# Econometric Methods of Program Evaluation

Manuel Arellano

CEMFI

February 2015

## I. Structural and treatment effect approaches

- The classic approach to quantitative policy evaluation in economics is the *structural approach*.
- Its goals are to specify a class of theory-based models of individual choice, choose the one within the class that best fits the data, and use it for ex-post or ex-ante policy simulation.
- During the last two decades the *treatment effect approach* has established itself as a formidable competitor that has introduced a different language, different priorities, techniques and practices in applied work.
- Not only that, it has also changed the perception of evidence-based economics among economists, public opinion, and policy makers.
- The ambition in a structural exercise is to use data from a particular context to identify, with the help of theory, deep rules of behavior that can be extrapolated to other contexts.

- A treatment effect (TE) exercise is context-specific and addresses less ambitious policy questions.
- The goal is to evaluate the impact of an *existing* policy by comparing the distribution of a chosen outcome variable for individuals affected by the policy (treatment group) with the distribution of unaffected individuals (control group).
- The aim is to choose the control and treatment groups in such a way that membership of one or the other, either results from randomization or can be regarded as if they were the result of randomization.
- In this way one hopes to achieve the standards of empirical credibility on causal evidence that are typical of experimental biomedical studies.

- The TE literature has expressed dissatisfaction with the existing structural approach along several dimensions:

1. Between theory, data, and estimable structural models there is a host of untestable functional form assumptions that undermine the force of structural evidence by:
    - Having unknown implications for results.
    - Giving researchers too much discretion.
    - Complexity affects transparency and replicability.
2. By being too ambitious on the policy questions we get very little credible evidence from data. Too much emphasis on "external validity" at the expense of the more basic "internal validity".

- The TE literature sees the role of empirical findings as one of providing bits and pieces of hard evidence that can help the assessment of future policies in an informal way.
- Main gains in empirical research are not expected to come from the use of formal theory or sophisticated econometrics, but from understanding the sources of variation in data with the objective of identifying policy parameters.

- Many policy interventions at the micro level have been evaluated:

1. training programs
2. welfare programs (e.g. unemployment insurance, worker's sickness compensation)
3. wage subsidies and minimum wage laws
4. tax-credit programs
5. effects of taxes on labor supply and investment
6. effects of Medicaid on health

- I will review the following contexts or research designs of evaluation:

1. social experiments
2. matching
3. instrumental variables
4. regression discontinuity
5. differences in differences

- I pay special attention to instrumental-variable methods and their connections with econometric models.

## Descriptive analysis vs. causal inference

- It is useful to distinguish between descriptive analysis and causal inference as two types of micro empirical research.
- Their boundaries overlap, but there are good examples clearly placed in each category.
- Their difference is not in the sophistication of the statistical techniques employed. Sometimes the term "descriptive" is associated with tables of means or correlations, whereas terms like "econometric" or "rigorous" are reserved for regression coefficients or more complex statistics of the same style.
- A simple comparison of means can be causal, whereas complex statistical analyses (like semiparametric censored quantile regression) can be descriptive.
- Perhaps the greatest successes of econometrics are descriptive analyses.
- Recent examples include trends in inequality and wage mobility, productivity measurement, or quality-adjusted inflation hedonic indices.
- A useful description is not a mechanical exercise. It is a valuable research activity, often associated with innovative ideas. The ideas have to do with the choice of aspects to describe, the way of doing it, and their interpretation.

## II. Potential outcomes and causality

- Association and causation have always been known to be different, but a mathematical framework for an unambiguous characterization of *statistical causal effects* is surprisingly recent (Rubin, 1974; despite precedents in statistics and economics, Neyman, 1923; Roy, 1951).
- Think of a population of individuals that are susceptible of treatment. Let $Y_1$ be the outcome for an individual if exposed to treatment and let $Y_0$ be the outcome for the same individual if not exposed. The treatment effect for that individual is $Y_1 - Y_0$.
- In general, individuals differ in how much they gain from treatment, so that we can imagine a distribution of gains over the population with mean

$$\alpha_{ATE} = E\left(Y_1 - Y_0\right).$$

- The average treatment effect so defined is a standard measure of the causal effect of treatment 1 relative to treatment 0 on the chosen outcome.
- Suppose that treatment has been administered to a fraction of the population, and we observe whether an individual has been treated or not ($D = 1$ or 0) and the person's outcome $Y$. Thus, we are observing $Y_1$ for the treated and $Y_0$ for the rest:

$$Y = (1 - D)\, Y_0 + DY_1.$$

- Because $Y_1$ and $Y_0$ can never be observed for the same individual, the distribution of gains lacks empirical entity. It is just a conceptual device that can be related to observables.
- This notion of causality is statistical because it is not interested in finding out causal effects for specific individuals. Causality is defined in an average sense.

*Connection with regression*

- A standard measure of association between $Y$ and $D$ is:

$$\beta = E(Y \mid D = 1) - E(Y \mid D = 0)$$

$$= E(Y_1 - Y_0 \mid D = 1) + \{E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)\}$$

- The second expression makes it clear that in general $\beta$ differs from the *average gain for the treated* (another standard measure of causality, that we call $\alpha_{TT}$).
- The reason is that treated and nontreated units may have different average outcomes in the absence of treatment.
- For example, this will be the case if treatment status is the result of individual decisions, and 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, but here $\alpha_{ATE}$ or $\alpha_{TT}$ have been directly defined with respect to the distribution of potential outcomes, so that relative to a 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 TE literature emphasizes "reduced form causal effects" as an intermediate category between predictive and structural effects.

*Social feedback*

- The potential outcome representation is predicated on the assumption that the effect of treatment is independent of how many individuals receive treatment, so that the possibility of different outcomes depending on the treatment received by other units is ruled out.
- This excludes general equilibrium or feedback effects, as well as strategic interactions among agents.
- So the framework is not well suited to the evaluation of system-wide reforms which are intended to have substantial equilibrium effects.

## III. Social experiments

- In the TE approach, a randomized field trial is regarded as the ideal research design.
- Observational studies 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:

  $$(Y_0, Y_1) \perp D$$

  In such a case, $F(Y_1 \mid D=1) = F(Y_1)$ and $F(Y_0 \mid D=0) = F(Y_0)$. The implication is $\alpha_{ATE} = \alpha_{TT} = \beta$.
- Analysis of data takes a simple form: An unbiased estimate of $\alpha_{ATE}$ is the difference between the average outcomes for treatments and controls:

  $$\widehat{\alpha}_{ATE} = \overline{Y}_T - \overline{Y}_C$$

- In a randomized setting, there is no need to "control" for covariates, rendering multiple regression unnecessary, except if interested in effects for specific groups.

*Experimental testing of welfare programs in the US*

- Long history of randomized field trials in social welfare in the US, beginning in the 1960s.
- Moffitt (2003) provides a lucid assessment.
- Early experiments had many flaws due to lack of experience in designing experiments and in data analysis.
- During the 1980s the US 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 US 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.
- Randomization is not appropriate for evaluating reforms with major spillovers from which the control group cannot be isolated.
- But it is an effective means of testing incremental reforms and searching for policy designs "that reveal what works and for whom." (Moffitt).

*Example 1: Employment effect of a subsidized job program.*

- The NSW program was designed in the US in the mid 70's 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: To be unemployed, a long-term AFDC recipient, and have no preschool children.
- Participants were randomly assigned to treatment & control groups in 1976-77. Experiment took place in 7 cities.
- Ham–LaLonde data: 275 women in treatment group and 266 controls. All volunteered in 1976. Averages: Age 34, 10 years of schooling, 70% H.S. dropout, 2 children, 65% married, 85% black.
- Thanks to randomization, a simple comparison between the employment rates of treatments and controls gives an unbiased estimate of the effect of the program.
- Figure 1 taken from Ham–LaLonde shows the effects.

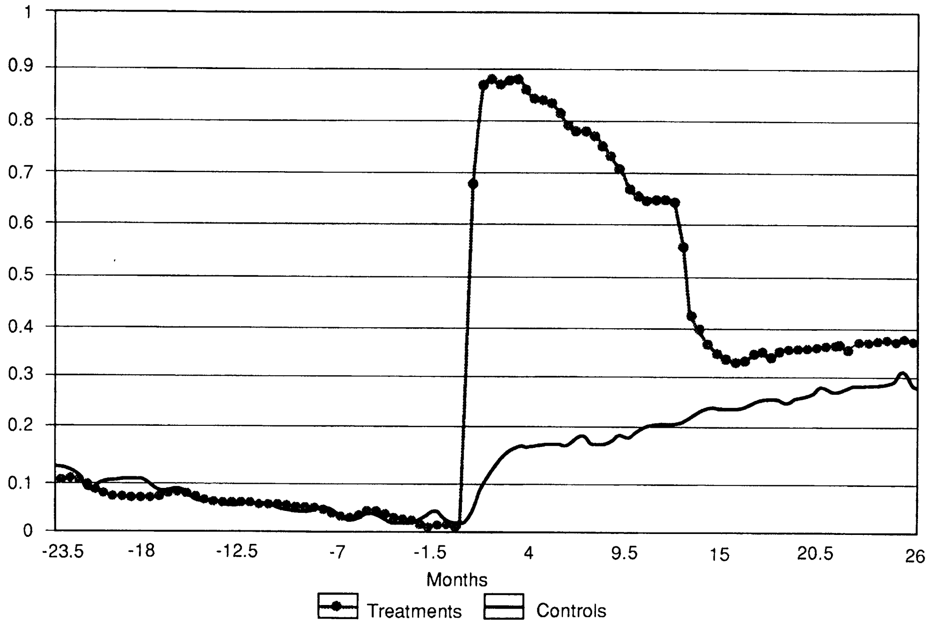

FIGURE 1.—Employment rates of AFDC women in the NSW Demonstration.

experimental evaluation shows that at least in the short run, NSW substantially improved the employment prospects of AFDC participants.

The NSW demonstration achieved these employment gains by helping trainees to hold on to their jobs longer and/or to find jobs faster, thereby increasing the length of their employment spells and/or reducing the length of their unemployment spells. To begin our analysis of these effects of training, we examine the Kaplan-Meier survivor functions for the treatments' and controls' employment and unemployment spells in Table I.[6] The first two columns of the table indicate that 65 percent of the trainees' employment spells lasted six or more months compared with only 57.3 percent of the controls' spells. When we follow standard practice and compare the experience of treatments and controls in fresh unemployment spells in columns three and five of Table I, we see that 73 percent of the treatments are still in an unemployment spell after a duration of 6 months compared to only 61.3 percent of the controls. Thus training appears to be a mixed blessing since it increases the length of both employment and unemployment spells.

Unfortunately, as previously noted, such a simple analysis of the treatments' and controls' employment histories may be misleading First, the possibility that the treatments and controls faced different demand conditions is particularly

[6] In practice many of the employment and unemployment spells are not completed during the sample period (i.e., they are right censored). Therefore, we cannot simply compare their mean durations, especially because the treatments spend on average half the sampling frame in training.

- The growth in the employment rates of the controls is just a reflection of the program's eligibility criteria.
- The conclusion from the experimental evaluation is that, at least in the short run, the NSW substantially improved the employment prospects of participants (a difference of 9 percentage points in employment rates).

*Covariates and job histories*

- At admission time, information collected on age, education, high-school dropout status, children, marital status, race, and labor history for the previous two years.
- Job histories following entry into the program: Treatments and controls were interviewed at 9 month intervals, collecting information on employment status. In this way employment and unemployment spells were constructed for more than two years following the baseline (26 months).

## *The Ham–LaLonde critique of experimental data*
## *A) Effects on wages*

- A direct comparison of mean wages for treatments and controls gives a biased estimate of the effect of the program on wages. This will happen as long as training has an impact on the employment rates of the treated.
- Let $W$ =wages, let $Y = 1$ if employed and $Y = 0$ if unemployed, $\eta = 1$ if high skill and $\eta = 0$ otherwise.
- Suppose that treatment increases the employment rates of high and low skill workers:

$$\begin{aligned}\Pr(Y = 1 \mid D = 1, \eta = 0) &> \Pr(Y = 1 \mid D = 0, \eta = 0)\\ \Pr(Y = 1 \mid D = 1, \eta = 1) &> \Pr(Y = 1 \mid D = 0, \eta = 1)\end{aligned}$$

- but the effect is of less intensity for the high skill group:

$$\frac{\Pr(Y = 1 \mid D = 1, \eta = 0)}{\Pr(Y = 1 \mid D = 0, \eta = 0)} > \frac{\Pr(Y = 1 \mid D = 1, \eta = 1)}{\Pr(Y = 1 \mid D = 0, \eta = 1)}.$$

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

$$\Pr(\eta = 0 \mid Y = 1, D = 1) > \Pr(\eta = 0 \mid Y = 1, D = 0),$$

  i.e. $\eta$ 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:

$$\Delta_f = E\left(W \mid Y = 1, D = 1\right) - E\left(W \mid Y = 1, D = 0\right),$$

whereas the effects of interest of $D$ on $W$ are:

  * For low skill individuals:

$$\begin{aligned} \Delta_0 &= E\left(W \mid Y = 1, D = 1, \eta = 0\right) \\ &\quad - E\left(W \mid Y = 1, D = 0, \eta = 0\right), \end{aligned}$$

  * for high skill:

$$\begin{aligned} \Delta_1 &= E\left(W \mid Y = 1, D = 1, \eta = 1\right) \\ &\quad - E\left(W \mid Y = 1, D = 0, \eta = 1\right) \end{aligned}$$

  * and the overall effect:

$$\Delta_s = \Delta_0 \Pr\left(\eta = 0\right) + \Delta_1 \Pr\left(\eta = 1\right).$$

- In general, we shall have that $\Delta_f < \Delta_s$.
- 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 $\eta$.

*B) Effects on durations*

- Effects on employment duration: similar to wages, the experimental comparison of exit rates from employment may be misleading. Let $T_e$ be the duration of an employment spell. An experimental comparison is

$$\Pr\left(T_e = t \mid T_e \geq t, D = 1\right) - \Pr\left(T_e = t \mid T_e \geq t, D = 0\right)$$

  but we are interested in

$$\Pr\left(T_e = t \mid T_e \geq t, D = 1, \eta\right) - \Pr\left(T_e = t \mid T_e \geq t, D = 0, \eta\right).$$

- $D$ is correlated with $\eta$ given $T_e \geq t$ for various reasons. e.g. If treatment especially helps to find a job those with $\eta = 0$ , the frequency of $\eta = 0$'s in the group $\{T_e \geq t, D = 1\}$ will increase relative to $\{T_e \geq t, D = 0\}$.
- Similar problems arise with unemployment durations. Ham and LaLonde's solution is to use an econometric model of labor histories with unobserved heterogeneity.
- The problem with wages and spells is one of censoring. It could be argued that the causal question is not well posed in these examples.
- Suppose that we wait until every individual completes an employment spell, and we consider the causal effect of treatment on the duration of such spell. This generates the problem that if the spells of controls and treatments tend to occur at different points in time, the economic environment is not held constant by the experimental design.

# Using experiments and models for ex ante evaluation

## Ex post and ex ante policy evaluation

- Ex post policy evaluation happens after the policy has been implemented.
  - The evaluation makes use of existing policy variation.
  - Experimental and nonexperimental 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 (Marschak 1953).
- Extrapolation requires a model (structural or nonstructural).
- The following discussion closely follows Todd and Wolpin (2006) and Wolpin (2007).

*Example 1: Ex ante evaluation of a school attendance subsidy program*

- Consider the following two situations:
  - Case (a): school tuition $p$ varies exogenously across counties in the range $\left(\underline{p}, \overline{p}\right)$.
  - Case (b): schools are free: $p = 0$.
- In case (a) it is possible to estimate a relationship between school attendance $s$ and tuition cost $p$, but in case (b) it is not.
- Suppose that $s$ also depends on a set of observed factors $X$ and it is possible to estimate nonparametrically
$$s = f\left(p, X\right) + v.$$
- 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.

*Example 2: Identifying subsidy effects from variation in the child market wage when $p = 0$*

- 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:

$$w < w^*$$

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

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

- To obtain separate estimates of $\alpha, \sigma$ 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

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

- 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.

## Combining experiments and structural estimation (Todd and Wolpin 2006)

*The PROGRESA school subsidy program*

- 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.

*Todd and Wolpin's model*

- Experimental effects 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.

*Todd and Wolpin's model (continued)*

- 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.

**Model selection: data mining (structural or otherwise)**

- 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.

## Holding out data: pros and cons

- 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.

*A Bayesian framework (Schorfheide and Wolpin 2012)*

- 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.
- The odd posterior ratio between two models is given by the odd prior ratio times the likelihood ratio:
$$\frac{\Pr\left(M_j \mid y\right)}{\Pr\left(M_\ell \mid y\right)} = \frac{\Pr\left(M_j\right)}{\Pr\left(M_\ell\right)} \frac{f\left(y \mid M_j\right)}{f\left(y \mid M_\ell\right)}$$
where $f\left(y \mid M_j\right) = \int f\left(y \mid \theta_j, M_j\right) \pi\left(\theta_j \mid M_j\right) d\theta_j$.
- The Schwarz approximation to the marginal ratio contains a correction factor for the difference in the number of parameters:
$$\frac{f\left(y \mid M_j\right)}{f\left(y \mid M_\ell\right)} \approx \frac{f\left(y \mid \widehat{\theta}_j, M_j\right)}{f\left(y \mid \widehat{\theta}_\ell, M_\ell\right)} \times n^{-[\dim(\theta_j) - \dim(\theta_\ell)]/2}.$$
- The overall posterior distribution of a treatment effect or predictor $\Delta$ is
$$p\left(\Delta \mid y\right) = \sum_j \Pr\left(M_j \mid y\right) p\left(\Delta \mid y, M_j\right)$$
where $p\left(\Delta \mid y, M_j\right)$ is the posterior density of $\Delta$ calculated under model $M_j$.

*A Bayesian framework (continued)*

- 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*.
  - Example: the Smet-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.

## A principal-agent framework

- 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 $\Delta$, then they have an incentive to reveal their subjective beliefs truthfully.
    - i.e. to report the posterior density of $\Delta$ 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.

# Matching Methods

## IV. 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.
- Experiments guarantee the independence condition

$$(Y_1, Y_0) \perp D$$

but with observational data it is not very plausible.

- A less demanding condition for nonexperimental data is:

$$(Y_1, Y_0) \perp D \mid X.$$

- Conditional independence implies

$$\begin{aligned} E(Y_1 \mid X) &= E(Y_1 \mid D = 1, X) = E(Y \mid D = 1, X) \\ E(Y_0 \mid X) &= E(Y_0 \mid D = 0, X) = E(Y \mid D = 0, X). \end{aligned}$$

Therefore, for $\alpha_{ATE}$ we can calculate (and similarly for $\alpha_{TT}$):

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

- The following is a matching expression for $\alpha_{TT} = E(Y_1 - Y_0 \mid D = 1)$:

$$E[Y - E(Y_0 \mid D = 1, X) \mid D = 1] = E[Y - \mu_0(X) \mid D = 1]$$

where $\mu_0(X) = E(Y \mid D = 0, X)$ is used as an imputation for $Y_0$.

*Relation with multiple regression*

- If we specify $E(Y \mid D, X)$ as a linear regression on $D$, $X$ and $D \times X$ we have

  $$E(Y \mid D, X) = \beta D + \gamma X + \delta DX$$

  and

  $$E(Y \mid D = 1, X) - E(Y \mid D = 0, X) = \beta + \delta X.$$

  $$\begin{aligned} \alpha_{ATE} &= \beta + \delta E(X) \\ \alpha_{TT} &= \beta + \delta E(X \mid D = 1), \end{aligned}$$

  which can be easily estimated using linear regression.
- Alternatively, we can treat $E(Y \mid D = 1, X)$ and $E(Y \mid D = 0, X)$ as nonparametric functions of $X$.
- The last approach is closer in spirit to the *matching* literature, which has emphasized direct comparisons, free from functional form assumptions and extrapolation.

*Sample-average vs population-level treatment effects*

- Sample-average versions of $\alpha_{ATE}$ and $\alpha_{TT}$ are

$$\begin{aligned}\alpha^S_{ATE} &= \frac{1}{N}\sum_{i=1}^{N}\left(Y_{1i}-Y_{0i}\right)\\ \alpha^S_{TT} &= \frac{1}{\sum_{i=1}^{N} D_i}\sum_{i=1}^{N} D_i\left(Y_{1i}-Y_{0i}\right)\end{aligned}$$

- If treatment gains were directly observed $\alpha^S_{ATE}$ and $\alpha^S_{TT}$ would be calculated without estimation error.
- A good estimate of $\alpha_{ATE}$ will also be a good estimate of $\alpha^S_{ATE}$ (and similarly for TTs), but one has to take a stand on what is being estimated because standard errors will be different in each case.
- One can estimate $\alpha^S_{ATE}$ at least as accurately as $\alpha_{ATE}$ and typically more so.
- This distinction matters because confidence intervals for one problem may be very different to those for the other (Imbens, 2004).

*Distributional effects and quantile treatment effects*

- Most of the literature focused on average effects, but the matching assumption also works for distributional comparisons.
- 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 $\alpha_{ATE}$ but also $E(Y_1)$ and $E(Y_0)$:

$$E(Y_1) = \int E(Y_1 \mid X)\, dF(X) = \int E(Y \mid D = 1, X)\, dF(X)$$

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

$$E[h(Y_1)] = \int E[h(Y_1) \mid X]\, dF(X) = \int E[h(Y) \mid D = 1, X]\, dF(X)$$

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

$$E[\mathbf{1}(Y_1 \leq r)] = \Pr(Y_1 \leq r) = \int \Pr(Y \leq r \mid D = 1, X)\, dF(X)$$

  and similarly for $\Pr(Y_0 \leq r)$.
- Given identification of the *cdf*s 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.

*The common support condition*

- Suppose for the sake of the argument that $X$ is a single covariate whose support lies in the range $\{X_{MIN}, X_{MAX}\}$.
- The support for the subpopulation of the treated $(D = 1)$ is $\{X_{MIN}, X_I\}$ whereas the support for the controls $(D = 0)$ is $\{X_0, X_{MAX}\}$ and $X_0 < X_I$, so that

$$\Pr(D = 1 \mid X \in \{X_{MIN}, X_0\}) = 1$$

$$0 < \Pr(D = 1 \mid X \in \{X_0, X_I\}) < 1$$

$$\Pr(D = 1 \mid X \in \{X_I, X_{MAX}\}) = 0$$

- The implication is that $E(Y \mid D = 1, X)$ is only identified for values of $X$ in the range $\{X_{MIN}, X_I\}$ and $E(Y \mid D = 0, X)$ is only identified for values of $X$ in the range $\{X_0, X_{MAX}\}$.
- Thus, we can only calculate the difference $[E(Y \mid D = 1, X) - (Y \mid D = 0, X)]$ for values of $X$ in the intersection range $\{X_0, X_I\}$, which implies that $\alpha_{ATE}$ is not identified. Only the average treatment effect of units with $X \in \{X_0, X_I\}$ is identified.
- If we want to ensure identification, in addition to conditional independence we need the overlap assumption:

$$0 < \Pr(D = 1 \mid X) < 1 \qquad \text{for all } X \text{ in its support}$$

*Lack of common support and parametric assumptions: a cautionary tale*

- Suppose that $E(Y_1 \mid X) = E(Y_0 \mid X) = m(X)$ for all $X$ but the support is as before. We can only hope to establish that $E(Y_1 - Y_0 \mid X = r) = 0$ for $r \in \{X_0, X_I\}$.
- Conditional independence holds, so $E(Y \mid D = 1, X) = E(Y \mid D = 0, X) = m(X)$, which in our example is a nonlinear function of $X$.
- Suppose that we use linear projections in place of conditional expectations:

$$\begin{aligned} E^*(Y \mid D = 0, X) &= \beta_0 + \beta_1 X \\ E^*(Y \mid D = 1, X) &= \gamma_0 + \gamma_1 X \end{aligned}$$

  where

$$\begin{aligned} (\beta_0, \beta_1) &= \arg\min_{b_0, b_1} E_{X \mid D=0} \left\{ \left[ E(Y \mid D = 0, X) - b_0 - b_1 X \right]^2 \right\} \\ (\gamma_0, \gamma_1) &= \arg\min_{g_0, g_1} E_{X \mid D=1} \left\{ \left[ E(Y \mid D = 1, X) - g_0 - g_1 X \right]^2 \right\} \end{aligned}$$

- Given the form of $m(X)$, $f(X \mid D = 0)$ and $f(X \mid D = 1)$ in the example, we shall get $\beta_1 > \gamma_1$. If we now project outside the observable ranges, we find a spurious negative treatment effect for large $X$ and a spurious positive effect for small $X$.
- So $\alpha_{ATE}$ calculated as $(\gamma_0 - \beta_0) + (\gamma_1 - \beta_1) E(X)$ may be positive, negative or close to zero depending on the form of the distributions involved, despite the fact that not only $E(Y_1 - Y_0) = 0$ but also $E(Y_1 - Y_0 \mid X) = 0$ for all values of $X$.

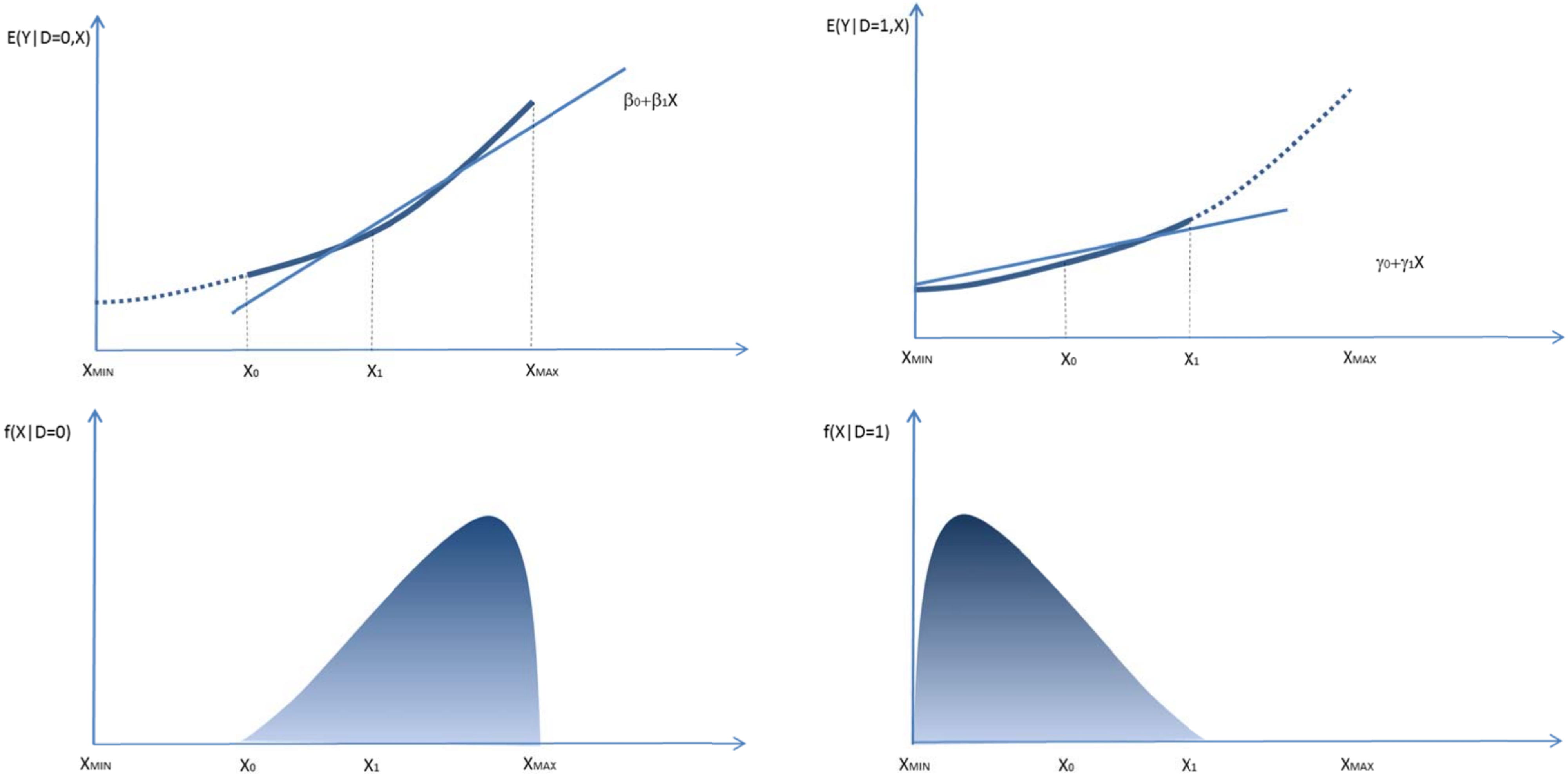
E(Y|D=0,X)
β0+β1X
XMIN
X0
X1
XMAX
E(Y|D=1,X)
γ0+γ1X
XMIN
X0
X1
XMAX
f(X|D=0)
XMIN
X0
X1
XMAX
f(X|D=1)
XMIN
X0
X1
XMAX

Figure 1

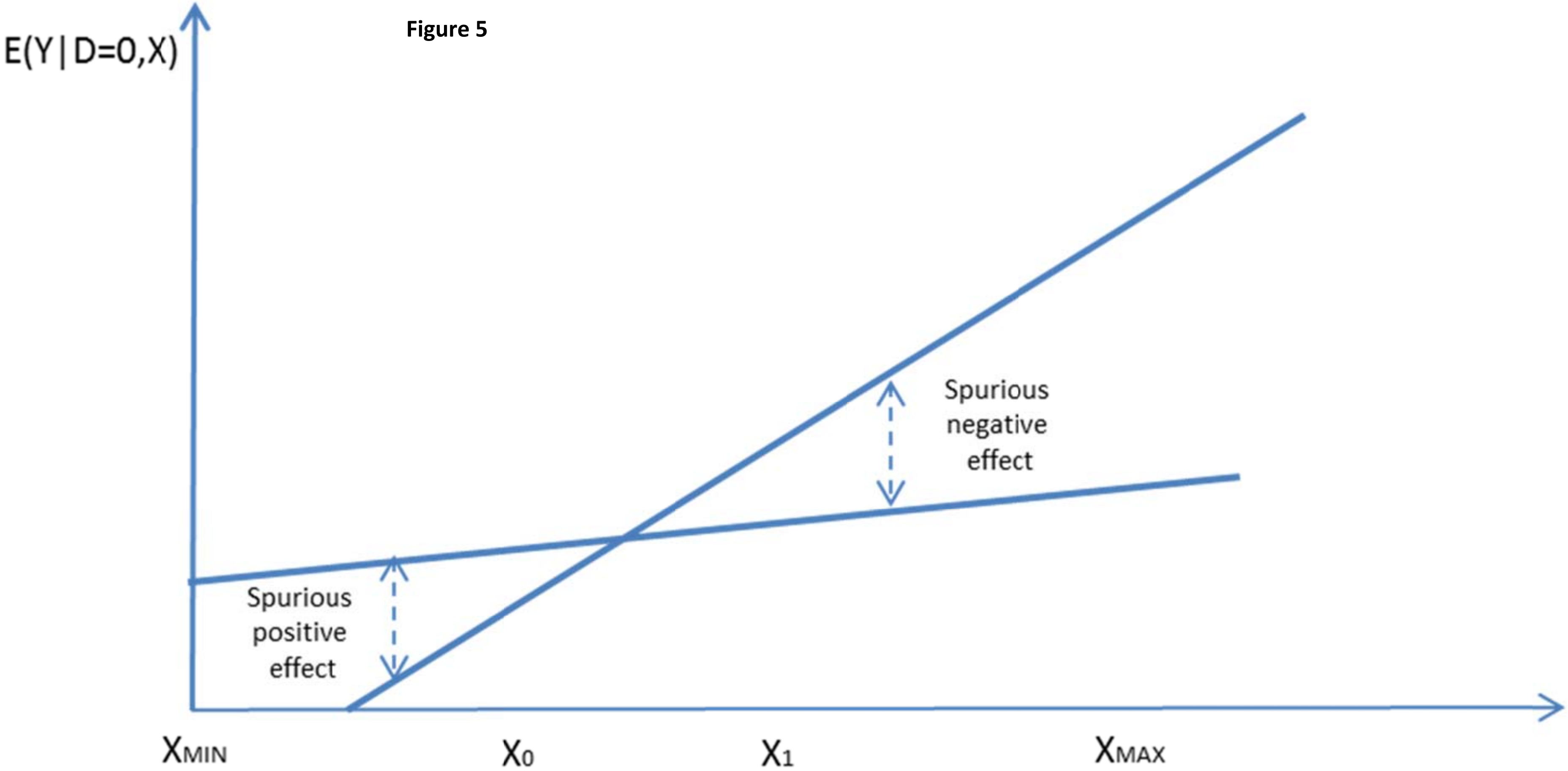
Figure 5
E(Y|D=0,X)
Spurious
negative
effect
Spurious
positive
effect
$X_{MIN}$
$X_0$
$X_1$
$X_{MAX}$

Figure 2

*Imputing missing outcomes (discrete $X$)*

- Suppose $X$ is discrete, takes on $J$ values $\left\{\xi_j\right\}_{j=1}^{J}$ and we have a sample $\{X_i\}_{i=1}^{N}$. Let

$$\begin{aligned} N^j &= \text{number of observations in cell } j. \\ N_\ell^j &= \text{number of observations in cell } j \text{ with } D = \ell. \\ \overline{Y}_\ell^j &= \text{mean outcome in cell } j \text{ for } D = \ell. \end{aligned}$$

- Thus, $\left(\overline{Y}_1^j - \overline{Y}_0^j\right)$ is the sample counterpart of

$$E\left(Y \mid D = 1, X = \xi_j\right) - E\left(Y \mid D = 0, X = \xi_j\right),$$

  which can be used to get the estimates

$$\widehat{\alpha}_{ATE} = \sum_{j=1}^{J}\left(\overline{Y}_1^j - \overline{Y}_0^j\right)\frac{N^j}{N}, \qquad \widehat{\alpha}_{TT} = \sum_{j=1}^{J}\left(\overline{Y}_1^j - \overline{Y}_0^j\right)\frac{N_1^j}{N_1}$$

- The formula for $\widehat{\alpha}_{TT}$ can also be written in the form

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

  where $j\left(i\right)$ is the cell of $X_i$. Thus, $\widehat{\alpha}_{TT}$ matches the outcome of each treated unit with the mean of the nontreated units in the same cell.
- To see this note that $E\left[E\left(Y \mid D = 1, X\right) - E\left(Y \mid D = 0, X\right) \mid D = 1\right] = E\left[Y - E\left(Y \mid D = 0, X\right) \mid D = 1\right]$.

*Imputing missing outcomes (continuous $X$)*

- A matching estimator can be regarded as a way of constructing imputations for missing potential outcomes so that gains $Y_{1i} - Y_{0i}$ can be estimated for each unit.
- In the discrete case
$$\widehat{Y}_{0i} = \overline{Y}_0^{j(i)} \equiv \sum_{k \in (D=0)} \frac{\mathbf{1}(X_k = X_i)}{\sum_{\ell \in (D=0)} \mathbf{1}(X_\ell = X_i)} Y_k$$
- In general
$$\widehat{Y}_{0i} = \sum_{k \in (D=0)} w(i,k) Y_k$$
- Different matching estimators use different weighting schemes.
- Nearest neighbor matching:
$$w(i,k) = \begin{cases} 1 \text{ if } X_k = \min_i \|X_k - X_i\| \\ 0 \text{ otherwise} \end{cases}$$
with perhaps matching restricted to cases where $\|X_i - X_k\| < \varepsilon$ for some $\varepsilon$. Usually applied in situations where the interest is in $\alpha_{TT}$ but also applicable to $\alpha_{ATE}$.
- Kernel matching:
$$w(i,k) = \frac{1}{\sum_{\ell \in (D=0)} K\left(\frac{X_\ell - X_i}{\gamma_{N_0}}\right)} K\left(\frac{X_k - X_i}{\gamma_{N_0}}\right)$$
where $K(.)$ is a kernel that downweights distant observations and $\gamma_{N_0}$ is a bandwidth parameter. Local linear approaches provide a generalization.

*Methods based on the propensity score*

- Rosenbaum and Rubin called "propensity score" to

$$\pi(X) = \Pr(D = 1 \mid X)$$

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

$$(Y_1, Y_0) \perp D \mid \pi(X)$$

  provided $0 < \pi(X) < 1$ for all $X$.

- We want to prove that provided $(Y_1, Y_0) \perp D \mid X$ then $\Pr(D = 1 \mid Y_1, Y_0, \pi(X)) = \Pr(D = 1 \mid \pi(X)) \equiv \pi(X)$. Using the law of iterated expectations:

$$\begin{aligned} E(D \mid Y_1, Y_0, \pi(X)) &= E[E(D \mid Y_1, Y_0, X) \mid Y_1, Y_0, \pi(X)] \\ &= E[E(D \mid X) \mid Y_1, Y_0, \pi(X)] = \pi(X) \end{aligned}$$

- The result tells us that we can match units with very different values of $X$ as long as they have similar values of $\pi(X)$.
- These results suggest two-step procedures in which we begin by estimating the propensity score.

*Weighting on the propensity score*

- Under unconditional independence

$$\alpha_{ATE} = E(Y \mid D = 1) - E(Y \mid D = 0) = \frac{E(DY)}{\Pr(D = 1)} - \frac{E[(1 - D)Y]}{\Pr(D = 0)}$$

- Similarly, under conditional independence

$$\begin{aligned} E(Y_1 - Y_0 \mid X) &= E(Y \mid D = 1, X) - E(Y \mid D = 0, X) \\ &= \frac{E(DY \mid X)}{\Pr(D = 1 \mid X)} - \frac{E[(1 - D)Y \mid X]}{\Pr(D = 0 \mid X)} \\ &= E\left(\frac{DY}{\pi(X)} - \frac{(1 - D)Y}{1 - \pi(X)} \mid X\right) \end{aligned}$$

so that

$$\alpha_{ATE} = E\left(\frac{DY}{\pi(X)} - \frac{(1 - D)Y}{1 - \pi(X)}\right) = E\left(Y \frac{[D - \pi(X)]}{\pi(X)[1 - \pi(X)]}\right)$$

- A simple estimator is

$$\widehat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{\widehat{\pi}(X_i)} - \frac{(1 - D_i)Y_i}{1 - \widehat{\pi}(X_i)}\right)$$

where $\widehat{\pi}(X_i)$ is a nonparametric series estimator of the propensity score (Hirano, Imbens, and Ridder, 2003).

*Quantile treatment effects (Firpo 2007)*

- Let $(Y_1, Y_0)$ be potential outcomes with marginal cdfs $F_1(r), F_0(r)$ and quantile functions $Q_{1\tau} = F_1^{-1}(\tau), Q_{0\tau} = F_0^{-1}(\tau)$. The QTE is defined to be

$$\theta_0 = Q_{1\tau} - Q_{0\tau}$$

- Under conditional exogeneity $F_j(r) = \int \Pr(Y \le r \mid D = j, X)\, dG(X)$, $(j = 0, 1)$. Moreover, $Q_{1\tau}, Q_{0\tau}$ satisfy the moment conditions:

$$\begin{aligned} E\left[\frac{D}{\pi(X)} 1(Y \le Q_{1\tau}) - \tau\right] &= 0 \\ E\left[\frac{1-D}{1-\pi(X)} 1(Y \le Q_{0\tau}) - \tau\right] &= 0 \end{aligned}$$

  and

$$Q_{1\tau} = \arg\min_q E\left[\frac{D}{\pi(X)} \rho_\tau(Y - q)\right], Q_{0\tau} = \arg\min_q E\left[\frac{1-D}{1-\pi(X)} \rho_\tau(Y - q)\right].$$

  where $\rho_\tau(u) = [\tau - 1(u < 0)] \times u$ is the "check" function.

- Firpo's method is a two-step weighting procedure in which the propensity score $\pi(X)$ is estimated first.

*Differences between matching and OLS*

- Matching avoids functional form assumptions and emphasizes the common support condition.
- Matching focuses on a single parameter at a time, which is obtained through explicit aggregation.

*The requirement of random variation in outcomes*

- Matching works on the presumption that for $X = x$ there is random variation in $D$, so that we can observe both $Y_1$ and $Y_0$. It fails if $D$ is a deterministic function of $X$.
- 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 $D$ for given $X$ is an indication that exogeneity is at fault.

*Example 2: Monetary incentives and schooling in the UK*

- The pilot of the *Education Maintenance Allowance* (EMA) program started in Sept. 1999. EMA paid youths aged 16–18 that continued in full time education (after 11 compulsory grades) a weekly stipend of £ 30 to 40, plus final bonuses for good results up to £140.
- Eligibility (and amounts paid) depends on household characteristics. Eligible for full payments if annual income under £13000. Those above £30000, not eligible.
- Dearden, Emmerson, Frayne & Meghir (2002) participated in the design of the pilot and did the evaluation.
- No experimental design for political reasons, but one defining treatment and control areas, both rural and urban.
- Basic question asked is whether more education results from this policy. The worry is that families fail to decide optimally due to liquidity constraints or misinformation.
- They use propensity scores. Probit estimates of $\pi\left(X\right)$ with family, local, and school characteristics. For each treated observation they construct a counterfactual mean using *kernel* regression and *bootstrap* standard errors.
- EMA increased participation in year 12 by 5.9% for eligible individuals, and by 3.7% for the whole population. Only significant results for full-payment recipients.

## Appendix: Local Linear Regression

- Let us consider estimating the regression function $g(x) = E(Y \mid X = x)$ from given observations $\{Y_i, X_i\}_{i=1}^n$.
- A linear approximation to $g(x)$ at a fixed point $r$ is

$$g(x) \approx a(r) + b(r)'(x - r)$$

  where $a(r) = g(r)$ and $b(r) = \partial g(r)/\partial r$ for $x$ in a neighborhood of $r$.
- Thus, locally, the problem of finding $g(r)$ is equivalent to finding the intercept of the approximating regression line.
- The local neighborhood may be determined by a kernel function $K$ and a smoothing parameter $\gamma_n$, which suggests using the least squares criterion

$$\sum_{i=1}^{n} K\left(\frac{X_i - r}{\gamma_n}\right)\left[Y_i - a - b'(X_i - r)\right]^2.$$

- Minimization with respect to $a$ and $b$ gives an estimate $\left[\widehat{a}(r), \widehat{b}(r)\right]$ of $g(r)$ and $\partial g(r)/\partial r$.

*Local Linear Regression (continued)*

- Letting $K_i(r) = K\left(\frac{X_i - r}{\gamma_n}\right)$ and

$$\begin{aligned}\overline{Y}(r) &= \frac{1}{\sum_{i=1}^n K_i(r)} \sum_{i=1}^n K_i(r)\, Y_i \\ \overline{D}_X(r) &= \frac{1}{\sum_{i=1}^n K_i(r)} \sum_{i=1}^n K_i(r)\,(X_i - r) \\ \widetilde{Y}_i(r) &= Y_i - \overline{Y}(r) \\ \widetilde{D}_{Xi}(r) &= (X_i - r) - \overline{D}_X(r),\end{aligned}$$

  the estimates are

$$\begin{aligned}\widehat{b}(r) &= \left(\sum_{i=1} K_i(r)\, \widetilde{D}_{Xi}(r)\, \widetilde{D}_{Xi}(r)'\right)^{-1} \sum_{i=1} K_i(r)\, \widetilde{D}_{Xi}(r)\, \widetilde{Y}_i(r) \\ \widehat{a}(r) &= \overline{Y}(r) - \overline{D}_X(r)'\, \widehat{b}(r).\end{aligned}$$

- The Nadaraya-Watson (NW) estimate of $g(r)$ is $\overline{Y}(r)$.
- If the distribution of the $X$'s in a neighborhood of $r$ is symmetric around $r$, then $\overline{D}_X(r) \approx 0$ and $\widehat{a}(r) \approx \overline{Y}(r)$ (i.e. the NW and local linear regression estimates of $g(r)$ will be close to each other).

*Local Linear Regression (continued)*

- However, if the $X$'s in a neighborhood of $r$ are mostly below (above) $r$ then $\overline{D}_X(r)$ will be negative (positive). In such case the local linear regression estimate applies a first-order correction to $\overline{Y}(r)$ using the local slope estimate $\widehat{b}(r)$.
- Thus, NW can be regarded as a local regression approximation to $g(r)$ of order zero, whereas $\widehat{a}(r)$ is a similar approximation of order one.
- Note that in the case where $X_i$ is discrete and $K\left(\frac{X_i-r}{\gamma_n}\right) = 1(X_i = r)$, the criterion boils down to $\sum_{X_i=r}(Y_i - a)^2$ which is minimized by the sample mean of $Y_i$ for the observations with $X_i = r$.
- Jianqing Fan (*JASA*, 1992) showed that local linear regression avoids the drawbacks of other types of kernel estimators such as NW.
- Local linear regression adapts to various types of designs (random, highly clustered, nearly uniform) and reduces boundary effects.

# Instrumental Variable Methods

## V. Instrumental variables

### 1. Instrumental variable assumptions

- Suppose we have non-experimental data with covariates, but cannot assume conditional independence as in matching:

$$(Y_1, Y_0) \perp D \mid X.$$

- Suppose, however, that we have a variable $Z$ that is an "exogenous source of variation in $D$" in the sense that it satisfies the *independence assumption*:

$$(Y_1, Y_0) \perp Z \mid X$$

  and the *relevance assumption*:

$$Z \not\perp D \mid X.$$

- Matching can be regarded as a special case of IV in which $Z = D$, i.e. all variation in $D$ is exogenous given $X$.

## 2. Instrumental-variable examples

### Example 1: Non-compliance in randomized trials

- In a classic example, $Z$ indicates assignment to treatment in an experimental design. Therefore, $(Y_1, Y_0) \perp Z$.
- However, "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 in general.
- 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, *Econometrica*, 2004).

## Example 2: Ethnic enclaves and immigrant outcomes

- Interest in the effect of leaving in a 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 (Edin, Fredriksson and Åslund, *QJE*, 2003).
- The causal effect is ambiguous. 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.
- Swedish governments of 1985-1991assigned initial areas of residence to refugee immigrants. Motivated by the belief that dispersing immigrants promotes integration.
- Let $Z$ indicate initial assignment (8 years before measuring ethnic enclave indicator $D$). Edin et al. assumed that $Z$ is independent of potential earnings $Y_0$ and $Y_1$.
- IV estimates implied a 13% gain for low-skill immigrants associated with one std. 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? (Angrist, 1990).
- Here we have:
  - Instrumental variable: draft lottery eligibility.
  - Treatment variable: Veteran status.
  - Outcome variable: Log earnings.
  - Data: $N = 11637$ white men born 1950–1953.
  - March Population Surveys 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 to serve up to a ceiling determined every year by the Department of Defense.
- Abadie (2002) uses as instrument an indicator for lottery numbers lower than 100.
- 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".
- 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.

## 3. Identification of causal effects in IV settings

- The question is whether the availability of an instrumental variable identifies causal effects. To answer it, I consider a binary $Z$, and abstract from conditioning.

### Homogeneous effects

- If the causal effect is the same for every individual

$$Y_{1i} - Y_{0i} = \alpha$$

  the availability of an IV allows us to identify $\alpha$. This is the traditional situation in econometric models with endogenous explanatory variables.

- In the homogeneous case

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

- Also, taking into account that $Y_{0i} \perp Z_i$

$$\begin{aligned} E(Y_i \mid Z_i = 1) &= E(Y_{0i}) + \alpha E(D_i \mid Z_i = 1) \\ E(Y_i \mid Z_i = 0) &= E(Y_{0i}) + \alpha E(D_i \mid Z_i = 0). \end{aligned}$$

- Subtracting both equations we obtain

$$\alpha = \frac{E(Y_i \mid Z_i = 1) - E(Y_i \mid Z_i = 0)}{E(D_i \mid Z_i = 1) - E(D_i \mid Z_i = 0)}$$

  which determines $\alpha$ as long as

$$E(D_i \mid Z_i = 1) \neq E(D_i \mid Z_i = 0).$$

- Get the effect of $D$ on $Y$ through the effect of $Z$ because $Z$ only affects $Y$ through $D$.

## Heterogeneous effects

*Summary*

- In the heterogeneous case the availability of IVs is not sufficient to identify a causal effect.
- 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 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 $Z$ (local average treatment effect or LATE).

**Indicator of potential treatment status**

- In preparation for the discussion below let us introduce the following notation:

$$D = \begin{cases} D_0 \text{ if } Z = 0 \\ D_1 \text{ if } Z = 1 \end{cases}$$

- Given data on $(Y, D)$ there are 4 observable groups but 8 underlying groups, which can be classified as never-takers, compliers, defiers, and always-takers.

*Example*

- Consider two levels of schooling ($D = 0, 1$, high school and college) with associated potential wages $(Y_0, Y_1)$, so that individual returns are $Y_1 - Y_0$. Also consider an exogenous determinant of schooling $Z$ with associated potential schooling levels $(D_0, D_1)$. The IV $Z$ is exogenous in the sense that it is independent of $(Y_0, Y_1, D_0, D_1)$.
- An example of $Z$ is proximity to college:
  - $Z = 0$ college far away
  - $Z = 1$ college nearby
  - Defier with $D = 1, Z = 0$ (ie. $D_1 = 0$): Person who goes to college when is far but would not go if it was near.
  - Defier with $D = 0, Z = 1$ (ie. $D_0 = 1$): Person does not go to college when it is near but would go if it was far.

Table 1

Observable and Latent Types

| | $Z$ | $D$ | $D_0$ | $D_1$ | | |
|---|---|---|---|---|---|---|
| Type 1 | 0 | 0 | 0 | 0 | Type 1A | Never-taker |
| | | | | 1 | Type 1B | Complier |
| Type 2 | 0 | 1 | 1 | 0 | Type 2A | Defier |
| | | | | 1 | Type 2B | Always-taker |
| Type 3 | 1 | 0 | 0 | 0 | Type 3A | Never-taker |
| | | | 1 | | Type 3B | Defier |
| Type 4 | 1 | 1 | 0 | 1 | Type 4A | Complier |
| | | | 1 | | Type 4B | Always-taker |

## Availability of IV is not sufficient by itself to identify causal effects

- Note that since

$$\begin{aligned} E\left(Y \mid Z=1\right) &= E\left(Y_0\right)+E\left[\left(Y_1-Y_0\right)D_1\right] \\ E\left(Y \mid Z=0\right) &= E\left(Y_0\right)+E\left[\left(Y_1-Y_0\right)D_0\right] \end{aligned}$$

  we have

$$E\left(Y \mid Z=1\right)-E\left(Y \mid Z=0\right)=E\left[\left(Y_1-Y_0\right)\left(D_1-D_0\right)\right]$$

$$\begin{aligned} = \quad & E\left(Y_1-Y_0 \mid D_1-D_0=1\right)\Pr\left(D_1-D_0=1\right) \\ & -E\left(Y_1-Y_0 \mid D_1-D_0=-1\right)\Pr\left(D_1-D_0=-1\right) \end{aligned}$$

- $E\left(Y \mid Z=1\right)-E\left(Y \mid Z=0\right)$ could be negative and yet the causal effect be positive for everyone, as long as the probability of defiers is sufficiently large.

**Additional assumption: Eligibility rules**

- An additional assumption that helps to identify $\alpha_{TT}$ is an eligibility rule of the form:

$$\Pr(D = 1 \mid Z = 0) = 0$$

i.e. individuals with $Z = 0$ are denied treatment.

- In this situation:

$$\begin{aligned} E(Y \mid Z = 1) &= E(Y_0) + E[(Y_1 - Y_0) D \mid Z = 1] \\ &= E(Y_0) + E(Y_1 - Y_0 \mid D = 1, Z = 1) E(D \mid Z = 1) \end{aligned}$$

and since $E(D \mid Z = 0) = 0$

$$E(Y \mid Z = 0) = E(Y_0) + E(Y_1 - Y_0 \mid D = 1, Z = 0) E(D \mid Z = 0) = E(Y_0)$$

- Therefore,

$$\text{Wald parameter} \equiv \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(D \mid Z = 1)} = E(Y_1 - Y_0 \mid D = 1, Z = 1).$$

- Moreover,

$$\alpha_{TT} \equiv E(Y_1 - Y_0 \mid D = 1) = E(Y_1 - Y_0 \mid D = 1, Z = 1).$$

This is so because $\Pr(Z = 1 \mid D = 1) = 1$. That is,

$$\begin{aligned} E(Y_1 - Y_0 \mid D = 1) &= E(Y_1 - Y_0 \mid D = 1, Z = 1) \Pr(Z = 1 \mid D = 1) \\ &\quad + E(Y_1 - Y_0 \mid D = 1, Z = 0)[1 - \Pr(Z = 1 \mid D = 1)]. \end{aligned}$$

- Thus, if $\Pr(D = 1 \mid Z = 0) = 0$ the IV coefficient coincides with the average treatment effect on the treated.

## 4. Local average treatment effects (LATE)
## Monotonicity and LATEs

- If we rule out defiers i.e. $\Pr(D_1 - D_0 = -1) = 0$, we have

$$E(Y \mid Z=1) - E(Y \mid Z=0) = E(Y_1 - Y_0 \mid D_1 - D_0 = 1) \Pr(D_1 - D_0 = 1)$$

  and

$$E(D \mid Z=1) - E(D \mid Z=0) = E(D_1) - E(D_0) = \Pr(D_1 - D_0 = 1).$$

- Therefore,

$$E(Y_1 - Y_0 \mid D_1 - D_0 = 1) = \frac{E(Y \mid Z=1) - E(Y \mid Z=0)}{E(D \mid Z=1) - E(D \mid Z=0)}$$

- Imbens and Angrist called this parameter “local average treatment effects” (LATE).
- Different IV's lead to different parameters, even under instrument validity, which is counter to standard GMM thinking.
- Policy relevance of a LATE parameter depends on the subpopulation of compliers defined by the instrument. Most relevant LATE's are those based on instruments that are policy variables (eg college fee policies or college creation).
- What happens if there are no compliers? In the absence of defiers, the probability of compliers satisfies

$$\Pr(D_1 - D_0 = 1) = E(D \mid Z=1) - E(D \mid Z=0).$$

  So, lack of compliers implies lack of instrument relevance, hence underidentification.

## Distributions of potential wages for compliers

- Imbens and Rubin (1997) showed that under monotonicity not only the average treatment effect for compliers is identified but also the entire marginal distributions of $Y_0$ and $Y_1$ for compliers.
- Abadie (2002) gives a simple proof that suggests a Wald calculation. For any function $h(.)$ let us consider

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

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

$$E(W_1 - W_0 \mid D_1 - D_0 = 1) = \frac{E(W \mid Z = 1) - E(W \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)},$$

  or substituting

$$E(h(Y_1) \mid D_1 - D_0 = 1) = \frac{E(h(Y)\,D \mid Z = 1) - E(h(Y)\,D \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)}.$$

- If we choose $h(Y) = 1(Y \leq r)$, the previous formula gives as an expression for the *cdf* of $Y_1$ for the compliers.

- Similarly, if we consider

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

  then

$$E(V_1 - V_0 \mid D_1 - D_0 = 1) = \frac{E(V \mid Z=1) - E(V \mid Z=0)}{E(1-D \mid Z=1) - E(1-D \mid Z=0)}$$

  or

$$E(h(Y_0) \mid D_1 - D_0 = 1) = \frac{E(h(Y)(1-D) \mid Z=1) - E(h(Y)(1-D) \mid Z=0)}{E(1-D \mid Z=1) - E(1-D \mid Z=0)}$$

  from which we can get the *cdf* of $Y_0$ for the compliers, again setting $h(Y) = 1(Y \leq r)$.
- To see the intuition, suppose $D$ is exogenous ($Z = D$), then the *cdf* of $Y \mid D = 0$ coincides with the *cdf* of $Y_0$, and the *cdf* of $Y \mid D = 1$ coincides with the *cdf* of $Y_1$.
- If we regress $h(Y)D$ on $D$, the OLS regression coefficient is

$$E[h(Y)D \mid D=1] - E[h(Y)D \mid D=0] = E[h(Y_1)]$$

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

$$E[h(Y)(1-D) \mid 1-D=1] - E[h(Y)(1-D) \mid 1-D=0] = E[h(Y_0)].$$

- 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.

## 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 $(Y_0, Y_1) \perp Z$ does not hold, but the following does:

$$\begin{array}{rcll} (Y_0, Y_1) & \perp & Z \mid X & \quad \text{(conditional independence)} \\ Z & \not\perp & D \mid X & \quad \text{(conditional relevance)} \end{array}$$

- For example, in the analysis of returns to college where $Z$ 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 given family background variables $X$.
- In a linear version of the problem:
  - First stage: Regress $D$ on $Z$ and $X$ → get $\widehat{D}$.
  - Second stage: Regress $Y$ on $\widehat{D}$ and $X$.
- In general we now have conditional LATE given $X$:

$$\gamma(X) = E(Y_1 - Y_0 \mid D_1 \neq D_0, X).$$

- On the other hand, we have conditional IV estimands:

$$\beta(X) = \frac{E(Y \mid Z = 1, X) - E(Y \mid Z = 0, X)}{E(D \mid Z = 1, X) - E(D \mid Z = 0, X)}$$

- What is the relevant aggregate effect? If the treatment effect is homogeneous given $X$

$$Y_1 - Y_0 = \beta(X),$$

then a parameter of interest is:

$$E[\beta(X)] = \int \beta(X)\, dF(X).$$

- However, in the case of heterogeneous effects, it makes sense to consider an average treatment effect for the overall subpopulation of compliers:

$$\beta_C = \int \beta(X)\, dF(X \mid compliers).$$

- Calculating $\beta_C$ appears problematic because $F(X \mid compliers)$ is unobservable, but

$$\begin{aligned}\beta_C &= \int \beta(X) \frac{\Pr(compliers \mid X)}{\Pr(compliers)} dF(X) \\ &= \int [E(Y \mid Z=1, X) - E(Y \mid Z=0, X)] \frac{1}{\Pr(compliers)} dF(X)\end{aligned}$$

where

$$\Pr(compliers) = \int [E(D \mid Z=1, X) - E(D \mid Z=0, X)]\, dF(X).$$

- Therefore,

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

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

## 5. Relating LATE to parametric models of the potential outcomes

### 5.1 The endogenous dummy explanatory variable probit model

- The model as usually written in terms of observables is

$$\begin{aligned} Y &= \mathbf{1}\left(\alpha + \beta D + U \geq 0\right) \\ D &= \mathbf{1}\left(\pi_0 + \pi_1 Z + V \geq 0\right) \end{aligned}$$

$$\begin{pmatrix} U \\ V \end{pmatrix} \mid Z \sim \mathcal{N}\left[0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

- 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:

$$\begin{aligned} Y_1 &= \mathbf{1}\left(\alpha + \beta + U \geq 0\right) \\ Y_0 &= \mathbf{1}\left(\alpha + U \geq 0\right) \end{aligned}$$

- The average probability effect of interest (ATE) is given by

$$\theta = E\left(Y_1 - Y_0\right) = \Phi\left(\alpha + \beta\right) - \Phi\left(\alpha\right).$$

- In less parametric specifications $E\left(Y_1 - Y_0\right)$ may not be point identified, but we may still be able to estimate LATE.

*Monotonicity is equivalent to the index model assumption for D*

- The equivalence between monotonicity and index models provides a link with economic assumptions.
- Consider the case where $Z$ is a scalar 0–1 instrument, so that there are only two potential values of $D$:

$$\begin{aligned} D_1 &= \mathbf{1}\left(\pi_0 + \pi_1 + V \geq 0\right) \\ D_0 &= \mathbf{1}\left(\pi_0 + V \geq 0\right). \end{aligned}$$

- Suppose without lack of generality that $\pi_1 \geq 0$. Then we can distinguish three subpopulations depending on an individual's value of $V$:
- Never-takers: Units with $V < -\pi_0 - \pi_1$. They have $D_1 = 0$ and $D_0 = 0$. Their mass is $1 - \Phi\left(\pi_0 + \pi_1\right)$.
- Compliers: Units with $V \geq -\pi_0 - \pi_1$ but $V < -\pi_0$. They have $D_1 = 1$ and $D_0 = 0$. Their mass is $\Phi\left(\pi_0 + \pi_1\right) - \Phi\left(\pi_0\right)$.
- Always-takers: Units with $V \geq -\pi_0$. They have $D_1 = 1$ and $D_0 = 1$. Their mass is $\Phi\left(\pi_0\right)$.

*LATE under joint probit assumptions*

- Let us obtain the average treatment effect for the subpopulation of compliers:

$$\theta_{LATE} = E\left(Y_1 - Y_0 \mid D_1 - D_0 = 1\right) \equiv E\left(Y_1 - Y_0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right).$$

- We have

$$E\left(Y_1 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right) = \Pr\left(\alpha + \beta + U \geq 0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right)$$
$$= 1 - \frac{\Pr\left(U \leq -\alpha - \beta, V \leq -\pi_0\right) - \Pr\left(U \leq -\alpha - \beta, V \leq -\pi_0 - \pi_1\right)}{\Pr\left(V \leq -\pi_0\right) - \Pr\left(V \leq -\pi_0 - \pi_1\right)}$$

  and similarly

$$E\left(Y_0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right) = \Pr\left(\alpha + U \geq 0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right)$$
$$= 1 - \frac{\Pr\left(U \leq -\alpha, V \leq -\pi_0\right) - \Pr\left(U \leq -\alpha, V \leq -\pi_0 - \pi_1\right)}{\Pr\left(V \leq -\pi_0\right) - \Pr\left(V \leq -\pi_0 - \pi_1\right)}.$$

- Finally,

$$\begin{aligned}\theta_{LATE} \quad = \quad & \frac{1}{\Phi\left(-\pi_0\right) - \Phi\left(-\pi_0 - \pi_1\right)}\left[\Phi_2\left(-\alpha, -\pi_0; \rho\right) - \Phi_2\left(-\alpha, -\pi_0 - \pi_1; \rho\right)\right. \\ & \left.-\Phi_2\left(-\alpha - \beta, -\pi_0; \rho\right) + \Phi_2\left(-\alpha - \beta, -\pi_0 - \pi_1; \rho\right)\right].\end{aligned}$$

  where $\Phi_2\left(r, s; \rho\right) = \Pr\left(U \leq r, V \leq s\right)$ is a standard normal bivariate probability.
- The nice thing about $\theta_{LATE}$ is that it is identified from the Wald formula in the absence of joint normality.
- In fact, it does not even require monotonicity in the relationship between $Y$ and $D$.

## 5.2 Models with additive errors: switching regressions

### The switching regression model with endogenous switch

- The model is as follows:

$$\begin{aligned} Y_i &= \alpha + \beta_i D_i + U_i \\ D_i &= 1\left(\gamma_0 + \gamma_1 Z_i + \varepsilon_i \geq 0\right) \end{aligned} \tag{1}$$

- The potential outcomes are

$$\begin{aligned} Y_{1i} &= \alpha + \beta_i + U_i \equiv \mu_1 + V_{1i} \\ Y_{0i} &= \alpha + U_i \equiv \mu_0 + V_{0i} \end{aligned}$$

  so that the treatment effect $\beta_i = Y_{1i} - Y_{0i}$ is heterogeneous.
- Traditional models assume that $\beta_i$ 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.
- $\beta_i$ may depend on unobservables and $D_i$ may be correlated with both $U_i$ and $\beta_i$.
- We assume the exclusion restriction holds in the sense that $\left(V_{1i}, V_{0i}, \varepsilon_i\right)$ or $\left(U_i, \beta_i, \varepsilon_i\right)$ are independent of $Z_i$.
- In terms of the alternative notation (letting $\alpha = \mu_0$ and $U_i = V_{0i}$):

$$Y_i = \mu_0 + \left(Y_{1i} - Y_{0i}\right) D_i + V_{0i} = \mu_0 + \left(\mu_1 - \mu_0\right) D_i + \left[V_{0i} + \left(V_{1i} - V_{0i}\right) D_i\right].$$

- Let us write the ATE as $\overline{\beta} = \mu_1 - \mu_0$ and $\xi_i = V_{1i} - V_{0i}$ so that $\beta_i = \overline{\beta} + \xi_i$.

**Example: Rosen and Willis (1979)**

- Consider the effect of education on earnings and the decision to become educated. We are interested in the decision of college education ($D = 1$) vs. high school ($D = 0$).
- The model consists of potential earnings with or without college education $(Y_1, Y_0)$ and a schooling decision rule:

$$D = 1\left(Y_1 - Y_0 > C\right).$$

- There are determinants of costs $(C)$ like distance to college, tuition fees, availability of scholarships, opportunity costs or borrowing constraints, which are potential instruments. $Y_1 - Y_0$ is the return to college education for a particular individual. Equation (1) can be regarded as a reduced form version of the schooling decision rule.
- In the Rosen & Willis model $Y_1 - Y_0$ may also depend on unobservables because they think of multiple abilities and comparative advantage. Moreover, the model suggests that $D_i$ may be correlated with both $U_i$ and $\beta_i$.

**Endogeneity and self-selection**

- Write

$$E\left(Y_i \mid Z_i\right) = \mu_0 + \left(\mu_1 - \mu_0\right) E\left(D_i \mid Z_i\right) + E\left(V_{1i} - V_{0i} \mid D_i = 1, Z_i\right) E\left(D_i \mid Z_i\right).$$

- If $\beta_i$ is mean independent of $D_i$

$$E\left(Y_i \mid Z_i\right) = \mu_0 + \left(\mu_1 - \mu_0\right) E\left(D_i \mid Z_i\right).$$

  so that $\overline{\beta} = Cov\left(Z, Y\right) / Cov\left(Z, D\right)$.
- Otherwise, $\overline{\beta}$ does not coincide with the IV estimand. A special case of mean independence of $\beta_i$ with respect to $D_i$ occurs when $\beta_i$ is constant.
- The failure of IV can be seen as the result of a missing variable. The model can be written as

$$Y_i = \alpha + \overline{\beta} D_i + \varphi\left(Z_i\right) D_i + \zeta_i$$

  where $\varphi\left(Z_i\right) = E\left(V_{1i} - V_{0i} \mid D_i = 1, Z_i\right)$. Note that $E\left(\zeta_i \mid Z_i\right) = 0$.
- When we do ordinary IV estimation we are not taking into account the variable $\varphi\left(Z_i\right) D_i$.
- $\varphi\left(z\right)$ is the average excess return for college-educated people with $Z_i = z$. In the distance to college example ($Z = 1$ if college near), we would expect $\varphi\left(1\right) \leq \varphi\left(0\right)$.
- The average treatment effect on the treated and the LATE are, respectively,

$$\alpha_{TT} = E\left(Y_{1i} - Y_{0i} \mid D_i = 1\right) = \overline{\beta} + E\left(V_{1i} - V_{0i} \mid D_i = 1\right),$$

$$\alpha_{LATE} = E\left(Y_{1i} - Y_{0i} \mid D_{1i} - D_{0i} = 1\right) = \overline{\beta} + E\left(V_{1i} - V_{0i} \mid -\gamma_0 - \gamma_1 \leq \varepsilon_i < -\gamma_0\right).$$

**The Gaussian model**

- The model is completed with the assumption

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

- 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

$$E\left(V_{1i} - V_{0i} \mid D_i = 1, Z_i\right) = \left(\sigma_{1\varepsilon} - \sigma_{0\varepsilon}\right)\lambda\left(\gamma_0 + \gamma_1 Z_i\right),$$

  so that we can do IV estimation in

$$Y_i = \alpha + \overline{\beta} D_i + \left(\sigma_{1\varepsilon} - \sigma_{0\varepsilon}\right)\lambda_i D_i + \zeta_i,$$

  or OLS estimation in:

$$Y_i = \alpha + \overline{\beta}\Phi_i + \left(\sigma_{1\varepsilon} - \sigma_{0\varepsilon}\right)\phi_i + \zeta_i^*.$$

**Identification without parametric distributional assumptions**

- The current model can be regarded as the combination of two generalized selection models. So the identification result for that model applies.
- Namely, with a continuous exclusion restriction $E\left(Y_{1i} \mid X_i\right)$ and $E\left(Y_{0i} \mid X_i\right)$ 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.

## 6. Marginal treatment effects

### Introduction

- When the support of $Z$ is not binary, there is a multiplicity of causal effects.
- What causal effects are relevant for evaluating a given policy?
- The natural experiment literature has been satisfied with identifying "causal effects", without paying much attention to their relevance.
- If $Z$ is continuous we can define a different LATE parameter for every pair $(z, z')$:

$$\alpha_{LATE}(z, z') = \frac{E(Y \mid Z = z) - E(Y \mid Z = z')}{E(D \mid Z = z) - E(D \mid Z = z')}.$$

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

### IV assumptions and monotonicity

- 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:
  - Independence: $(Y_1, Y_0, D_z) \perp Z$.
  - Relevance: $\Pr(D = 1 \mid Z = z) = P(z)$ is a nontrivial function of $z$.
- The monotonicity assumption for general $Z$ can be expressed as follows. For any pair of values $(z, z')$ either

$$D_{zi} \geq D_{z'i} \quad \text{or} \quad D_{zi} \leq D_{z'i}$$

  for all units in the population.

## Latent index representation

- Alternatively we can postulate an index model for $D_z$:

$$D_z = 1\left(\mu\left(z\right) - U > 0\right) \qquad \text{and } U \perp Z,$$

  which can be a useful way of organizing different LATEs (Heckman & 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 $\mu\left(z\right) = P\left(z\right)$ and take $U$ as uniformly distributed in the $\left(0, 1\right)$ interval. To see this note that

$$1\left(\mu\left(z\right) > U\right) = 1\left\{F_U\left[\mu\left(z\right)\right] > F_U\left(U\right)\right\} = 1\left(P\left(z\right) > \widetilde{U}\right)$$

  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 $P\left(0\right)$ and $P\left(1\right)$. Suppose that $P\left(0\right) < P\left(1\right)$. Always-takers have $U < P\left(0\right)$, compliers have a value of $U$ between $P\left(0\right)$ and $P\left(1\right)$, and never-takers have $U > P\left(1\right)$. A similar argument can be made for any pair $\left(z, z'\right)$ in the case of a general $Z$.
- So under monotonicity we can always invoke and index equation and imagine each member of the population as having a particular value of the unobserved variable $U$.

## Marginal Treatment Effect

- Using the propensity score $P(Z) = \Pr(D = 1 \mid Z)$ as instrument, LATE becomes

$$\alpha_{LATE}\left(P(z), P\left(z'\right)\right) = \frac{E\left(Y \mid P(Z) = P(z)\right) - E\left(Y \mid P(Z) = P(z')\right)}{P(z) - P(z')}.$$

- If $Z$ is binary this is equivalent to what we had in the first place, but if $Z$ is continuous, taking limits as $z \rightarrow z'$, we get a limiting form of LATE or MTE:

$$MTE\left(P(z)\right) = \frac{\partial E\left(Y \mid P(Z) = P(z)\right)}{\partial P(z)}.$$

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

$$\alpha_{LATE}\left(P(z), P\left(z'\right)\right) = E\left[Y_1 - Y_0 \mid P\left(z'\right) < U < P(z)\right]$$

- Similarly $MTE\left(P(z)\right)$ 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

$$MTE\left(P(z)\right) = E\left(Y_1 - Y_0 \mid U = P(z)\right)$$

- Integrating $MTE\left(P\left(z\right)\right)$ over different ranges of $U$ we can get other ATE measures. For example,

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

- Moreover,

$$\alpha_{ATE}=\int_{0}^{1}MTE\left(u\right)du,$$

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

**Policy-relevant treatment effects**

- Constructing suitably integrated $MTE\left(u\right)s$ 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.
- For example, 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.

## Estimation: Local IV method

- Heckman and Vytlacil suggest to estimate MTE by estimating the derivative of the conditional mean
$$E\left(Y \mid P\left(Z\right)=P\left(z\right), X=x\right)$$
using kernel-based local linear regression techniques.
- Note that in this context the propensity score plays a very different role to matching.
- *Testing for homogeneity (or absence of self-selection)*: A test of linearity on the propensity score (conditional on $X$) is a test of homogeneity of treatment effects.
- To see this use $Y=Y_0+\left(Y_1-Y_0\right)D$ and write
$$\begin{aligned}E\left(Y \mid P\left(Z\right)\right) &= E\left(Y_0 \mid P\left(Z\right)\right)+E\left(\left(Y_1-Y_0\right)D \mid P\left(Z\right)\right)\\ &= E\left(Y_0\right)+E\left[Y_1-Y_0 \mid D=1, P\left(Z\right)\right]P\left(Z\right)\end{aligned}$$
- The quantity $E\left[Y_1-Y_0 \mid D=1, P\left(Z\right)\right]$ is constant under homogeneity, so that the conditional mean $E\left(Y \mid P\left(Z\right)\right)$ is linear in $P\left(Z\right)$.

## Remarks about unobserved heterogeneity in IV settings

- How important is it?
  - The balance between observed and unobserved heterogeneity depends on how detailed information on agents is available (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.
- Warnings:
  - 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 (Deaton, 2009).
- Heterogeneity of gains vs. heterogeneity of treatments
  - 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.

# VI. Regression discontinuity methods

## 1. Introduction and examples

- In the matching context we make the conditional exogeneity assumption

$$(Y_1, Y_0) \perp D \mid X$$

whereas in the IV context we assume

$$\begin{aligned} (Y_1, Y_0) \quad & \perp \quad Z \mid X \qquad \text{(independence)} \\ D \quad & \not\perp \quad Z \mid X \qquad \text{(relevance).} \end{aligned}$$

The relevance condition can also be expressed as saying that for some $z \neq z'$

$$\Pr(D = 1 \mid Z = z) \neq \Pr\left(D = 1 \mid Z = z'\right).$$

- 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 often create usable discontinuities.

## Examples

- Effect of class size on test scores (“Maimonides’ rule” in Israel, Angrist & Lavy, 1999):

| | | |
|---|---|---|
| $Y_{is}$ | : | average score at class $i$ in school $s$ |
| $D_{is}$ | : | size of class $i$ (not binary) |
| $Z_s$ | : | beginning of year enrollment in school $s$ |

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.

## Examples (continued)

- Effect of financial aid offers on students' enrollment decisions (van der Klaauw, 2002)

  $Y_i$ : decision of student $i$ to enroll in college "X" (binary)

  $D_i$ : amount of financial aid offer to student $i$

  $Z_i$ : 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.

- Do parties matter for economic outcomes? (Petterson-Lidbom, 2006):

  $Y_i$ : economic outcome in area $i$

  $D_i$ : party control indicator in local government $i$

  $Z_i$ : vote share

## 2. The fundamental RD assumption

- We can now state the basic RD assumption more formally. Namely, discontinuity in treatment assignment but continuity in potential outcomes: There is at least a known value $z = z_0$ such that

$$\begin{aligned} \lim_{z \to z_0^+} \Pr\left(D = 1 \mid Z = z\right) &\neq \lim_{z \to z_0^-} \Pr\left(D = 1 \mid Z = z\right) && (2) \\ \lim_{z \to z_0^+} \Pr\left(Y_j \leq r \mid Z = z\right) &= \lim_{z \to z_0^-} \Pr\left(Y_j \leq r \mid Z = z\right) \qquad (j = 0, 1) && (3) \end{aligned}$$

  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.

### Sharp and fuzzy designs

- The early RD literature in psychology (Cook & Campbell 1979) distinguished between "sharp" and "fuzzy" designs. In the former, $D$ is a deterministic function of $Z$:

$$D = 1\left(Z \geq z_0\right)$$

  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

$$\lim_{z \to z_0^+} E\left(D \mid Z = z\right) = 1, \qquad \lim_{z \to z_0^-} E\left(D \mid Z = z\right) = 0.$$

### 3. Homogeneous treatment effects

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

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

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

$$\begin{aligned} \lim_{z \to z_0^+} E\left(Y \mid Z = z\right) &= \alpha \lim_{z \to z_0^+} E\left(D \mid Z = z\right) + \lim_{z \to z_0^+} E\left(Y_0 \mid Z = z\right) \\ \lim_{z \to z_0^-} E\left(Y \mid Z = z\right) &= \alpha \lim_{z \to z_0^-} E\left(D \mid Z = z\right) + \lim_{z \to z_0^-} E\left(Y_0 \mid Z = z\right). \end{aligned}$$

- The RD assumption then leads to consideration of the following RD parameter

$$\gamma = \frac{\lim_{z \to z_0^+} E\left(Y \mid Z = z\right) - \lim_{z \to z_0^-} E\left(Y \mid Z = z\right)}{\lim_{z \to z_0^+} E\left(D \mid Z = z\right) - \lim_{z \to z_0^-} E\left(D \mid Z = z\right)}$$

  which is determined provided the "relevance part" (2) of the RD assumption is satisfied, and equals $\alpha$ provided the "independence part" (3) of the RD assumption holds.

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

$$\gamma = \lim_{z \to z_0^+} E\left(Y \mid Z = z\right) - \lim_{z \to z_0^-} E\left(Y \mid Z = z\right), \tag{4}$$

  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.

## 4. Heterogeneous treatment effects

- Now suppose that

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

- In the *sharp design* since $D_i = 1(Z \geq z_0)$ we have

$$E(Y \mid Z = z) = E(\alpha \mid Z = z) 1(z \geq z_0) + E(Y_0 \mid Z = z).$$

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

$$k(z) = E(Y_0 \mid Z = z) + [E(\alpha \mid Z = z) - E(\alpha \mid Z = z_0)] 1(z \geq z_0)$$

  we have

$$E(Y \mid Z = z) = E(\alpha \mid Z = z_0) 1(z \geq z_0) + k(z)$$

  where $k(z)$ is continuous at $z = z_0$.

- Therefore, the OLS population coefficient on $D$ in the equation

$$Y = \gamma D + k(z) + w \tag{5}$$

  coincides with $\gamma$, which in turn equals $E(\alpha \mid Z = z_0)$.

- The control function $k(z)$ is nonparametrically identified. To see this, first note that $\gamma$ is identified from (4). Then $k(z)$ is identifiable as the nonparametric regression $E(Y - \gamma D \mid Z = z)$. Note that if the treatment effect is homogeneous $k(z)$ coincides with $E(Y_0 \mid Z = z)$, but not in general.

- If $\mu(z) \equiv E(Y_0 \mid Z = z)$ 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 $E(\alpha \mid z \geq z_0)$.
- In the *fuzzy design*, $D$ not only depends on $1(Z \geq z_0)$ but also on other unobserved variables. Thus, $D$ is an endogenous variable in equation (5). However, we can still use $1(Z \geq z_0)$ as an instrument for $D$ in such equation to identify $\gamma$, at least in the homogeneous case.
- The connection between the fuzzy design and the instrumental variables perspective was first made explicit in van der Klaauw (2002).
- Next, we discuss the interpretation of $\gamma$ in the fuzzy design with heterogeneous treatment effects, under two different assumptions.

## Conditional independence near $z_0$

- Let us first consider the weak conditional independence assumption

$$D \perp (Y_0, Y_1) \mid Z = z$$

for $z$ near $z_0$, i.e. for $z = z_0 \pm e$ where $e > 0$ denotes an arbitrarily small number, or

$$\Pr\left(Y_j \leq r \mid D = 1, Z = z_0 \pm e\right) = \Pr\left(Y_j \leq r \mid Z = z_0 \pm e\right) \qquad (j = 0, 1)\,.$$

- That is, we are assuming that treatment assignment is exogenous in a neighborhood of $z_0$. An implication is

$$E\left(\alpha D \mid Z = z_0 \pm e\right) = E\left(\alpha \mid Z = z_0 \pm e\right) E\left(D \mid Z = z_0 \pm e\right).$$

- Proceeding as before, we have

$$\begin{aligned}
\lim_{z \to z_0^+} E\left(Y \mid Z = z\right) &= \lim_{z \to z_0^+} E\left(\alpha \mid D = 1, Z = z\right) \lim_{z \to z_0^+} \Pr\left(D = 1 \mid Z = z\right) \\
&\quad + \lim_{z \to z_0^+} E\left(Y_0 \mid Z = z\right) \\
\lim_{z \to z_0^-} E\left(Y \mid Z = z\right) &= \lim_{z \to z_0^-} E\left(\alpha \mid D = 1, Z = z\right) \lim_{z \to z_0^-} \Pr\left(D = 1 \mid Z = z\right) \\
&\quad + \lim_{z \to z_0^-} E\left(Y_0 \mid Z = z\right)
\end{aligned}$$

and

$$\begin{aligned}
\lim_{z \to z_0^+} E\left(Y \mid Z = z\right) &= E\left(\alpha \mid Z = z_0\right) \lim_{z \to z_0^+} \Pr\left(D = 1 \mid Z = z\right) + \lim_{z \to z_0^+} E\left(Y_0 \mid Z = z\right) \\
\lim_{z \to z_0^-} E\left(Y \mid Z = z\right) &= E\left(\alpha \mid Z = z_0\right) \lim_{z \to z_0^-} \Pr\left(D = 1 \mid Z = z\right) + \lim_{z \to z_0^-} E\left(Y_0 \mid Z = z\right).
\end{aligned}$$

- Subtracting

$$\begin{aligned}
& \lim_{z \to z_0^+} E\left(Y \mid Z = z\right) - \lim_{z \to z_0^-} E\left(Y \mid Z = z\right) \\
= & \left[\lim_{z \to z_0^+} \Pr\left(D = 1 \mid Z = z\right) - \lim_{z \to z_0^-} \Pr\left(D = 1 \mid Z = z\right)\right] E\left(\alpha \mid Z = z_0\right).
\end{aligned}$$

- Thus, it emerges that

$$\gamma = E\left(Y_1 - Y_0 \mid Z = z_0\right).$$

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

## Monotonicity near $z_0$

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

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

- 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 $\gamma$ identifies the local average treatment effect at $z = z_0$:

$$\gamma = \lim_{\varepsilon \to 0^+} E\left(Y_1 - Y_0 \mid D_{z_0+\varepsilon} - D_{z_0-\varepsilon} = 1\right)$$

  i.e. 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.

### 5. Estimation strategies

- There are parametric and semiparametric strategies.

### A nonparametric Wald estimator

- Hahn-Todd-van der Klaauw suggested the following local Wald estimator. Let $S_i \equiv 1\left(z_0 - h < Z_i < z_0 + h\right)$ 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 $W_i \equiv 1\left(z_0 < Z_i < z_0 + h\right)$ as an instrument, applied to the subsample with $S_i = 1$:

$$\widehat{\gamma} = \frac{\widehat{E}\left(Y_i \mid W_i = 1, S_i = 1\right) - \widehat{E}\left(Y_i \mid W_i = 0, S_i = 1\right)}{\widehat{E}\left(D_i \mid W_i = 1, S_i = 1\right) - \widehat{E}\left(D_i \mid W_i = 0, S_i = 1\right)}.$$

- This estimator has nevertheless a poor boundary performance. An alternative suggested by HTV is a local linear regression method.

## Parametric and semiparametric alternatives

- Suppose

  $$E\left(D \mid Z\right) = g\left(Z\right) + \delta 1\left(Z \geq z_0\right)$$

  and

  $$E\left(Y_0 \mid Z\right) = k\left(Z\right).$$

- A control function regression-based approach is based in the control function augmented equation that replaces $D$ by the propensity score $E\left(D \mid Z\right)$:

  $$Y = \gamma E\left(D \mid Z\right) + k\left(Z\right) + w$$

- In a parametric approach, we assume functional forms for $g\left(Z\right)$ and $k\left(Z\right)$. van der Klaauw (2002) considered a semiparametric approach using a power series approximation for $k\left(Z\right)$.
- If $g\left(Z\right) = k\left(Z\right)$, then we can do 2SLS using as instrumental variables

  $$\left\{1\left(Z \geq z_0\right), g\left(Z\right)\right\},$$

  where $g\left(Z\right)$ is the "included" instrument and $1\left(Z \geq z_0\right)$ 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.

## 6. Distributional effects

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

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

- Using $h(Y) = 1(Y \leq r)$, the RD parameter for the outcome $W(r) = 1(Y \leq r)\,D$ delivers

$$\Pr(Y_1 \leq r \mid Z = z_0) = \frac{\lim_{z \to z_0^+} E(W(r) \mid Z = z) - \lim_{z \to z_0^-} E(W(r) \mid Z = z)}{\lim_{z \to z_0^+} E(D \mid Z = z) - \lim_{z \to z_0^-} E(D \mid Z = z)}$$

under the local conditional independence assumption.

- A similar strategy can be followed to obtain $\Pr(Y_0 \leq r \mid Z = z_0)$. In that case we consider

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

- The RD parameter for the outcome $V(r) = 1(Y \leq r)(1 - D)$ delivers

$$\Pr(Y_0 \leq r \mid Z = z_0) = \frac{\lim_{z \to z_0^+} E(V(r) \mid Z = z) - \lim_{z \to z_0^-} E(V(r) \mid Z = z)}{\lim_{z \to z_0^+} E(D \mid Z = z) - \lim_{z \to z_0^-} E(D \mid Z = z)}.$$

### 7. 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 Hoxby, QJE, 2000, 1239–1285, for an application).

## VII. Differences in differences

*Example: minimum wages and employment*

- 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) evaluated 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:
  - A large source of employment for low-wage workers.
  - They comply with minimum wage regulations (especially franchised restaurants).
  - Fairly homogeneous job, so good measures of employment and wages can be obtained.
  - Easy to get a sample frame of franchised restaurants (yellow pages) with high response rates.
  - Response rates 87% and 73% (less in Penn, because the interviewer was less persistent).

- The DID coefficient is

$$\begin{aligned}\beta &= \left[E\left(Y_2 \mid D=1\right)-E\left(Y_1 \mid D=1\right)\right] \\ &\quad -\left[E\left(Y_2 \mid D=0\right)-E\left(Y_1 \mid D=0\right)\right].\end{aligned}$$

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

- $\beta$ measures the difference between the average employment change in NJ and the average employment change in Penn.
- The key assumption in giving a causal interpretation to $\beta$ is that the temporal effect in the two states is the same in the absence of intervention.
- But 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.

*The context of difference in difference comparisons*

- 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.
- Two-period potential outcomes with treatment in $t = 2$:

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

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

$$E\left(Y_0(2) - Y_0(1) \mid D = 1\right) = E\left(Y_0(2) - Y_0(1) \mid D = 0\right).$$

- $Y_0(1)$ 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:

$$\beta = [E(Y_2 \mid D=1) - E(Y_1 \mid D=1)] - [E(Y_2 \mid D=0) - E(Y_1 \mid D=0)]$$

$$= E(Y_1(2) \mid D=1) - E(Y_0(1) \mid D=1) - [E(Y_0(2) \mid D=0) - E(Y_0(1) \mid D=0)]$$

- Now, adding and subtracting $E(Y_0(2) \mid D=1)$:

$$\begin{aligned}\beta &= E[Y_1(2) - Y_0(2) \mid D=1] \\ &\quad + \{E[Y_0(2) - Y_0(1) \mid D=1] - E[Y_0(2) - Y_0(1) \mid D=0]\},\end{aligned}$$

which as long as the last term vanishes it equals

$$\beta = E[Y_1(2) - Y_0(2) \mid D=1].$$

*Comments and problems*

- $\beta$ 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 $\beta$ from a regression of outcome changes on the treatment dummy. This is convenient for accounting for dependence between the two periods.
- Differences in the composition of the cross-sectional populations over time (especially problematic if not using panel data).
- The fundamental assumption might be satisfied conditionally given certain covariates, but identification vanishes if some of them are *unobservable*.

## VIII. Concluding remarks

*Empirical work and empirical content*

- 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.

*Quasi-experimental approaches in policy evaluation*

- 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 an 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.