experimental demonstration. Ham and LaLonde (1996) looked at the effects of the NSW on women that volunteered for training. NSW guaranteed to treated participants 12 months of subsidized employment (as trainees) in jobs with gradual increase in work standards. Eligibility requirements were to be unemployed, a long-term AFDC recipient, and have no preschool children. Participants were randomly assigned to treatment and control groups in 1976-1977. The experiment took place in 7 cities. Ham and LaLonde analyze data for 275 women in the treatment group and 266 controls. All volunteered in 1976.

Thanks to randomization, a simple comparison between employment rates of treatments and controls gives an unbiased estimate of the effect of the program on employment at different horizons. Figure 1 from Ham and LaLonde (1996) shows the effects. Initially, by construction there is a mechanical effect from the fact that treated women are offered a subsidized job. As apparent from the figure, compliance with the treatment is decreasing over time, as women can decide to drop from the subsidized job. The employment growth for controls is just a reflection of the program's eligibility criteria. Importantly, after the program ends, a 9 percentage points difference in employment rates is sustained in the

Figure 1. Employment rates of AFDC women in NSW Demonstration



Note: This figure corresponds to Figure 1 in Ham and LaLonde (1996)

medium run, at least until month 26 after the beginning of the program.

But Ham and LaLonde (1996) make an important additional point. Even though randomization allows researchers to evaluate the impact of the program on a particular outcome (employment) simply by comparing means, this is not true for any possible outcomes. In particular, if one is interested in the effect of the program on wages or on employment and unemployment durations, a comparison of means would provide a biased estimate of the effect of the program. This is because, as discussed above, the training program had an effect on employment rates of the treated.

To illustrate that, let W denote wages, let Y be an indicator variable that takes the value of one if the individual is employed, and zero if she is unemployed, and let η denote the ability type, with $\eta = 1$ if the individual is skilled, and $\eta = 0$ if she is unskilled. Suppose that the treatment increases the employment rates of high skill and low skill workers, but the effect is of less intensity for the high skilled (as they were more likely to find a job anyway without the training program):

$$P(Y=1|D=1,\eta=0) > P(Y=1|D=0,\eta=0), \tag{15}$$

$$P(Y=1|D=1,\eta=1) > P(Y=1|D=0,\eta=1), \tag{16}$$

and:

$$\frac{P(Y=1|D=1,\eta=0)}{P(Y=1|D=0,\eta=0)} > \frac{P(Y=1|D=1,\eta=1)}{P(Y=1|D=0,\eta=1)}. \tag{17}$$

This implies that the frequency of low skill will be greater in the group of employed treatments than in the employed controls:

$$P(\eta = 0|Y = 1, D = 1) > P(\eta = 0|Y = 1, D = 0), \tag{18}$$

which is a way to say that η, which is unobserved, is not independent of D given $Y = 1$, although, unconditionally, $\eta \perp D$. For this reason, a direct comparison of average wages between treatments and controls will tend to underestimate the effect of treatment on wages. In particular, consider the conditional effects:

$$\Delta_0 \equiv \mathbb{E}[W|Y = 1, D = 1, \eta = 0] - \mathbb{E}[W|Y = 1, D = 0, \eta = 0], \tag{19}$$

$$\Delta_1 \equiv \mathbb{E}[W|Y = 1, D = 1, \eta = 1] - \mathbb{E}[W|Y = 1, D = 0, \eta = 1]. \tag{20}$$

Our effect of interest is:

$$\Delta_{ATE} = \Delta_0 P(\eta = 0) + \Delta_1 P(\eta = 1), \tag{21}$$

whereas the comparison of average wages between treatments and controls delivers:

$$\Delta_W = \mathbb{E}[W|Y = 1, D = 1] - \mathbb{E}[W|Y = 1, D = 0]. \tag{22}$$

In general, we shall have $\Delta_W < \Delta_{ATE}$. Indeed, it may not be possible to construct an experiment to measure the effect of training the unemployed on subsequent wages, i.e. it does not seem possible to experimentally undo the conditional correlation between D and η.

A similar problem would occur with the comparison of exit rates from employment or unemployment:

$$\begin{aligned} &P(T = \tau|T \geq \tau, D = 1) - P(T = \tau|T \geq \tau, D = 0) \\ &\quad \neq P(T = \tau|T \geq \tau, D = 1, \eta) - P(T = \tau|T \geq \tau, D = 0, \eta). \end{aligned} \tag{23}$$

In particular, D is correlated with η given $T = \tau$ for various reasons (the argument is analogous to the classical discussion state dependence versus unobserved heterogeneity that you probably discuss in the Microeconometrics course).

IV. Matching

A. Selection based on observables and matching

There are many situations where experiments are too expensive, unfeasible, or unethical. A classical example is the analysis of the effects of smoking on mortality. In these situations, we have to rely on observational data, which is unlikely to satisfy independence. In some situations, we can arguably defend the assumption of conditional independence. We say that there is ***selection into treatment*** when the independence assumption is not satisfied. When we assume conditional independence, we say that we are assuming that there is ***selection based on observables***.

As discussed above, when there is selection based on observables, the simple comparison of treatment and control averages does deliver our treatment effects of interest. The problem is that the controls are not a counterfactual of treated in the absence of treatment, because the two groups differ in characteristics that are correlated with the outcome. However, Equations (11) and (11) delivers useful representations. For example, for the ATE:

$$\alpha_{ATE} = \int (\mathbb{E}[Y|D=1,X] - \mathbb{E}[Y|D=0,X])dF(X), \tag{24}$$

What the above expression does is to compare average outcomes for individuals with the same characteristics, and then integrate over the distribution of characteristics. This is called ***matching***, as it links each group of individuals in the treatment group with their counterparts in the control group. Provided that the selection is based on observables, for a given X, the assignment to treatment and control groups is random, and the above expression provides an unbiased computation of α_{ATE}. Similar arguments follow the expression for α_{TT}.

B. The common support condition

Note that an essential condition for matching is that, for each possible value of X, there are individuals in the treatment and control group for which we can average outcomes. This requirement is called as the ***common support condition***. Suppose, for the sake of the argument, that X is a single covariate whose support lies in the range (X_{min}, X_{max}). Suppose also that the support for the subpopulation of treated $(D = 1)$ is (X_{min}, X_1), and the support of the controls $(D = 0)$ is (X_0, X_{max}), with $X_0 < X_1$. In that case:

$$P(D=1|X) = \begin{cases} 1 & \text{if } X_{min} \leq X < X_0 \\ p \in (0,1) & \text{if } X_0 \leq X \leq X_1 \\ 0 & \text{if } X_1 < X \leq X_{max} \end{cases}. \tag{25}$$

The implication is that $\mathbb{E}[Y|D = 1, X]$ is only identified for values of X in the range (X_{min}, X_1), and $\mathbb{E}[Y|D = 0, X]$ is only identified for values of X in the range (X_0, X_{max}). Thus, we can only calculate the difference $\mathbb{E}[Y|D = 1, X] - \mathbb{E}[Y|D =$

$0, X]$ for values of X in the intersection range (X_0, X_1), which implies that α_{ATE} (and α_{TT}) is not identified. The bottom line is that the necessary condition to ensure identification, in addition to conditional independence, is the common support assumption, that can be stated as:

$$0 < P(D = 1|X) < 1 \quad \text{for all } X \text{ in its support.} \tag{26}$$

C. Estimation methods

Let us start from the discrete case. Suppose that X is discrete and takes on J possible values $\{x_j\}_{j=1}^{J}$, and we have a sample of N observations $\{X_i\}_{i=1}^{N}$. Let N^j be the number of observations in cell j, N_ℓ^j be the number of observations in cell j with $D = \ell$, and $\bar{Y}_\ell^j$ be the mean outcome in cell j for $D = \ell$. With this notation, $\bar{Y}_1^j - \bar{Y}_0^j$ is the sample counterpart of $\mathbb{E}[Y|D = 1, X = x_j] - \mathbb{E}[Y|D = 0, X = x_j]$, which can be used to get the following estimates:

$$\widehat{\alpha}_{ATE} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N^j}{N} \tag{27}$$

$$\widehat{\alpha}_{TT} = \sum_{j=1}^{J} \left(\bar{Y}_1^j - \bar{Y}_0^j\right) \frac{N_1^j}{N_1}. \tag{28}$$

Note that the formula for $\widehat{\alpha}_{TT}$ can also be written in the form:

$$\widehat{\alpha}_{TT} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - \bar{Y}_0^{j(i)}\right), \tag{29}$$

where $j(i)$ indicates the cell of X_i. Thus, $\widehat{\alpha}_{TT}$ matches the outcome of each treated unit with the mean of untreated units in the same cell, which is a way of imputing the missing outcome for the treated individuals, and compute the average treatment effect for them. Note that this expression is the sample analog of Equation (11).

In the continuous case, a matching estimator can be regarded as a way of constructing imputations for missing potential outcomes in a similar way, so that gains $Y_{1i} - Y_{0i}$ can be estimated for each unit. In the discrete case:

$$\widehat{Y}_{0i} = \bar{Y}_0^{j(i)} \equiv \sum_{k \in (D=0)} \frac{\mathbb{1}\{X_k = X_i\} Y_k}{\sum_{\ell \in (D=0)} \mathbb{1}\{X_\ell = X_i\}}, \tag{30}$$

and now we can generalize it to:

$$\widehat{Y}_{0i} = \sum_{k \in (D=0)} w(i, k) Y_k, \tag{31}$$

and different estimators will use different weighting schemes. For example, ***nearest neighbor matching*** uses the following weighting function:

$$w(i,k) = \mathbb{1}\{X_k = \min_i ||X_k - X_i||\}, \tag{32}$$

sometimes restricting the sample to cases in which $\min_i ||X_k - X_i|| < \varepsilon$ for some ε. This method is typically applied to compute α_{TT}, but it is also applicable to α_{ATE}.

An alternative weighting is given by ***kernel matching***:

$$w(i,k) = \frac{\kappa\left(\frac{X_k - X_i}{\gamma_{N_0}}\right)}{\sum_{\ell \in (D=0)} \kappa\left(\frac{X_\ell - X_i}{\gamma_{N_0}}\right)}, \tag{33}$$

where $\kappa(.)$ is a kernel that downweights distant observations, and γ_{N_0} us a bandwidth parameter.

Finally, a popular method for matching is the ***propensity score matching***. Rosenbaum and Rubin (1983) defined the propensity score as:

$$\pi(x) \equiv P(D = 1|X), \tag{34}$$

and proved that if $(Y_1, Y_0) \perp D|X$ then:

$$(Y_1, Y_0) \perp D|\pi(X), \tag{35}$$

provided that $0 < \pi(X) < 1$:

$$\begin{aligned} P(D = 1|Y_1, Y_0, \pi(X)) &= \mathbb{E}[D|Y_0, Y_1, \pi(X)] \\ &= \mathbb{E}[\mathbb{E}[D|Y_0, Y_1, X]|Y_0, Y_1, \pi(X)] \\ &= \mathbb{E}[\mathbb{E}[D|X]|Y_0, Y_1, \pi(X)] \\ &= \mathbb{E}[P(D = 1|X)|Y_0, Y_1, \pi(X)] \\ &= \mathbb{E}[\pi(X)|Y_0, Y_1, \pi(X)] \\ &= \pi(X), \end{aligned} \tag{36}$$

which proves the result in (35).

This result suggests two-step procedures to estimate the treatment effects where first we estimate the propensity score, and then create the appropriate weighting. To do so, let us rewrite α_{ATE} in terms of the propensity score. Under unconditional independence, we could write:

$$\alpha_{ATE} = \mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0] = \frac{\mathbb{E}[DY]}{P(D = 1)} - \frac{\mathbb{E}[(1 - D)Y]}{P(D = 0)}. \tag{37}$$

Thus, under conditional independence we can write:

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0|X] &= \mathbb{E}[Y|D=1,X] - \mathbb{E}[Y|D=0,X] \\ &= \frac{\mathbb{E}[DY|X]}{P(D=1|X)} - \frac{\mathbb{E}[(1-D)Y|X]}{P(D=0|X)} \\ &= \frac{\mathbb{E}[DY|X]}{\pi(X)} - \frac{\mathbb{E}[(1-D)Y|X]}{1-\pi(X)}, \end{aligned} \tag{38}$$

which implies:

$$\begin{aligned}\alpha_{ATE} &= \mathbb{E}\left[\mathbb{E}\left[\frac{DY}{\pi(X)} - \frac{(1-D)Y}{1-\pi(X)}\middle| X\right]\right] \\ &= \mathbb{E}\left[Y\frac{D-\pi(X)}{\pi(X)[1-\pi(X)]}\right]. \end{aligned} \tag{39}$$

Based on the sample analog of the above expression, Hirano, Imbens and Ridder (2003) propose the following estimator:

$$\widehat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i\left(\frac{D-\widehat{\pi}(X_i)}{\widehat{\pi}(X_i)[1-\widehat{\pi}(X_i)]}\right). \tag{40}$$

Note that is estimator is of the matching type described above, where the expression in parenthesis is the corresponding weight.

D. Advantages and disadvantages of matching

Given conditional independence one could alternatively estimate the treatment effect by means of a regression of the observed outcome Y on the treatment dummy D and the controls X. Controlling for X, D would be uncorrelated with the error term by assumption, and the coefficient of D would be a consistent estimate of α_{ATE} (if the assumption holds). Thus, it is natural to compare the two.

The main advantages of matching are that it avoids functional form assumptions and it emphasizes the common support condition. Matching focuses on a single parameter at a time, which is obtained through explicit aggregation. On the downside, matching works under the presumption that for $X = x$ there is random variation in D, so that we can observe both Y_0 and Y_1. Hence, it fails if D is a deterministic function of X, that is, if $\pi(X)$ is either 0 or 1. Additionally, there is a tension between the thought that if X is good enough then there may not be within-cell variation in D, and the suspicion that seeing enough variation in D given X is an indication that exogeneity is at fault.

E. Example: monetary incentives and schooling in the UK

As an illustration, Dearden, Emmerson, Frayne and Meghir (2009) analyze the effect of a conditional cash transfer on school participation in the UK. They participated in the design of the pilot and did the evaluation. The program was called *Education Maintenance Allowance* (EMA), and its pilot implementation started in September 2009. EMA paid youths aged 16-18 that continued in full time education (after 11 compulsory grades) a weekly stipend of £30 to £40, plus bonuses for good results up to £140. Eligibility (and amounts paid) depends on household characteristics. Eligible for full payments were households with annual income under £13,000; those above £30,000 were not eligible. For political reasons, there were no experimental design, but treatment and control areas were defined, both rural and urban.

The main question asked is whether more education results from this policy. The worry is that families fail to decide optimally due to liquidity constraints or missinformation. To address it, the authors use propensity scores. The reason is that, given that individuals in treatment and control areas can differ in characteristics, the unconditional independence may not hold.

To implement the matching approach, the authors estimate $\pi(X)$ using a Probit with family, local, and school characteristics. For each treated observation they construct a counterfactual mean using kernel regression and bootstrap standard errors. They find that EMA increased participation in grade 12 by 5.9% for eligible individuals, and by 3.7% for the whole population. They find estimated effects to be significantly different from zero only for full-payment recipients.

V. Instrumental Variables (IV)

A. Identification of causal effects in IV settings

Suppose that $(Y_1, Y_0) \not\perp D|X$, but we have an exogenous source of variation in D so that $(Y_1, Y_0) \perp Z|X$ that satisfies the relevance condition $Z \not\perp D|X$. In that situation, we can use the variation in Z to identify α_{ATE} under certain circumstances. Matching can be regarded as a special case in which $Z = D$, i.e. all the variation in D is exogenous given X. For simplicity, we do most of the analysis below considering a single binary instrument Z, and we abstract from including other covariates. We consider two different cases, depending on whether the treatment effects are homogeneous across individuals or not.

Homogeneous treatment effects. In this case, the causal effect is the same for every individual:

$$\widehat{Y}_{1i} - \widehat{Y}_{0i} = \alpha. \tag{41}$$

In this case, the availability of an instrumental variable allows us to identify α. This is the traditional situation in econometric models with endogenous explanatory variables (IV regression). To see it, note that:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i = Y_{0i} + \alpha D_i. \tag{42}$$

Taking into account that $Y_{0i} \perp Z_i$:

$$\alpha = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]}, \tag{43}$$

which is derived from:

$$\begin{aligned}
\alpha &= \frac{\mathrm{Cov}(Y_i, Z_i)}{\mathrm{Cov}(D_i, Z_i)} \\
&= \frac{\mathbb{E}[Y_i Z_i] - \mathbb{E}[Y_i]\,\mathbb{E}[Z_i]}{\mathbb{E}[D_i Z_i] - \mathbb{E}[D_i]\,\mathbb{E}[Z_i]} \\
&= \frac{\mathbb{E}[Y_i|Z_i = 1]P(Z_i = 1) - \{\mathbb{E}[Y_i|Z_i = 1]P(Z_i = 1) + \mathbb{E}[Y_i|Z_i = 0]P(Z_i = 0)\}P(Z_i = 1)}{\mathbb{E}[D_i|Z_i = 1]P(Z_i = 1) - \{\mathbb{E}[D_i|Z_i = 1]P(Z_i = 1) + \mathbb{E}[D_i|Z_i = 0]P(Z_i = 0)\}P(Z_i = 1)} \\
&= \frac{\mathbb{E}[Y_i|Z_i = 1]P(Z_i = 1)(1 - P(Z_i = 1)) - \mathbb{E}[Y_i|Z_i = 0]P(Z_i = 0)P(Z_i = 1)}{\mathbb{E}[D_i|Z_i = 1]P(Z_i = 1)(1 - P(Z_i = 1)) - \mathbb{E}[D_i|Z_i = 0]P(Z_i = 0)P(Z_i = 1)} \\
&= \frac{\mathbb{E}[Y_i|Z_i = 1]P(Z_i = 1)P(Z_i = 0) - \mathbb{E}[Y_i|Z_i = 0]P(Z_i = 0)P(Z_i = 1)}{\mathbb{E}[D_i|Z_i = 1]P(Z_i = 1)P(Z_i = 0) - \mathbb{E}[D_i|Z_i = 0]P(Z_i = 0)P(Z_i = 1)} \\
&= \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]}.
\end{aligned} \tag{44}$$

An alternative way to see it is that, given $Y_{0i} \perp Z_i$:

$$\left.\begin{aligned}\mathbb{E}[Y_i|Z_i = 1] = \mathbb{E}[Y_{0i}] + \alpha\,\mathbb{E}[D_i|Z_i = 1] \\ \mathbb{E}[Y_i|Z_i = 0] = \mathbb{E}[Y_{0i}] + \alpha\,\mathbb{E}[D_i|Z_i = 0]\end{aligned}\right\} \Rightarrow \alpha = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]}. \tag{45}$$

Identification obviously requires that $\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0] \neq 0$, which is the relevance condition. All in all, we are getting the effect of D on Y through the effect of Z because Z only affects Y through D (exclusion restriction).

Heterogeneous effects. In the heterogeneous case, the availability of instrumental variables is not sufficient to identify a causal effect (e.g. α_{ATE}). An additional assumption that helps identification of causal effects is the following ***monotonicity condition***: any person that was willing to treat if assigned to the control group would also be prepared to treat if assigned to the treatment group. The plausibility of this assumption depends on the context of the application. Under monotonicity, the IV coefficient coincides with the average treatment effect for those whose value of D would change when changing the value of Z, which is known as the ***local average treatment effect*** (LATE).

Let D_0 denote D when $Z = 0$, and let D_1 denote D when $Z = 1$. As we only observe D_ℓ, for ℓ either equal to one or to zero, the combination of treatment and instrument define four observable groups. However, using a potential outcome interpretation, we have eight potential groups, depending on the value of the unobserved treatment status $D_{-\ell}$:

Obs. type	Z	D	D_0	D_1	**Latent type**
Type 1	0	0	0	0	Never-taker
				1	Complier
Type 2	0	1	1	0	Defier
				1	Always-taker
Type 3	1	0	0	0	Never-taker
			1		Defier
Type 4	1	1	0	1	Complier
			1		Always-taker

To illustrate this classification, consider the following example. We are interested in the effect of going to college (treatment) on wages (outcome). Because individuals with higher ability may be more likely to go to college and, with a given educational level, to earn higher wages, independence does not hold neither conditionally nor unconditionally. Hence, to be able to identify a causal effect of college on wages, we need an instrument. We consider proximity to a college as an exogenous source of variation: it is associated with the cost of education, but plausibly uncorrelated with later outcomes. To make it dichotomous, we distinguish between being *far* and *close* from school. A complier is an individual that goes to school if she lives close to school, but would have not gone had she lived far, or one that does not go to school because she leaves far, but would have gone had she lived close. An individual that goes to school whether she lives close or far is an always-taker, and one that does not go to school whether she lives close or far is a never-taker. Defiers are those individuals that go to school being far, but would have not gone had they been close, or those who do not go being close, but would have gone had they been far.

To see that the availability of an instrumental variable is not enough to identify causal effects, consider the second derivation of the treatment effect for the homogeneous effects descried in Equation (45). Now we have:

$$\begin{aligned}\mathbb{E}[Y|Z=1] &= \mathbb{E}[Y_0] + \mathbb{E}[(Y_1 - Y_0)D_1]\\ \mathbb{E}[Y|Z=0] &= \mathbb{E}[Y_0] + \mathbb{E}[(Y_1 - Y_0)D_0],\end{aligned} \tag{46}$$

which implies:

$$\begin{aligned}\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0] &= \mathbb{E}[(Y_1 - Y_0)(D_1 - D_0)]\\ &= \mathbb{E}[Y_1 - Y_0|D_1 - D_0 = 1]P(D_1 - D_0 = 1)\\ &\quad - \mathbb{E}[Y_1 - Y_0|D_1 - D_0 = -1]P(D_1 - D_0 = -1).\end{aligned} \tag{47}$$

Thus, $\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]$ could be negative and yet the causal effect be positive for everyone, as long as the probability of defiers is sufficiently large.

One possibility is that we assume an ***eligibility rule*** of the form:

$$P(D=1|Z=0) = 0, \tag{48}$$

that is, individuals with $Z = 0$ are denied treatment (observable types 2, 3B, and 4B are ruled out). This situation is not implausible, as, in some designs treatment is offered to a subpopulation who endogenously select whether to comply with the treatment or not. In this case:

$$\begin{aligned}\mathbb{E}[Y|Z=1] &= \mathbb{E}[Y_0] + \mathbb{E}[(Y_1 - Y_0)D|Z=1]\\ &= \mathbb{E}[Y_0] + \mathbb{E}[Y_1 - Y_0|D=1, Z=1]P(D=1|Z=1),\end{aligned} \tag{49}$$

and, since $P(D = 1|Z = 0) = 0$:

$$\begin{aligned}\mathbb{E}[Y|Z=0] &= \mathbb{E}[Y_0] + \mathbb{E}[(Y_1 - Y_0)D|Z=0]\\ &= \mathbb{E}[Y_0] + \mathbb{E}[Y_1 - Y_0|D=1, Z=0]P(D=1|Z=0)\\ &= \mathbb{E}[Y_0].\end{aligned} \tag{50}$$

Therefore:

$$\begin{aligned}\alpha_{TT} &= \mathbb{E}[Y_1 - Y_0|D=1]\\ &= \mathbb{E}[Y_1 - Y_0|D=1, Z=1]\\ &= \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{P(D=1|Z=1)},\end{aligned} \tag{51}$$

where the second equality is the result of the fact that $P(Z = 1|D = 1) = 1$. Thus, if the eligibility condition in Equation (48) holds, the IV coefficient coincides with the treatment effect on the treated.

B. *Local average treatment effects (LATE)*

If we rule out defiers (which implies monotonicity), that is $P(D_1 - D_0 = -1) = 0$, Equation (47) reduces to:

$$\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0] = \mathbb{E}[Y_1 - Y_0|D_1 - D_0 = 1]P(D_1 - D_0 = 1), \tag{52}$$

and:

$$\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0] = \mathbb{E}[D_1 - D_0] = P(D_1 - D_0 = 1). \tag{53}$$

Thus, the causal effect that we can identify is:

$$\alpha_{LATE} \equiv \mathbb{E}[Y_1 - Y_0|D_1 - D_0 = 1] = \frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]}. \tag{54}$$

Imbens and Angrist (1994) called this parameter the local average treatment effect, because it is the average treatment effects on the subsample of compliers. Importantly, different instrumental variables lead to different parameters, even under instrument validity, which is counter to standard GMM thinking. This concept changed radically the way we think of and understand IV.

As noted, the identified coefficient is the average treatment effect for compliers. Thus, when selecting an instrument, on top of thinking about relevance and orthogonality conditions, the researcher needs to think about the potential group of compliers selected by the instrument. Most relevant LATEs are those based on instruments that are policy variables (e.g. college fee policies or college creation). Thus, in our example above on the returns to college, the identified effect is not going to be a good measurement of the average return to education in the overall population, but it would be a very good measure of how outcomes would react to a college subsidy.

As a final remark, what happens if there are no compliers? In the absence of defiers, the probability of compliers satisfies $P(D_1 - D_0 = 1) = \mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]$, so the lack of compliers implies lack of instrument relevance, and, hence, underidentification. This is natural, because if the population is formed of never-takers and always-takers, there is no role to be played by the instrument.

C. Conditional estimation with instrumental variables

So far we abstracted from the fact that the validity of the instrument may only be conditional on X: it may be that $(Y_0, Y_1) \not\perp Z$, but the following does:

$$\begin{gathered}(Y_0, Y_1) \perp Z|X \text{ (conditional independence)} \\ Z \not\perp D|X \text{ (conditional relevance)} .\end{gathered} \tag{55}$$

For example, in the analysis of the returns to college, Z is an indicator of proximity to college. The problem is that Z is not randomly assigned but chosen by parents, and this choice may depend on characteristics that subsequently affect wages. The validity of Z may be more credible if we can condition on family background, X.

In the linear version of the problem we can estimate using a two-stage procedure: first regress D on Z and X, so that we get $\widehat{D}$, and in the second stage we regress Y on $\widehat{D}$ and X. In general, we now have a conditional LATE given X:

$$\gamma(X) \equiv \mathbb{E}[Y_1 - Y_0 | D_1 - D_0 = 1, X], \tag{56}$$

and a conditional IV estimator:

$$\beta(X) \equiv \frac{\mathbb{E}[Y|Z=1,X] - \mathbb{E}[Y|Z=0,X]}{\mathbb{E}[D|Z=1,X] - \mathbb{E}[D|Z=0,X]}. \tag{57}$$

To get an aggregate effect, we proceed differently depending on whether the effects are homogeneous or heterogeneous. In the homogeneous case:

$$Y_1 - Y_0 = \beta(X). \tag{58}$$

In the heterogeneous case, it makes sense to consider an average treatment effect for the overall subpopulation of compliers:

$$\begin{aligned}\beta_C &\equiv \int \beta(X) \frac{P(compliers|X)}{P(compliers)} dF(X) \\ &= \int \{\mathbb{E}[Y|Z=1,X] - \mathbb{E}[Y|Z=0,X]\} \frac{1}{P(compliers)} dF(X),\end{aligned} \tag{59}$$

where:

$$P(compliers) = \int \{\mathbb{E}[D|Z=1,X] - \mathbb{E}[D|Z=0,X]\} \, dF(X). \tag{60}$$

Therefore:

$$\beta_C = \frac{\int \{\mathbb{E}[Y|Z=1,X] - \mathbb{E}[Y|Z=0,X]\} \, dF(X)}{\int \{\mathbb{E}[D|Z=1,X] - \mathbb{E}[D|Z=0,X]\} \, dF(X)}, \tag{61}$$

which can be estimated as a ratio of matching estimators (Frölich, 2003).

The endogenous dummy explanatory variable probit model. The model as usually written in terms of observables is:

$$\begin{aligned} Y &= \mathbb{1}\{\beta_0 + \beta_1 D + U \geq 0\} \\ D &= \mathbb{1}\{\pi_0 + \pi_1 Z + V \geq 0\} \end{aligned} \qquad \begin{pmatrix} U \\ V \end{pmatrix} \Bigg| Z \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{62}$$

In this model D is an endogenous explanatory variable as long as $\rho \neq 0$. D is exogenous if $\rho = 0$. In this model, there are only two potential outcomes:

$$\begin{aligned} Y_1 &= \mathbb{1}\{\beta_0 + \beta_1 + U \geq 0\} \\ Y_0 &= \mathbb{1}\{\beta_0 + U \geq 0\}. \end{aligned} \tag{63}$$

The ATE is given by:

$$\theta = \mathbb{E}[Y_1 - Y_0] = \Phi(\beta_0 + \beta_1) - \Phi(\beta_0). \tag{64}$$

In less parametric specifications, $\mathbb{E}[Y_1 - Y_0]$ may not be point identified, but we may still be able to estimate a LATE.

The index model for the treatment equation already imposes monotonicity. For example, consider the case in which Z is binary, so that there are only two potential values of D:

$$\begin{aligned} D_1 &= \mathbb{1}\{\pi_0 + \pi_1 + V \geq 0\} \\ D_0 &= \mathbb{1}\{\pi_0 + V \geq 0\}. \end{aligned} \tag{65}$$

Suppose, without loss of generality, that $\pi_1 \geq 0$. Then, we can distinguish three subpopulations depending on an individual's value of V:

Group	Conditon			Probability mass
Never-takers	$V < -\pi_0 - \pi_1$	$\Rightarrow$	$D_1 = 0, D_0 = 0$	$1 - \Phi(\pi_0 + \pi_1)$
Compliers	$-\pi_0 - \pi_1 \leq V \leq -\pi_0$	$\Rightarrow$	$D_1 = 1, D_0 = 0$	$\Phi(\pi_0 + \pi_1) - \Phi(\pi_0)$
Always-takers	$V \geq -\pi_0$	$\Rightarrow$	$D_1 = 1, D_0 = 1$	$\Phi(\pi_0)$

We can obtain the average treatment effect for the subpopulation of compliers:

$$\theta_{LATE} = \mathbb{E}[Y_1 - Y_0 | D_1 - D_0 = 1] = \mathbb{E}[Y_1 - Y_0 | -\pi_0 - \pi_1 \leq V < -\pi_0]. \tag{66}$$

We have:

$$\begin{aligned} \mathbb{E}[Y_1 | -\pi_0 - \pi_1 \leq V < -\pi_0] &= P(\beta_0 + \beta_1 + U \geq 0 | -\pi_0 - \pi_1 \leq V < -\pi_0) \\ &= 1 - \frac{P(U \leq -\beta_0 - \beta_1, V \leq -\pi_0) - P(U \leq -\beta_0 - \beta_1, V \leq -\pi_0 - \pi_1)}{P(V \leq -\pi_0) - P(V \leq -\pi_0 - \pi_1)}, \end{aligned} \tag{67}$$

and similarly:

$$\begin{aligned}\mathbb{E}[Y_0| - \pi_0 - \pi_1 \leq V < -\pi_0] &= P(\beta_0 + U \geq 0| - \pi_0 - \pi_1 \leq V < -\pi_0) \\ &= 1 - \frac{P(U \leq -\beta_0, V \leq -\pi_0) - P(U \leq -\beta_0, V \leq -\pi_0 - \pi_1)}{P(V \leq -\pi_0) - P(V \leq -\pi_0 - \pi_1)}. \qquad (68)\end{aligned}$$

Finally:

$$\theta_{LATE} = \frac{\left\{ \begin{array}{c} [\Phi_2(-\beta_0 - \beta_1, -\pi_0 - \pi_1; \rho) - \Phi_2(-\beta_0 - \beta_1, -\pi_0; \rho)] \\ -[\Phi_2(-\beta_0, -\pi_0 - \pi_1; \rho) - \Phi_2(-\beta_0, -\pi_0; \rho)] \end{array} \right\}}{\Phi(-\pi_0) - \Phi(-\pi_0 - \pi_1)}, \qquad (69)$$

where $\Phi_2(r, s; \rho) \equiv P(U \leq r, V \leq s)$ is the standard normal bivariate probability. The nice thing about θ_{LATE} is that it is identified also in the absence of joint normality. In fact, it does not even require monotonicity in the relationship between Y and D.

Models with additive errors: switching regressions. Consider the following switching regression model with endogenous switch:

$$\begin{aligned} Y_i &= \beta_0 + \beta_{1i} D_i + U_i \\ D_i &= \mathbb{1}\{\gamma_0 + \gamma_1 Z_i + \varepsilon_i \geq 0\}. \qquad (70)\end{aligned}$$

The potential outcomes are:

$$\begin{aligned} Y_{1i} &= \beta_0 + \beta_{1i} + U_i \equiv \mu_1 + V_{1i} \\ Y_{0i} &= \beta_0 + U_i \equiv \mu_0 + V_{0i}, \qquad (71)\end{aligned}$$

so that the treatment effect $\beta_{1i} = Y_{1i} - Y_{0i}$ is heterogeneous. Traditional models assume β_{1i} is constant or that it varies only with observable characteristics. In these models, D may be exogenous (independent of U) or endogenous (correlated with U), but in either case $Y_1 - Y_0$ is constant, at least given controls. In our model, β_{1i} may depend on unobservables and D_i may be correlated with both U_i and β_{1i}. We assume the exclusion restriction holds, in the sense that $(V_{1i}, V_{0i}, \varepsilon_i)$ or $(U_i, \beta_{1i}.\varepsilon_i)$ are independent of Z_i.

In terms of the alternative notation:

$$Y_i = \mu_0 + (Y_{1i} - Y_{0i}) D_i + V_{0i} = \mu_0 + (\mu_1 - \mu_0) D_i + [V_{0i} + (V_{1i} - V_{0i}) D_i], \qquad (72)$$

and we can write the ATE as $\bar{\beta}_1 \equiv \mu_1 - \mu_0$, and $\xi_i \equiv V_{1i} - V_{0i}$, so that $\beta_{1i} = \bar{\beta}_1 + \xi_i$. Thus:

$$\mathbb{E}[Y_i|Z_i] = \mu_0 + (\mu_1 - \mu_0)\, \mathbb{E}[D_i|Z_i] + \mathbb{E}[V_{1i} - V_{0i}|D_i = 1, Z_i]\, \mathbb{E}[D_i|Z_i]. \qquad (73)$$

If β_{1i} is mean independent of D_i, then $\mathbb{E}[V_{1i} - V_{0i}|D_i = 1, Z_i] = 0$ and:

$$\mathbb{E}[Y_i|Z_i] = \mu_0 + (\mu_1 - \mu_0)\,\mathbb{E}[D_i|Z_i], \tag{74}$$

so that $\bar{\beta}_1 = \operatorname{Cov}(Z, Y)/\operatorname{Cov}(Z, D)$, which is the IV coefficient. Otherwise, $\bar{\beta}_1$ does not coincide with the IV coefficient. A special case of mean independence of β_{1i} with respect to D_i occurs when β_{1i} is constant.

The failure of IV can be seen as the result of a missing variable. The model can be written as:

$$Y_i = \beta_0 + \bar{\beta}_1 D_i + \varphi(Z_i)D_i + \zeta_i, \tag{75}$$

where $\varphi(Z_i) \equiv E[V_{1i} - V_{0i}|D_i = 1, Z_i]$. Note that $\mathbb{E}[\zeta_i|Z_i] = 0$. When we are doing IV estimation we are not taking into account the variable $\varphi(Z_i)D_i$. In the example of returns to education, $\varphi(z)$ is the average excess return for college-educated people with $Z_i = z$. If $Z_i = 1$ if the individual leaves near to college, we would expect $\varphi(1) \leq \varphi(0)$.

The average treatment effect on the treated and the LATE are respectively:

$$\begin{aligned} \alpha_{TT} &= \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] = \bar{\beta}_1 + \mathbb{E}[V_{1i} - V_{0i}|D_i = 1] & (76)\\ \alpha_{LATE} &= \mathbb{E}[Y_{1i} - Y_{0i}|D_{1i} - D_{0i} = 1] = \bar{\beta}_1 + \mathbb{E}[V_{1i} - V_{0i}| - \gamma_0 - \gamma_1 \leq \varepsilon_i \leq -\gamma_0]. \end{aligned}$$

The model is completed with the assumption:

$$\begin{pmatrix} V_{1i} \\ V_{0i} \\ \varepsilon_i \end{pmatrix} \Bigg| Z_i \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{10} & \sigma_{1\varepsilon} \\ & \sigma_0^2 & \sigma_{0\varepsilon} \\ & & 1 \end{pmatrix}\right]. \tag{77}$$

In this case, we have a parametric likelihood model that can be estimated by ML. We can also consider a variety of two-step methods. Note that:

$$\mathbb{E}[V_{1i} - V_{0i}|D_i = 1, Z_i] = (\sigma_{1\varepsilon} - \sigma_{0\varepsilon})\lambda_i, \tag{78}$$

where $\lambda_i \equiv \lambda(\gamma_0 + \gamma_1 Z_i)$ and $\lambda(.)$ is the inverse Mills ratio, so that we can do IV estimation in:

$$Y_i = \beta_0 + \bar{\beta}_1 D_i + (\sigma_{1\varepsilon} - \sigma_{0\varepsilon})\lambda_i D_i + \zeta_i, \tag{79}$$

or OLS estimation in:

$$Y_i = \beta_0 + \bar{\beta}_1 \Phi_i + (\sigma_{1\varepsilon} - \sigma_{0\varepsilon})\phi_i + \zeta_i^*. \tag{80}$$

The current model can be regarded as the combination of two generalized selection models; therefore, the identification result for that model applies. Namely, with a continuous exclusion restriction, $\mathbb{E}[Y_{1i}|X_i]$ and $\mathbb{E}[Y_{0i}|X_i]$ are identified up to a constant (X_i denotes controls that so far we omitted for simplicity). However, the constants are important because they determine the average treatment effect of D on Y. Unfortunately, they require an identification at infinity argument.

Willis and Rosen (1979) provide an illustration for this model. In their paper, the authors propose a switching regression model, where individuals decide whether to engage in college education or stay with high school. There are determinants of the costs, like distance to college, tuition fees, availability of scholarships, opportunity costs or borrowing constraints which are potential instruments.

E. Continuous instruments: marginal treatment effects (MTE)

When the support of Z is not binary, there is a multiplicity of causal effects. Then, the question is which of these causal effects are relevant for evaluating a given policy. The natural experiment literature has been satisfied with identifying "causal effects" in a a broad sense, without paying much attention to their relevance. But the reality is that some causal effects are more informative than others.

If Z is continuous, we can define a different LATE parameter for every pair (z, z'):

$$\alpha_{LATE}(z, z') \equiv \frac{\mathbb{E}[Y|Z = z] - \mathbb{E}[Y|Z = z']}{\mathbb{E}[D|Z = z] - \mathbb{E}[D|Z = z']}. \tag{81}$$

The multiplicity is even higher when there is more than one instrument.

For a general instrument vector Z, there are as many potential treatment status indicators D_z as possible values z of the instrument. The IV assumptions become:

$$\begin{array}{ll} (Y_0, Y_1, D_z) \perp Z & \text{(independence)} \\ P(D = 1|Z = z) \equiv P(z) \text{ is a nontrivial function of } z & \text{(relevance)}\ . \end{array} \tag{82}$$

The monotonicity assumption for general Z can be expressed as follows. For any pair of values (z, z'), all units in the population satisfy either:

$$D_{zi} \geq D_{z'i} \text{ or } D_{zi} \leq D_{z'i}. \tag{83}$$

Alternatively, we can postulate an index model for D_z:

$$D_z = \mathbb{1}\{\mu(z) - U > 0\} \text{ and } U \perp Z, \tag{84}$$

which can be a useful way of organizing different LATEs (Heckman and Vytlacil, 2005). Note that the observed D is $D = D_z$. Monotonicity and index model assumptions are equivalent (Vytlacil, 2002). This result connects LATE thinking with econometric selection models. Without loss of generality, we can set $\mu(z) = P(z)$, and take U as uniformly distributed in the $(0, 1)$ interval. To see this, note that:

$$\mathbb{1}\{\mu(z) > U\} = \mathbb{1}\{F_U(\mu(z)) > F_U(U)\} = \mathbb{1}\{P(z) > \widetilde{U}\}, \tag{85}$$

where $\widetilde{U}$ is uniformly distributed.

To connect with the earlier discussion, if Z is a 0-1 scalar instrument, there are only two values of the propensity score $P(0)$ and $P(1)$. Suppose that $P(0) < P(1)$. Always-takers have $U < P(0)$, compliers have U between $P(0)$ and $P(1)$, and never-takers have $U > P(1)$. A similar argument can be made for any pair (z, z') in the case of a general Z. Therefore, under monotonicity we can always invoke an index equation and imagine each member of the population as having a particular value of the unobserved variable U.

Using the propensity score $P(Z) \equiv P(D = 1|Z)$ as instrument, LATE becomes:

$$\alpha_{LATE}(P(z), P(z')) = \frac{\mathbb{E}[Y|P(Z) = P(z)] - \mathbb{E}[Y|P(Z) = P(z')]}{P(z) - P(z')}. \tag{86}$$

If Z is binary, this is equivalent to what we had in the first place, but if Z is continuous, taking limits as $z \to z'$, we get a limiting form of LATE, which we refer to as ***marginal treatment effect*** (MTE):

$$\alpha_{MTE}(P(z)) = \frac{\partial\, \mathbb{E}[Y|P(Z) = P(z)]}{\partial P(z)}. \tag{87}$$

$\alpha_{LATE}(P(z), P(z'))$ gives the ATE for individuals who would change schooling status from changing $P(Z)$ from $P(z)$ to $P(z')$:

$$\alpha_{LATE}(P(z), P(z')) = \mathbb{E}[Y_1 - Y_0|P(z') < U < P(z)]. \tag{88}$$

Similarly, $\alpha_{MTE}(P(z))$ gives the ATE for individuals who would change schooling status following a marginal change in $P(z)$ or, in other words, who are indifferent between schooling choices at $P(Z) = P(z)$. Using the error term in the index model, we can say that:

$$\alpha_{MTE}(P(z)) = \mathbb{E}[Y_1 - Y_0|U = P(z)]. \tag{89}$$

Integrating $\alpha_{MTE}(U)$ over different ranges of U we can get other ATE measures. For example:

$$\alpha_{LATE}(P(z), P(z')) = \frac{\displaystyle\int_{P(z')}^{P(z)} \alpha_{MTE}(u)du}{P(z) - P(z')}, \tag{90}$$

and:

$$\alpha_{ATE} = \int_0^1 \alpha_{MTE}(u)du, \tag{91}$$

which makes it clear that to be able to identify α_{ATE} we need identification of $\alpha_{MTE}(u)$ over the entire $(0,1)$ range.

Constructing suitably integrated marginal treatment effects, it may be possible to identify policy relevant treatment effects. LATE gives the per capita effect of the policy in those induced to change by the policy when the instrument is precisely an indicator of the policy change (e.g. policies that change college fees or distance to school, under the assumption that the policy change affects the probability of participation but not the gain itself).

Heckman and Vytlacil (2005) suggest to estimate MTE by estimating the derivative of the conditional mean $\mathbb{E}[Y|P(Z) = P(z), X = x]$ using kernel-based local linear regression techniques. Hence, in this context, the propensity score plays a very different role to matching.

Homogeneity (or absence of self-selection) can be tested by testing linearity of the conditional mean outcome on the propensity score. To see it, use $Y = Y_0 + (Y_1 - Y_0)D$ and write:

$$\begin{aligned}\mathbb{E}[Y|P(Z)] &= \mathbb{E}[Y_0|P(Z)] + \mathbb{E}[(Y_1 - Y_0)D|P(Z)] \\ &= \mathbb{E}[Y_0|P(Z)] + \mathbb{E}[Y_1 - Y_0|P(Z), D = 1]P(Z). \end{aligned} \tag{92}$$

The quantity $\mathbb{E}[Y_1 - Y_0|P(Z), D = 1]$ is constant under homogeneity, so that the conditional mean is linear in $P(Z)$.

F. *Some remarks about unobserved heterogeneity in IV settings*

Applied researchers are often concerned about the implications of unobserved heterogeneity. The balance between observed and unobserved heterogeneity depends on how detailed information on agents is available, which ultimately is an empirical issue. The worry for IV-based identification of treatment effects is not heterogeneity *per se*, but the fact that heterogeneous gains may affect program participation.

In the absence of an economic model or a clear notional experiment, it is often difficult to interpret what IV estimates estimate. Knowing that IV estimates can be interpreted as averages of heterogeneous effects is not very useful if understanding the heterogeneity itself is first order. This is clearly a drawback of the approach.

Heterogeneity of treatments may be more important. For example, the literature has found significant differences in returns to different college majors. A problem of aggregating educational categories is that returns are less meaningful. Sometimes education outcomes are aggregated into just two categories, because some techniques are only well developed for binary explanatory variables. A methodological emphasis may offer new opportunities but also impose constraints.

G. Some examples

Example 1: Non-compliance in randomized trials. In a classic example, Z indicates assignments to treatment in an experimental design. Therefore, $(Y_0, Y_1) \perp Z$. However, the "actual treatment" D differs from Z because some individuals in the treatment group decide not to treat (non-compliers). Z and D will be correlated by construction. We mentioned this example to illustrate the plausibility of the eligibility rule that allows us to identify α_{TT}.

Assignment to treatment is not a valid instrument in the presence of externalities that benefit members of the treatment group, even if they are not treated themselves. In such case, the exclusion restriction fails to hold. An example of this situation arises in a study of the effect of deworming on school participation in Kenya using school-level randomization (Miguel and Kremer, 2004).

Example 2: Ethnic Enclaves and Immigrant Outcomes. Edin, Fredriksson and Åslund (2003) are interested in the effect of living in highly concentrated ethnic area on labor success. In Sweden, 11% of the population was born abroad. Of those, more than 40% live in an ethnic enclave. The question is, then, whether they perform worse than the other immigrants because they live in an enclave.

The causal effect is ambiguous *ex-ante*. Residential segregation lowers the acquisition rate of local skills, preventing access to good jobs, but enclaves act as opportunity-increasing networks by disseminating information to new immigrants.

Immigrants in ethnic enclaves have 5% lower earnings, after controlling for age, education, gender, family background, country of origin, and year of immigration. But this association may not be causal if the decision to live in an enclave depends on expected opportunities. As a result, the authors use an exogenous source of variation as an instrument. Motivated by the belief that dispersing immigrants promotes integration, Swedish governments of 1985-1991 assigned initial areas of residence to refugee immigrants. Let Z indicate initial assignment (8 years before measuring the ethnic enclave indicator D). Edin, Fredriksson and Åslund (2003) assume that Z is independent of potential earnings Y_0 and Y_1. IV estimates imply a 13% gain for low-skill immigrants associated with one standard deviation increase in ethnic concentration. For high-skill immigrants, there was no effect.

Example 3: Vietnam veterans and civilian earnings. Did military service in Vietnam have a negative effect on earnings? This is the question analyzed by Angrist (1990). In this example, he uses draft lottery eligibility as the instrumental variable, Veteran status as treatment variable, and log earnings as the outcome. He uses administrative records for 11,637 white men born in 1950-1953 linked with March CPS of 1979 and 1981-1985.

This lottery was conducted annually during 1970-1974. It assigned numbers from 1 to 365 to dates of birth in the cohorts being drafted. Men with lowest numbers were called up to a ceiling determined every year by the department of defense. The fact that draft eligibility affected the probability of enrollment along with its random nature makes this variable a good candidate to instrument "veteran status". The need for instrumentation is because there was a strong selection process in the military during the Vietnam period: some volunteered, while others avoided enrollment using student or job deferments. Presumably, enrollment was influenced by future potential earnings.

VI. Regression Discontinuity (RD)

A. *The fundamental RD assumption*

In the matching context we make the conditional exogeneity assumption $(Y_1, Y_0) \perp D|X$ whereas in the IV context we assume $(Y_1, Y_0) \perp Z|X$ (orthogonality of the instrument) and $D \not\perp Z|X$ (relevance). The relevance condition can also be expressed as saying that for some $z \neq z'$, the following inequality is satisfied: $P(D = 1|Z = z) \neq P(D = 1|Z = z')$. In regression discontinuity we consider a situation where there is a continuous variable Z that is not necessarily a valid instrument (it does not satisfy the exogeneity assumption), but such that treatment assignment is a discontinuous function of Z. The basic asymmetry on which identification rests is discontinuity in the dependence of D on Z but continuity in the dependence of (Y_1, Y_0) on Z. RD methods have much potential in economic applications because geographic boundaries or program rules (e.g. eligibility thresholds) often create usable discontinuities.

More formally, discontinuity in treatment assignment but continuity in potential outcomes means that there is at least a known value $z = z_0$ such that:

$$\lim_{z \to z_0^+} P(D = 1|Z = z) \neq \lim_{\to z_0^-} P(D = 1|Z = z) \tag{93}$$

$$\lim_{z \to z_0^+} P(Y_j \leq r|Z = z) = \lim_{z \to z_0^-} P(Y_j \leq r|Z = z) \quad (j = 0, 1) \tag{94}$$

Implicit regularity conditions are: (i) the existence of the limits, and (ii) that Z has positive density in a neighborhood of z_0. We abstract from conditioning covariates for the time being for simplicity.

Early RD literature in Psychology (Cook and Campbell, 1979) distinguishes between ***sharp*** and ***fuzzy*** designs. In the former, D is a deterministic function of Z:

$$D = \mathbb{1}\{Z \geq z_0\}, \tag{95}$$

whereas in the latter is not. The sharp design can be regarded as a special case of the fuzzy design, but one that has different implications for identification of treatment effects. In the sharp design:

$$\begin{aligned} \lim_{z \to z_0^+} \mathbb{E}[D|Z = z] = 1 \\ \lim_{z \to z_0^-} \mathbb{E}[D|Z = z] = 0. \end{aligned} \tag{96}$$

B. Homogeneous treatment effects

Like in the IV setting, the case of homogeneous treatment effects is useful to present the basic RD estimator. Suppose that $\alpha = Y_1 - Y_0$ is constant, so that:

$$Y_i = \alpha D_i + Y_{0i} \tag{97}$$

Taking conditional expectations given $Z = z$ and left- and right-side limits:

$$\begin{aligned} \lim_{z \to z_0^+} \mathbb{E}[Y|Z = z] = \alpha \lim_{z \to z_0^+} \mathbb{E}[D|Z = z] + \lim_{z \to z_0^+} \mathbb{E}[Y_0|Z = z] \\ \lim_{z \to z_0^-} \mathbb{E}[Y|Z = z] = \alpha \lim_{z \to z_0^-} \mathbb{E}[D|Z = z] + \lim_{z \to z_0^-} \mathbb{E}[Y_0|Z = z], \end{aligned} \tag{98}$$

which leads to the consideration of the following RD parameter:

$$\alpha_{RD} = \frac{\lim_{z \to z_0^+} \mathbb{E}[Y|Z = z] - \lim_{z \to z_0^-} \mathbb{E}[Y|Z = z]}{\lim_{z \to z_0^+} \mathbb{E}[D|Z = z] - \lim_{z \to z_0^-} \mathbb{E}[D|Z = z]}, \tag{99}$$

which is determined provided the relevance condition in Equation (93) is satisfied, and equals α provided the independence condition in Equation (94) holds.

In the case of a sharp design, the denominator is unity so that:

$$\alpha_{RD} = \lim_{z \to z_0^+} \mathbb{E}[Y|Z = z] - \lim_{z \to z_0^-} \mathbb{E}[Y|Z = z], \tag{100}$$

which can be regarded as a matching-type situation, in the same way that the general case can be regarded as an IV-type situation. So the basic idea is to obtain a treatment effect by comparing the average outcome left of the discontinuity with the average outcome to the right of discontinuity, relative to the difference between the left and right propensity scores. Intuitively, considering units within a small interval around the cutoff point is similar to a randomized experiment at the cutoff point.

C. Heterogeneous treatment effects

Now suppose that:

$$Y_i = \alpha_i D_i + Y_{0i} \tag{101}$$

In the sharp design since $D = \mathbb{1}\{Z \geq z_0\}$ we have:

$$\mathbb{E}[Y|Z = z] = \mathbb{E}[\alpha|Z = z]\, \mathbb{1}\{z \geq z_0\} + \mathbb{E}[Y_0|Z = z]. \tag{102}$$

Therefore, the situation is one of selection on observables. That is, letting:

$$k(z) \equiv \mathbb{E}[Y_0|Z = z] + \{\mathbb{E}[\alpha|Z = z] - \mathbb{E}[\alpha|Z = z_0]\}\, \mathbb{1}\{z \geq z_0\} \tag{103}$$

we have:

$$\mathbb{E}[Y|Z = z] = \mathbb{E}[\alpha|Z = z_0]\, \mathbb{1}\{z \geq z_0\} + k(z) \tag{104}$$

where $k(z)$ is continuous at $z = z_0$. Therefore, the OLS population coefficient on D in the equation:

$$Y = \alpha_{RD} D + k(z) + w \tag{105}$$

coincides with α_{RD}, which in turn equals $\mathbb{E}[\alpha|Z = z_0]$. The control function $k(z)$ is nonparametrically identified. To see this, first note that α_{RD} is identified from Equation (102). Then $k(z)$ is identifiable as the nonparametric regression $\mathbb{E}[Y - \alpha_{RD} D|Z = z]$. Note that if the treatment effect is homogeneous, $k(z)$ coincides with $\mathbb{E}[Y_0|Z = z]$, but not in general.

If $\mu(z) \equiv \mathbb{E}[Y_0|Z = z]$ was known (e.g. using data from a setting in which no program was present) then we could consider a regression of Y on D and $\mu(z)$. It turns out that the coefficient on D in such a regression is $\mathbb{E}[\alpha|z \geq z_0]$.

In the fuzzy design, D not only depends on $\mathbb{1}\{Z \geq z_0\}$, but also on other unobserved variables. Thus, D is an endogenous variable in Equation (105). However, we can still use $\mathbb{1}\{Z \geq z_0\}$ as an instrument for D in such equation to identify α_{RD}, at least in the homogeneous case. The connection between the fuzzy design and the instrumental variables perspective was first made explicit in van der Klaaw (2002).

Next, we discuss the interpretation of α_{RD} in the fuzzy design with heterogeneous treatment effects, under two different assumptions. Let us first consider the weak conditional independence assumption:

$$D \perp (Y_0, Y_1)|Z = z \tag{106}$$

for z near z_0, i.e. for $z = z_0 \pm e$, where e is an arbitrarily small positive number, or:

$$P(Y_j \leq r|D = 1, Z = z_0 \pm e) = P(Y_j \leq r|Z = z_0 \pm e) \quad (j = 0, 1). \tag{107}$$

That is, we are assuming that treatment assignment is exogenous in the neighborhood of z_0. An implication is:

$$\mathbb{E}[\alpha D|Z = z_0 \pm e] = \mathbb{E}[\alpha|Z = z_0 \pm e] \, \mathbb{E}[D|Z = z_0 \pm e]. \tag{108}$$

Proceeding as before, we have:

$$\begin{aligned} \lim_{z \to z_0^+} \mathbb{E}[Y|Z = z] &= \lim_{z \to z_0^+} \mathbb{E}[\alpha|Z = z] \, \mathbb{E}[D|Z = z] + \lim_{z \to z_0^+} \mathbb{E}[Y_0|Z = z] \\ \lim_{z \to z_0^-} \mathbb{E}[Y|Z = z] &= \lim_{z \to z_0^-} \mathbb{E}[\alpha|Z = z] \, \mathbb{E}[D|Z = z] + \lim_{z \to z_0^-} \mathbb{E}[Y_0|Z = z]. \end{aligned} \tag{109}$$

Noting that $\lim_{z \to z_0^+} \mathbb{E}[\alpha|Z = z] = \lim_{z \to z_0^-} \mathbb{E}[\alpha|Z = z] = \mathbb{E}[\alpha|Z = z_0]$, subtracting one equation from the other, and rearranging the terms we obtain:

$$\begin{aligned} \mathbb{E}[\alpha|Z = z_0] &= \mathbb{E}[Y_1 - Y_0|Z = z_0] \\ &= \frac{\lim_{z \to z_0^+} \mathbb{E}[Y|Z = z] - \lim_{z \to z_0^-} \mathbb{E}[Y|Z = z]}{\lim_{z \to z_0^+} \mathbb{E}[D|Z = z] - \lim_{z \to z_0^-} \mathbb{E}[D|Z = z]} = \alpha_{RD}. \end{aligned} \tag{110}$$

That is, the RD parameter can be interpreted as the average treatment effect at z_0.

Hahn, Todd and van der Klaaw (2001) also consider an alternative LATE-type of assumption. Let D_z be the potential assignment indicator associated with $Z = z$, and for some $\bar{\varepsilon} > 0$ and any pair $(z_0 - \varepsilon, z_0 + \varepsilon)$ with $0 < \varepsilon < \bar{\varepsilon}$ suppose the local monotonicity assumption:

$$D_{z_0+\varepsilon} \geq D_{z_0-\varepsilon} \text{ for all units in the population.} \tag{111}$$

An example is a population of cities where Z denotes voting share and D_z is an indicator of party control when $Z = z$. In this case the local conditional independence assumption could be problematic but the monotonicity assumption is not. In such case, it can be shown that α_{RD} identifies the local average treatment effect at $z = z_0$:

$$\alpha_{RD} = \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_1 - Y_0|D_{z_0+\varepsilon} - D_{z_0-\varepsilon} = 1] \tag{112}$$

that is, the ATE for the units for whom treatment changes discontinuously at z_0. If the policy is a small change in the threshold for program entry, the LATE parameter delivers the treatment effect for the subpopulation affected by the change, so that in that case it would be the parameter of policy interest.

D. Estimation strategies

There are parametric and semiparametric estimation strategies. Hahn, Todd and van der Klaaw (2001) suggested the following local estimator. Let $S_i \equiv \mathbb{1}\{z_0 - h < Z_i < z_0 + h\}$ where $h > 0$ denotes the bandwidth, and consider the subsample such that $S_i = 1$. The proposed estimator is the IV regression of Y_i on D_i using $W_i \equiv \mathbb{1}\{z_0 < Z_i < z_0 + h\}$ as an instrument, applied to the subsample with $S_i = 1$:

$$\widehat{\alpha}_{RD} = \frac{\widehat{\mathbb{E}}[Y_i|W_i = 1, S_i = 1] - \widehat{\mathbb{E}}[Y_i|W_i = 0, S_i = 1]}{\widehat{\mathbb{E}}[D_i|W_i = 1, S_i = 1] - \widehat{\mathbb{E}}[D_i|W_i = 0, S_i = 1]}. \tag{113}$$

This estimator has nevertheless a poor boundary performance. An alternative suggested by Hahn, Todd and van der Klaaw (2001) is a local linear regression method. Suppose:

$$\mathbb{E}[D|Z] = g(Z) + \delta\, \mathbb{1}\{Z \geq z_0\} \tag{114}$$

and:

$$\mathbb{E}[Y_0|Z] = k(Z). \tag{115}$$

A control function regression-based approach is based in the control function augmented equation that replaces D by the propensity score $\mathbb{E}[D|Z]$:

$$Y = \alpha_{RD}\, \mathbb{E}[D|Z] + k(Z) + w \tag{116}$$

In a parametric approach, we assume functional forms for $g(Z)$ and $k(Z)$. van der Klaaw (2002) considered a semiparametric approach using a power series approximation for $k(Z)$. If $g(Z) = k(Z)$, then we can do 2SLS using as instrumental variables $\mathbb{1}\{Z \geq z_0\}$ and $g(Z)$, where $g(Z)$ is the *included* instrument and $\mathbb{1}\{Z \geq z_0\}$ is the *excluded* instrument. These methods of estimation, which are not local to data points near the threshold, are implicitly predicated on the assumption of homogeneous treatment effects.

E. Conditioning on covariates

Even if the RD assumption is satisfied unconditionally, conditioning on covariates may mitigate the heterogeneity in treatment effects, hence contributing to the relevance of RD estimated parameters. Covariates may also make the local conditional exogeneity assumption more credible. This would also be true of within-group estimation in a panel data context (see application in Hoxby, 2000).

F. Examples

Effect of class size on test scores. Angrist and Lavy (1999) analyze the effect of class size on test scores using the "Maimonides' rule" in Israel. The Maimonides rule divides students in classes of less than a given maximum number of students (40). Maimonides' rule allows enrollment cohorts of 1-40 to be grouped in a single class, but enrollment groups of 41-80 are split into two classes of average size 20.5-40, enrollment groups of 81-120 are split into three classes of average size 27-40, etc. (in practice, the rule was not exact: class size predicted by the rule differed from actual size). Angrist and Lavy (1999) use this discontinuity to analyze the effect of class size on school outcomes. Their outcome variable is the average test score at a class i in school s, the treatment variable (not binary) is the size of class i, and the instrument is the total enrollment at the beginning of an academic year at school s.

Effect of financial aid offers on students' enrollment decisions. This is the interest of van der Klaaw (2002). His outcome of interest is the decision of student i to enroll in college a given college (binary), the treatment is the amount of financial aid offer to student i, and the instrument is the index that aggregates SAT score and high school GPA: applicants for aid were divided into four groups on the basis of the interval the index Z fell into. Average aid offers as a function of Z contained jumps at the cutoff points for the different ranks, with those scoring just below a cutoff point receiving much less on average than those who scored just above the cutoff.

VII. Differences-in-Differences (DID)

In this section, we start straight from the example. In March 1992 the state of New Jersey increased the legal minimum wage by 19%, whereas the bordering state of Pennsylvania kept it constant. Card and Krueger (1994) evaluate the effect of this change on the employment of low wage workers. In a competitive model the result of increasing the minimum wage is to reduce employment. They conducted a survey to some 400 fast food restaurants from the two states just before the NJ reform, and a second survey to the same outlets 7-8 months after. Characteristics of fast food restaurants: (i) a large source of employment for low-wage workers; (ii) they comply with minimum wage regulations (especially franchised restaurants);

(iii) fairly homogeneous job, so good measures of employment and wages can be obtained; (iv) easy to get a sample frame of franchised restaurants (yellow pages) with high response rates (response rates 87% and 73% —less in Pennsylvania, because the interviewer was less persistent).

The DID coefficient is:

$$\beta = \{\mathbb{E}[Y_2|D=1] - \mathbb{E}[Y_1|D=1]\} - \{\mathbb{E}[Y_2|D=0] - \mathbb{E}[Y_1|D=0]\}, \tag{117}$$

where Y_1 and Y_2 denote employment before and after the reform, $D = 1$ denotes a store in New Jersey (treatment group) and $D = 0$ denotes one in Pennsylvania (control group).

β measures the difference between the average employment change in New Jersey and the average employment change in Pennsylvania. The key assumption in giving a causal interpretation to β is that the temporal effect in the two states is the same in the absence of intervention.

It is possible to generalize the comparison in several ways, for example controlling for other variables. Card and Krueger found that rising the minimum wage increased employment in some of their comparisons but in no case caused an employment reduction. This article originated much economic and political debate. DID estimation has become a very popular method of obtaining causal effects, especially in the US, where the federal structure provides cross state variation in legislation.

If we observe outcomes before and after treatment, we could use the treated before treatment as controls for the treated after treatment. The problem of this comparison is that it can be contaminated by the effect of events other than the treatment that occurred between the two periods. Suppose that only a fraction of the population is exposed to treatment. In such a case, we can use the group that never receives treatment to identify the temporal variation in outcomes that is not due to exposure to treatment. This is the basic idea of the DID method.

To see identification more formally, consider the two-period potential outcomes representation with treatment in $t = 2$:

$$\begin{aligned} Y_1 &= Y_0(1) \\ Y_2 &= (1-D)Y_0(2) + DY_1(2) \end{aligned} \tag{118}$$

The fundamental identifying assumption is that the average changes in the two groups are the same in the absence of treatment:

$$\mathbb{E}[Y_0(2) - Y_0(1)|D=1] = \mathbb{E}[Y_0(2) - Y_0(1)|D=0]. \tag{119}$$

$Y_0(1)$ is always observed but $Y_0(2)$ is counterfactual for units with $D = 1$. Under such identification assumption, the DID coefficient coincides with the average treatment effect for the treated. To see this note that the DID parameter in general is equal to:

$$\begin{aligned} \beta &= \{\mathbb{E}[Y_2|D=1] - \mathbb{E}[Y_1|D=1]\} - \{\mathbb{E}[Y_2|D=0] - \mathbb{E}[Y_1|D=0]\} \qquad (120) \\ &= \{\mathbb{E}[Y_1(2)|D=1] - \mathbb{E}[Y_0(1)|D=1]\} - \{\mathbb{E}[Y_0(2)|D=0] - \mathbb{E}[Y_0(1)|D=0]\}. \end{aligned}$$

Now, adding and subtracting $\mathbb{E}[Y_0(2)|D = 1]$:

$$\beta = \mathbb{E}[Y_1(2) - Y_0(2)|D = 1] + \{\mathbb{E}[Y_0(2) - Y_0(1)|D = 1] - \mathbb{E}[Y_0(2) - Y_0(1)|D = 0]\}, \qquad (121)$$

which as long as the last term vanishes it equals:

$$\beta = \mathbb{E}[Y_1(2) - Y_0(2)|D = 1]. \qquad (122)$$

There are some relevant comments to make. β can be obtained as the coefficient of the interaction term in a regression of outcomes on treatment and time dummies. To obtain the DID parameter we do not need panel data (except if e.g. we regard the Card-Krueger data as an aggregate panel with two units and two periods), just cross-sectional data for at least two periods. With panel data, we can estimate β from a regression of outcome changes on the treatment dummy. This is convenient for accounting for dependence between the two periods.

This approach has also caveats. β is obtained from differences in averages in the two periods and two groups. If the composition of the cross-sectional populations change over time, estimates will be biased (especially problematic if not using panel data). Still, the fundamental assumption might be satisfied conditionally given certain covariates, but identification vanishes if some of them are unobservable.

VIII. Distributional Effects and Quantile Treatment Effects

A. Distributional effects. Matching

Most of the literature focused on average effects, but the results seen in previous sections also hold for distributional comparisons. First consider conditional independence. Under conditional independence, the full marginal distributions of Y_1 and Y_0 can be identified. To see this, first note that we can identify not just α_{ATE} but also $\mathbb{E}[Y_1]$ and $\mathbb{E}[Y_0]$:

$$\mathbb{E}[Y_1] = \int \mathbb{E}[Y_1|X]dF(X) = \int \mathbb{E}[Y|D = 1, X]dF(X) \qquad (123)$$

and similarly for $\mathbb{E}[Y_0]$. Next, we can equally identify the expected value of any function of the outcomes $\mathbb{E}[h(Y_1)]$ and $\mathbb{E}[h(Y_0)]$:

$$\mathbb{E}[h(Y_1)] = \int \mathbb{E}[h(Y_1)|X]dF(X) = \int \mathbb{E}[h(Y_1)|D = 1, X]dF(X) \tag{124}$$

Thus, setting $h(Y_1) = \mathbb{1}\{Y_1 \leq r\}$ we get:

$$\mathbb{E}[\mathbb{1}\{Y_1 \leq r\}] = P(Y_1 \leq r) = \int [P(Y \leq r|D = 1, X)dF(X), \tag{125}$$

and similarly for $P(Y_0 \leq r)$. Given identification of the cdfs we can also identify quantiles of Y_1 and Y_0. Quantile treatment effects are differences in the marginal quantiles of Y_1 and Y_0. More substantive objects are the joint distribution of (Y_1, Y_0) or the distribution of gains Y_1, Y_0, but their identification requires stronger assumptions.

Firpo (2007) proposes a quantile treatment effects under the matching assumptions. Let (Y_1, Y_0) be potential outcomes with marginal cdfs $F_1(r)$ and $F_0(r)$, and quantile functions $Q_{1\tau} = F_1^{-1}(\tau)$ and $Q_{0\tau} = F_0^{-1}(\tau)$. The QTE is defined to be:

$$\alpha_\tau \equiv Q_{1\tau} - Q_{0\tau} \tag{126}$$

Under conditional exogeneity $F_j(r) = P(Y \leq r|D = j, X)dG(X)$ for $j = 0, 1$. Moreover, $Q_{1\tau}$ and $Q_{0\tau}$ satisfy the moment conditions:

$$\begin{aligned} \mathbb{E}\left[\tfrac{D}{\pi(X)}\mathbb{1}\{Y \leq Q_{1\tau}\} - \tau\right] &= 0 \\ \mathbb{E}\left[\tfrac{1-D}{1-\pi(X)}\mathbb{1}\{Y \leq Q_{0\tau}\} - \tau\right] &= 0, \end{aligned} \tag{127}$$

and

$$\begin{aligned} Q_{1\tau} &= \arg\min_q \mathbb{E}\left[\tfrac{D}{\pi(X)}\rho_\tau(Y - q)\right] \\ Q_{0\tau} &= \arg\min_q \mathbb{E}\left[\tfrac{1-D}{1-\pi(X)}\rho_\tau(Y - q)\right], \end{aligned} \tag{128}$$

where $\rho_\tau(u) \equiv [\tau - \mathbb{1}\{u < 0\}]u$ is the "check" function. Firpo's method is a two-step weighting procedure in which the propensity score $\pi(X)$ is estimated in a first stage.

B. Instrumental Variables

Imbens and Rubin (1997) show that if conditional independence does not hold, but valid instruments that satisfy monotonicity are available, not only the average treatment effect for compliers is identified but also the entire marginal distributions of Y_0 and Y_1 for them. Abadie (2002) gives a simple proof that suggests a similar calculation to the one done for average treatment effects. For any function $h(.)$ consider:

$$W \equiv h(Y)D = \begin{cases} W_1 \equiv h(Y_1) & \text{if } D = 1 \\ W_0 \equiv 0 & \text{if } D = 0 \end{cases} \tag{129}$$

Because (W_1, W_0, D_1, D_0) are independent of Z, we can apply the LATE formula to W and get:

$$\mathbb{E}[W_1 - W_0 | D_1 - D_0 = 1] = \frac{\mathbb{E}[W|Z=1] - \mathbb{E}[W|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]}, \tag{130}$$

or substituting:

$$\mathbb{E}[h(Y_1)|D_1 - D_0 = 1] = \frac{\mathbb{E}[h(Y)D|Z=1] - \mathbb{E}[h(Y)D|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]}, \tag{131}$$

If we choose $h(Y) = \mathbb{1}\{Y \leq r\}$, the previous formula gives as an expression for the cdf of Y_1 for the compliers. Similarly, if we consider:

$$V \equiv h(Y)(1-D) = \begin{cases} V_1 \equiv h(Y_0) & \text{if } 1 - D = 1 \\ V_0 \equiv 0 & \text{if } 1 - D = 0 \end{cases} \tag{132}$$

then:

$$\mathbb{E}[V_1 - V_0 | D_1 - D_0 = 1] = \frac{\mathbb{E}[V|Z=1] - \mathbb{E}[V|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]}, \tag{133}$$

or:

$$\mathbb{E}[h(Y_0)|D_1 - D_0 = 1] = \frac{\mathbb{E}[h(Y)(1-D)|Z=1] - \mathbb{E}[h(Y)(1-D)|Z=0]}{\mathbb{E}[D|Z=1] - \mathbb{E}[D|Z=0]}, \tag{134}$$

from which we obtain the cdf of Y_0 for the compliers setting $h(Y) = \mathbb{1}\{Y \leq r\}$.

To see the intuition, suppose D is exogenous ($Z = D$), then the cdf of $Y|D = 0$ coincides with the cdf of Y_0, and the cdf of $Y|D = 1$ coincides with the cdf of Y_1. If we regress $h(Y)D$ on D, the OLS regression coefficient is:

$$\mathbb{E}[h(Y)D|D=1] - \mathbb{E}[h(Y)D|D=0] = E[h(Y_1)], \tag{135}$$

which for $h(Y) = \mathbb{1}\{Y \leq r\}$ gives us the cdf of Y_1. Similarly, if we regress $h(Y)(1-D)$ on $(1-D)$, the regression coefficient is:

$$\mathbb{E}[h(Y)(1-D)|1-D=1] - \mathbb{E}[h(Y)(1-D)|1-D=0] = E[h(Y_0)]. \tag{136}$$

In the IV case, we are running similar IV (instead of OLS) regressions using Z as instrument and getting expected $h(Y_1)$ and $h(Y_0)$ for compliers.

What does this parameter tell us? Consider, again the college example. We want to disentangle what is the effect of increasing college on the distribution of wages. Using again a binary indicator of distance as an instrument, our quantile comparison of interest is not between the distribution of wages for individuals who effectively attended college and the one of individuals who did not, but, instead, the distribution of wages of individuals that went to college because it was close to home, but that would have not gone if it had been further away, and that of individuals that did not go because it was far, but would have gone if it had been close. This comparison is what we can identify using this instrument.

Abadie (2003) suggests a weighting procedure that is useful to estimate IV quantile treatment effects. If our instrument Z satisfies the standard assumptions given X, for any measurable function of (Y, X, D) with finite expectation, $h(Y, X, D)$:

$$\mathbb{E}[h(Y, X, D)|D_1 - D_0 = 1] = \frac{\mathbb{E}[\kappa h(Y, X, D)]}{\mathbb{E}[\kappa]}, \tag{137}$$

where:

$$\kappa = 1 - \frac{D(1 - Z)}{1 - P(Z = 1|X)} - \frac{(1 - D)Z}{P(Z = 1|X)}. \tag{138}$$

The main idea is that the operator κ “finds compliers”, given that:

$$\mathbb{E}[\kappa|Y, X, D] = \Pr(D_1 - D_0 = 1|Y, X, D). \tag{139}$$

The intuition behind this is that individuals with $D(1 - Z) = 1$ are *always-takers* as $D_0 = 1$ for them; similarly, individuals with $(1 - D)Z = 1$ are *never-takers*, as $D_1 = 0$ for them; hence, the left-out are the compliers.

Given this result, Abadie et al. (2002) developed the IV quantile treatment effects estimator as the sample analog to:

$$(\alpha_\tau, Q_{0\tau}) = \arg\min_{(a,q)} \mathbb{E}\left[\rho_\tau(Y - aD - q)|D_1 - D_0 = 1\right] = \arg\min_{(a,q)} \mathbb{E}\left[\kappa\rho_\tau(Y - aD - q)\right] \tag{140}$$

There are several aspects that worth a mention. First is that κ needs to be estimated (and standard errors should take this into account —bootstrapped standard errors (including the estimation of κ in the bootstrapping) will). Second, κ is negative when $D \neq Z$ (instead of zero), which makes the regression minimand non-convex. To solve this problem, we can apply the law of iterated expectations, so that we transform the problem into:

$$(\alpha_\tau, Q_{0\tau}) = \arg\min_{(a,q)} \mathbb{E}\left[\mathbb{E}[\kappa|Y, X, D]\rho_\tau(Y - aD - q)\right] \tag{141}$$

this solves the problem provided that $\mathbb{E}[\kappa|Y,X,D] = P(D_1 - D_0 = 1|Y,X,D)$ is a probability and, hence, lies between zero and one.

This later trick makes indeed the problem very easy to implement in practice. Note that:

$$\mathbb{E}[\kappa|Y,X,D] = 1 - \frac{D(1-\mathbb{E}[Z|Y,X,D=1])}{1-P(Z=1|X)} - \frac{(1-D)\,\mathbb{E}[Z|Y,X,D=0]}{\Pr(Z=1|X)}. \tag{142}$$

A very simple two-stage method consists of the following two steps:

1) Estimate $\mathbb{E}[Z_i|Y_i,X_i,D_i]$ with a Probit of Z_i on Y_i and X_i separately for $D_i = 0$ and $D_i = 1$ subsamples, and $P(Z_i = 1|X_i)$ with a Probit of Z_i on X_i with the whole sample. Construct $\hat{\mathbb{E}}[\kappa_i|Y_i,X_i,D_i]$ using the fitted values from the previous expressions.[1]

2) Estimate the quantile regression model with the standard procedure using these predicted kappas as weights.

One should then compute the correct standard errors taking into account that the weights are estimated instead of the true weights as we discussed above.

C. Regression discontinuity

For some function $h(.)$, consider the outcome:

$$W \equiv h(Y)D = \begin{cases} W_1 \equiv h(Y_1) & \text{if } D = 1 \\ W_0 \equiv 0 & \text{if } D = 0 \end{cases} \tag{143}$$

as defined above. Using $h(Y) = \mathbb{1}\{Y \leq r\}$ the RD parameter for the outcome $W(r) = \mathbb{1}\{Y \leq r\}D$ delivers:

$$P(Y_1 \leq r|Z = z_0) = \frac{\lim\limits_{z \to z_0^+} \mathbb{E}[W(r)|Z=z] - \lim\limits_{z \to z_0^-} \mathbb{E}[W(r)|Z=z]}{\lim\limits_{z \to z_0^+} \mathbb{E}[D|Z=z] - \lim\limits_{z \to z_0^-} \mathbb{E}[D|Z=z]}, \tag{144}$$

under the local conditional independence assumption. A similar strategy can be followed to obtain $P(Y_0 \leq r|Z = z_0)$. In that case we consider:

$$V \equiv h(Y)(1-D) = \begin{cases} V_1 \equiv h(Y_0) & \text{if } 1 - D = 1 \\ V_0 \equiv 0 & \text{if } 1 - D = 0, \end{cases} \tag{145}$$

[1] It may happen that, for some observations, the predicted value goes below 0 or above 1; in this case, replace the values below 0 by 0 and the values above 1 by 1.

and the RD parameter for the outcome $V(r) = \mathbb{1}\{Y \leq r\}(1 - D)$ delivers:

$$P(Y_0 \leq r|Z = z_0) = \frac{\lim\limits_{z \to z_0^+} \mathbb{E}[V(r)|Z = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[V(r)|Z = z]}{\lim\limits_{z \to z_0^+} \mathbb{E}[D|Z = z] - \lim\limits_{z \to z_0^-} \mathbb{E}[D|Z = z]}, \tag{146}$$

Comparing the two, we obtain the distributional treatment effects.

IX. A Bridge Between Structural and Reduced Form Methods

A. *Ex-post and ex-ante policy evaluation*

Ex-post policy evaluation happens after the policy has been implemented. This is the context in which we typically implement the techniques seen so far in the course. Specifically, the evaluation makes use of existing policy variation; experimental and non-experimental methods are used. *Ex-ante* evaluation concerns interventions which have not taken place. These include treatment levels outside those in the range of existing programs, other modifications to existing programs, or new programs altogether. *Ex-ante* evaluation requires an extrapolation from (i) existing policy or (ii) policy-relevant variation. Extrapolation requires a model (structural or nonstructural). In this section we draw from Todd and Wolpin (2006), Wolpin (2007), and Chetty (2009).

Consider the *ex-ante* evaluation of a school attendance subsidy program in a development country. Consider the following two possible situations:

1) School tuition p varies exogenously across countries in the range $(\underline{p}, \bar{p})$.

2) Schools are free: $p = 0$.

In case 1) it is possible to estimate a relationship between school attendance s and tuition cost p, but in case 2) it is not. Suppose that s also depends on a set of observed factors X and it is possible to estimate non-parametrically:

$$s = f(p, X) + v. \tag{147}$$

Then it is possible to estimate the effect of the subsidy b on s for all households i in which tuition net of the subsidy $p_i - b$ is in the support of p. Because some values of net tuition must be outside of the support, it is not possible to estimate the entire response function, or to obtain population estimates of the impact of the subsidy in the absence of a parametric assumption.

In case 2), we need to look at the opportunity cost of school. Consider a household with one child making a decision about whether to send the child to school or to work. Suppose the household chooses to have the child attend school ($s = 1$) if w is below some reservation wage w^*, where w^* represents the utility gain for the household if the child goes to school:

$$w < w^* \tag{148}$$

If $w^* \sim \mathcal{N}(\alpha, \sigma^2)$, we get a standard probit model:

$$P(s = 1) = 1 - P(w^* < w) = \Phi\left(\frac{\alpha - w}{\sigma}\right) \tag{149}$$

To obtain separate estimates of α and σ we need to observe child wage offers (not only the wages of children who work). Under the school subsidy the child goes to school if $w < w^* + b$ so that the probability that a child attends school will increase by:

$$\Phi\left(\frac{b + \alpha - w}{\sigma}\right) - \Phi\left(\frac{\alpha - w}{\sigma}\right). \tag{150}$$

The conclusion is that variation in the opportunity cost of attending school (the child market wage) serves as a substitute for variation in the tuition cost of schooling.

B. Combining experiments and structural estimation

In an influential paper, Todd and Wolpin (2006) evaluate the effects of the PROGRESA school subsidy program on schooling of girls in rural Mexico. The Mexican government conducted a randomized social experiment between 1997 and 1999, in which 506 rural villages were randomly assigned to either treatment (320) or control groups (186). Parents of eligible treatment households were offered substantial payments contingent on their children's regular attendance at school. The benefit levels represented about 1/4 of average family income. The subsidy increased with grade level up to grade 9 (age 15). Eligibility was determined on the basis of a poverty index.

Experimental treatment effects on school attendance rates one year after the program showed large gains, ranging from about 5 to 15 percentage points depending on age and sex. These effects, however, assessed the impact only of the particular subsidy that was implemented. From the PROGRESA experiment alone it is not possible to determine the size and structure of the subsidy that achieves the policy goals at the lowest cost, or to assess alternative policy tools to achieve the same goals.

Todd and Wolpin use a structural model of parental fertility and schooling choices to compare the efficacy of the PROGRESA program with that of alternative policies that were not implemented. They estimate the model using control households only, exploiting child wage variation and, in particular, distance to the nearest big city for identification. They use the treatment sample for model validation and presumably also for model selection. The model specifies choice rules to determine pregnancies and school choices of parents for their children from the beginning of marriage throughout mother's fertile period and children until aged 15. These rules come from intertemporal expected utility maximization. Parents are uncertain about future income (both their own and their children) and their own future preferences for schooling.

The response functions lack a closed form expression, so that the model needs to be solved numerically. They estimate the model by maximum likelihood. The model is further complicated by including unobserved household heterogeneity (discrete types). The downside of their model is the numerical complication. The advantage is the interpretability of its components, even if some of them may be unrealistic such as the specification of household uncertainty.

They emphasize that social experiments provide an opportunity for out-of-sample validation of models that involve extrapolation outside the range of existing policy variation. This is true of both structural and nonstructural estimation.

C. Model selection: data mining and Bayesian estimation

Once the researcher has estimated a model, she can perform diagnostics, like tests of model fit and tests of overidentifying restrictions. If the model does not provide a good fit, the researcher will change the model in the directions in which the model poorly fits the data. Formal methods of model selection are no longer applicable because the model is the result of repeated pretesting. Estimating a fixed set of models and employing a model selection criterion (like AIC) is also unlikely to help because models that result from repeated pretesting will tend to be very similar in terms of model fit.

Imagine a policy maker concerned on how best to use the data (experimental program data on control and treatment households) for an ex ante policy evaluation. The policy maker selects several researchers, each of whose task is to develop a model for ex ante evaluation. One possibility is to give the researcher all the data. The other possibility is to hold out the post-program treatment households, so that the researcher only has access to control households. Is there any gain in holding out the data on the treated households? That is, is there a gain that compensates for the information loss from estimating the model on a smaller sample with less variation?

The problem is that after all the pre-testing associated with model building it is not a viable strategy to try to discriminate among models on the basis of within-sample fit because all the models are more or less indistinguishable. So we need some other criterion for judging the relative success of a model. One is assessing a model's predictive accuracy for a hold out sample.

Weighting models on the basis of posterior model probabilities in a Bayesian framework in principle seems the way to go because posterior model probabilities carry an automatic penalty for overfitting. This is proposed by Schorfheide and Wolpin (2012). The odd posterior ratio between two models is given by the odd prior ratio times the likelihood ratio:

$$\frac{P(M_j|y)}{P(M_\ell|y)} = \frac{P(M_j)f(y|M_j)}{P(M_\ell)f(y|M_\ell)}, \tag{151}$$

where $f(y|M_j) = \int f(y|\theta_j, M_j)\pi(\theta_j|M_j)d\theta_j$. The Schwarz approximation to the marginal ratio contains a correction factor for the difference in the number of parameters:

$$\frac{f(y|M_j)}{f(y|M_\ell)} \approx \frac{f(y|\widehat{\theta}_j, M_j)}{f(y|\widehat{\theta}_j, M_j)} \times N^{-[\dim(\theta_j)-\dim(\theta_\ell)]/2}. \tag{152}$$

The overall posterior distribution of a treatment effect or predictor Δ is:

$$P(\Delta|y) = \sum_j P(\Delta|y, M_j)P(M_j|y) \tag{153}$$

where $P(\Delta|y, M_j)$ is the posterior density of Δ calculated under model M_j.

From a Bayesian perspective the use of holdout samples is suboptimal because the computation of posterior probabilities should be based on the entire sample y and not just a subsample. Schorfheide and Wolpin argue that the problem with the Bayesian perspective is that the set of models under consideration is not only incomplete but that the collection of models that are analyzed is data dependent. That is, the researcher will start with some model, inspect the data, reformulate the model, consider alternative models based on the previous data inspection and so on. This is a process of data mining (e.g. the Smets and Wouters (2007) DSGE model widely used in macro policy evaluation).

The problem with such data mining is the prior distribution is shifted towards models that fit the data well whereas other models that fit slightly worse are forgotten. So these data dependent priors produce marginal likelihoods that (i) overstate the fit of the reported model and also (ii) the posterior distribution understates the parameter uncertainty. There is no viable commitment from the modelers not to look at data that are stored on their computers!

Schorfheide and Wolpin (2012, 2014) develop a principal-agent framework to address this trade-off. Data mining generates an impediment for the implementation of the ideal Bayesian analysis. In their analysis there is a policy maker (the principal) and two modelers (the agents). The modelers can each fit a structural model to whatever data they get from the policy maker and provide predictions of the treatment effect. The modelers are rewarded based on the fit of the model that they are reporting. So they have an incentive to engage in data mining.

In the context of a holdout sample, modelers are asked by the policy maker to predict features of the sample that is held out for model evaluation. If the modelers are rewarded such that their payoff is proportional to the log of the reported predictive density for Δ, then they have an incentive to reveal their subjective beliefs truthfully (i.e. to report the posterior density of Δ given their model and the data available to them). They provide a formal rationale for holding out samples in situations where the policy maker is unable to implement the full Bayesian analysis.

D. Sufficient statistics

There is a third alternative to evaluate public policies: the ***sufficient statistics*** approach, reviewed in Chetty (2009). This approach provides a middle ground between more structural and reduced form approaches. These papers develop sufficient-statistic formulas that combine the advantages of reduced-form empirics (transparent and credible identification) with an important advantage of structural models (the ability to make precise statements about welfare). The idea is to derive formulas for the welfare consequences of policies that are functions of high-level elasticities rather than deep primitives. Even though there are multiple combinations of primitives that are consistent with the inputs to the formulas, all these combinations have the same welfare implications (Chetty, 2009).

For example, in our school subsidy example, it can be that the increase in welfare produced by the subsidy can be expressed purely in terms of the elasticity of school enrollment to changes in tuition (and maybe some other elasticity), despite the fact that the subsidy can affect later decisions of the individual in terms of future education and employment, as well as parent's fertility decisions.

Provided that the program-evaluation estimates can provide a value to these elasticities, this approach allows to give economic meaning to what might otherwise be viewed as atheoretical statistical estimates.

X. Concluding Remarks

Empirical papers have become more central to economics than they used to. This reflects the new possibilities afforded by technical change in research and is a sign of scientific maturity of Economics. In an empirical paper the econometric strategy is often paramount, i.e. what aspects of data to look at and how to interpret them. This typically requires a good understanding of both relevant theory and sources of variation in data. Once this is done there is usually a more or less obvious estimation method available and ways of assessing statistical error.

Statistical issues like quality of large sample approximations or measurement error may or may not matter much in a particular problem, but a characteristic of a good empirical paper is the ability to focus on the econometric problems that matter for the question at hand. The quasi-experimental approach is also having a contribution to reshaping structural econometric practice. It is increasingly becoming standard fare a reporting style that distinguishes clearly the roles of theory and data in getting the results.

Experimental and quasi-experimental approaches have an important but limited role to play in policy evaluation. There are relevant quantitative policy questions that cannot be answered without the help of economic theory. In Applied Microeconomics there has been a lot of excitement in recent years in empirically establishing causal impacts of interventions (from field and natural experiments and the like). This is understandable because in principle causal impacts are more useful for policy than correlations. However, there is an increasing awareness of the limitations due to heterogeneity of responses and interactions and dynamic feedback. Addressing these matters require more theory. A good thing of the treatment effect literature is that it has substantially raised the empirical credibility hurdle.

A challenge for the coming years is to have more theory-based or structural empirical models that are structural not just because the author has written down the model as derived from a utility function but because he/she has been able to establish empirically invariance to a particular class of interventions, which therefore lends credibility to the model for ex ante policy evaluation within this class.

REFERENCES

Abadie, Alberto, "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, March 2002, *97* (457), 284–292.

_ , "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, April 2003, *113* (2), 231–263.

_ **, Joshua D. Angrist, and Guido W. Imbens**, "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, January 2002, *70* (1), 91–117.

Angrist, Joshua D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, June 1990, *80* (3), 313–336.

_ **and Victor Lavy**, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, May 1999, *114* (2), 533–575.

Card, David E. and Alan B. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, September 1994, *84* (4), 772–293.

Chetty, Raj, "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Review of Economics*, September 2009, *1* (1), 451–488.

Cook, Thomas D. and Donald T. Campbell, *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Chicago: Rand McNally College Pub. Co., 1979.

Dearden, Lorraine, Carl Emmerson, Christine Frayne, and Costas Meghir, "Conditional Cash Transfer and School Dropout Rates," *Journal of Human Resources*, Fall 2009, *44* (4), 827–857.

Edin, Per-Anders, Peter Fredriksson, and Olof Åslund, "Ethnic Enclaves and the Economic Success of Immigrants — Evidence from a Natural Experiment," *Quarterly Journal of Economics*, February 2003, *118* (1), 329–357.

Firpo, Sergio, "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, January 2007, *75* (1), 259–276.

Frölich, Markus, *Program Evaluation and Treatment Choice*, Berlin-Heidelberg: Springer-Verlag, 2003.

Hahn, Jinyong, Petra E. Todd, and Wilbert van der Klaaw, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, January 2001, *69* (1), 201–209.

Ham, John C. and Robert J. LaLonde, "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training," *Econometrica*, January 1996, *64* (1), 175–205.

Heckman, James J. and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, May 2005, *73* (3), 669–738.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, July 2003, *71* (4), 1161–1189.

Hoxby, Caroline M., "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, November 2000, *115* (4), 1239–1285.

Imbens, Guido W. and Donald B. Rubin, "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, October 1997, *64* (4), 555–574.

_ **and Joshua D. Angrist**, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, March 1994, *62* (2), 467–475.

Lucas, Robert E., "Econometric Policy Evaluation: A Critique," *Carnegie-Rochester Conference Series on Public Policy*, 1976, *1*, 16–46.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, January 2004, *72* (1), 159–217.

Moffitt, Robert A., *Means-Tested Transfer PProgram in the United States*, Chicago: The University of Chicago Press, 2003.

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

Schorfheide, Frank and Kenneth I. Wolpin, "On the Use of Houldout Samples for Model Selection," *American Economic Review*, May 2012, *102* (3), 477–481.

_ **and** _ , "To Hold Out or Not to Hold Out," NBER Working Paper N. 16565, July 2014.

Smets, Frank and Rafael Wouters, "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, June 2007, *97* (3), 586–606.

Todd, Petra E. and Kenneth I. Wolpin, "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," *American Economic Review*, December 2006, *96* (5), 1384–1417.

_ **and** _ , "Structural Estimation and Policy Evaluation in Developing Countries," *Annual Review of Economics*, September 2010, *2* (1), 21–50.

van der Klaauw, Wilbert, "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," *International Economic Review*, November 2002, *43* (4), 1249–1287.

Vytlacil, Edward, "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, January 2002, *70* (1), 331–341.

Willis, Robert J. and Sherwin Rosen, "Education and Self-Selection," *Journal of Political Economy*, October 1979, *87* (5), S7–S36.

Wolpin, Kenneth I., "Ex Ante Policy Evaluation, Structural Estimation and Model Selection," *American Economic Review*, May 2007, *97* (2), 48–52.