CHAPTER 1

Causal Inference in Urban and Regional Economics

Nathaniel Baum-Snow*, Fernando Ferreira†
*Department of Economics, Brown University, Providence, RI, USA
†The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Contents

Abstract

Recovery of causal relationships in data is an essential part of scholarly inquiry in the social sciences. This chapter discusses strategies that have been successfully used in urban and regional economics for recovering such causal relationships. Essential to any successful empirical inquiry is careful consideration of the sources of variation in the data that identify parameters of interest. Interpretation of such parameters should take into account the potential for their heterogeneity as a function of both observables and unobservables.

Keywords

Casual inference, Urban economics, Regional economics, Research design, Empirical methods, Treatment effects

Handbook of Regional and Urban Economics, Volume 5A
ISSN 1574-0080, http://dx.doi.org/10.1016/B978-0-444-59517-1.00001-5

JEL Classification Code

R1

1.1. INTRODUCTION

The field of urban and regional economics has become much more empirically oriented over recent decades. In 1990, 49% of publications in the *Journal of Urban Economics* were empirical, growing to 71% in 2010. Moreover, the set of empirical strategies that are most commonly employed has changed. While most empirical papers in 1990 only used cross-sectional regressions, articles in 2010 were more likely to use instrumental variables (IV), panel data, and nonlinear models. Furthermore, special attention is now paid to the employment of research designs that can plausibly handle standard omitted variable bias problems. While only a handful of papers attempted to deal with these problems in 1990, more than half of the empirical publications in 2010 used at least one research design that is more sophisticated than simple ordinary least squares (OLS), such as difference in differences (DD), matching, and IV, to recover causal parameters. However, the credibility of estimates generated with these more sophisticated techniques still varies. While, in general, the credibility of empirical work in urban economics has improved markedly since 1990, many studies continue to mechanically apply empirical techniques while omitting important discussions of the sources of identifying variation in the data and of which treatment effects, if any, are being recovered. Table 1.1 details the percentages of publications in the *Journal of Urban Economics* that were empirical and the distribution of empirical methods used for the years 1980, 1990, 2000, and 2010.

This chapter discusses the ways that researchers have successfully implemented empirical strategies that deliver the most credible treatment effect estimates from data sets that describe urban and regional phenomena. Our treatment emphasizes the importance of randomization, which has been more broadly recognized in other fields, most notably development economics. Randomized trials are an important tool to recover treatment effects, especially those of interest for policy evaluation (Duflo et al., 2008). However, it is typically more challenging and expensive to implement field

Table 1.1 Prevalence of empirical methods in the *Journal of Urban Economics*, 1980–2010

		As percentages of empirical papers						
Year	Empirical	OLS	IV	Logit/ probit	Panel data	Difference in differences	Randomization	Matching
1980	57%	87%	10%	3%	0%	0%	0%	0%
1990	49%	79%	17%	13%	4%	0%	0%	0%
2000	62%	64%	32%	36%	14%	4%	0%	0%
2010	71%	77%	46%	26%	62%	8%	3%	5%

Notes: Authors calculations from all published articles in the *Journal of Urban Economics* in the indicated years.

experiments in settings of interest to urban and regional economists, as it is in other fields such as labor economics. General equilibrium effects, which contaminate control groups with influences of treatment, are more likely to arise in urban settings. Moreover, the nature of such general equilibrium effects is more likely to be the object of inquiry by urban and regional researchers. Labor economists have typically adopted higher standards for evaluating the credibility of estimated causal effects in research that uses nonexperimental data. Here we explore identification strategies that have been successfully used to recover credible estimates of treatment effects, typically in the absence of experimental variation. These include DD, various fixed effects methods, propensity score matching, IV, and regression discontinuity (RD) identification strategies. We also discuss treatment effect heterogeneity and how differences in results across identification strategies may simply reflect different causal relationships in the data. We emphasize that especially without experimental variation (and even often with experimental variation), no one identification strategy is ever perfect. Moreover, when considering causal effects of treatments, it is useful to think in the context of a world in which a distribution of treatment effects exists. Selection into treatment (on both observable and unobservable characteristics) and treatment effect heterogeneity makes empirical work complicated.

One recurring theme of this chapter is the following principle, which applies to all empirical strategies: it is crucial to consider the sources of variation in the treatment variables that are used to recover parameters of interest. Distinguishing this "identifying variation" allows the researcher to consider two central questions. First, could there be unobserved variables that both influence the outcome and are correlated with this identifying variation in the treatment variable? If such omitted variables exist, coefficients on the treatments are estimated as biased and inconsistent. We typically label such situations as those with an "endogeneity problem." Second, how representative of the population is the subset of the data for which such identifying variation exists? If clean identification exists only in a small unrepresentative subset of the population, coefficients on treatment variables apply only narrowly and are unlikely to generalize to other populations.

Throughout the chapter, we discuss the key properties of various identification strategies mostly assuming a simple linear data-generating process which allows for heterogeneous treatment effects. Each section cites articles from the literature for readers interested in the details of more complex applications. This structure allows us to easily explain the relationships between different empirical strategies while leaving space to cover applications in urban and regional economics. In each section, we illustrate best practices when implementing the research design by discussing several recent examples from the literature.

Given the importance of the use of economic models to aid in the specification of empirical models and interpret treatment effect estimates, we view the material on structural empirical modeling in Chapter 2 as complementary to the material discussed

in this chapter. Chapter 2 also considers the recovery of causal relationships in urban and regional data, but through making use of model formulations that are more involved than those considered in this chapter. The advantage of the structural approach is that it allows for the recovery of parameters that could never be identified with observational or experimental data alone. Estimates of a model's "deep" parameters facilitate evaluation of more sophisticated counterfactual simulations of potential policy changes than is possible with the less specific treatment effect parameters considered in this chapter. However, structural models are by their very natures full of assumptions that are most often stronger than the assumptions needed to make use of randomization to recover treatment effects. Additionally, because models can always be misspecified, such theory-derived treatment effects may be less credible than those whose data-based identification we discuss here. When possible, we present a unified treatment of causal relationships that can be interpreted in the context of an economic model or as stand-alone parameters.

While the field of urban economics has made considerable progress recently in improving its empirical methods, we hope that this chapter promotes further advances in the credibility of our empirical results by encouraging researchers to more carefully consider which particular treatment effects are being identified and estimated. In defense of our field, it is fortunately no longer acceptable to report regression results without any justification for the econometric identification strategy employed. Nonetheless, we hope we can go beyond this admittedly low bar. This includes dissuading ourselves from simply trying several instruments and hoping for the best without careful thought about the conditions under which each instrument tried is valid or the different causal effects (or combinations thereof) that each instrument may be capturing.

This chapter proceeds as follows. Section 1.2 develops an empirical framework as a basis for discussion, defines various treatment effects, and considers the importance of randomization. Section 1.3 briefly considers some of the consequences of using spatially aggregated data. Section 1.4 considers methods for recovering causal effects from purely observational data. Section 1.5 considers various ways of handling nonrandom sorting on unobservables leading up to a discussion of IV estimators. Section 1.6 describes the use of various types of RD designs. Finally, Section 1.7 concludes the chapter.

1.2. A FRAMEWORK FOR EMPIRICAL INVESTIGATION

In this section, we lay out an empirical framework that we use throughout this chapter as a basis for discussion and development. Our specification of the nature of the data-generating process facilitates consideration of the fundamental problem of causal inference. In particular, we emphasize the importance of determining the sources of variation in treatment variables that identify causal relationships of interest. Making use of explicit or pseudo random sources of variation in treatment variables is essential to credible

identification of causal relationships in any data set. We also consider the implications of the potential existence of heterogeneous causal effects of treatment variables on outcomes of interest.

In general, we are interested in causal relationships between a vector of "treatment" variables T and an outcome y. A flexible data-generating process for the outcome y can be represented by the following linear equation which holds for each observation i:

$$y_i = T_i\beta_i + X_i\delta_i + U_i + e_i. \tag{1.1}$$

For now, we think of observations as individuals, households, or firms rather than geographic regions. There is a vector of "control" variables X, which are observed. The vector U incorporates all unobserved components that also influence the outcome of interest. One can think of U as $W\rho$, where W is a vector of unobserved variables, and ρ is a set of coefficients that are never identified under any circumstances. We collapse $W\rho$ into U for ease of exposition. Given the existence of U, any remaining stochasticity e in the outcome y can be thought of as classical (uncorrelated) measurement error or, equivalently for statistical purposes, as fundamental stochasticity which may come from an underlying economic model and is uncorrelated with T, X, and U. We are also not interested in recovery of the coefficients δ_i on X_i, but it is useful for expositional purposes to define these coefficients separately from the coefficients of interest β_i.

Note that we express the relationships between predictors and the outcome of interest in a very general way by allowing coefficients to be indexed by i. In order to make progress on recovering the parameters of interest β_i for each individual, some further assumptions will be required. The linearity of (1.1) may incorporate nonlinear relationships by including polynomials of treatment variables and treatment-control interactions in T and polynomials of control variables in X.

It is often useful to think of (1.1) as being the "structural" equation describing the outcome of interest y, generated from an economic model of individual or firm behavior. For some outcomes such as firms' output or value added, this structural equation may result from a mechanical model such as a production function. More often for urban and regional questions, (1.1) can be thought of as an equilibrium condition in a theoretical model of human or firm behavior. In either type of model, we typically treat T, X, and U as "exogenous." This means that these variables are determined outside the model and do not influence each other through the model.

While the linearity in (1.1) may come from additive separability in the equilibrium condition, typically after a log transformation, we can more generally justify linearity in the empirical representation of a static model's equilibrium condition through implicit differentiation with respect to time. That is, if some model of individual behavior generates the equilibrium condition $y = f(T, X, U, e)$, differentiation yields an equation resembling (1.1) as an approximation, with partial derivatives of f represented by coefficients and each variable measured in first differences. That is,

$$\Delta y_i \approx \Delta T_i \frac{\partial f(T_i, X_i, U_i, e_i)}{\partial T} + \Delta X_i \frac{\partial f(T_i, X_i, U_i, e_i)}{\partial X} + \Delta U_i \frac{\partial f(T_i, X_i, U_i, e_i)}{\partial U} + \Delta e_i \frac{\partial f(T_i, X_i, U_i, e_i)}{\partial e},$$

in which Δ indicates differences over time. Note that this expression can be equivalently stated in semilog or elasticity form depending on the context. If the treatment status for every agent is the same in the base period and $\widetilde{X}_i$ includes 1, ΔX_i, X_i in the base period, and various interactions, this expression thus reduces to

$$\Delta y_i = \Delta T_i B(X_i, U_i) + \widetilde{X}_i D(U_i) + \varepsilon_i. \tag{1.2}$$

(1.2) closely resembles (1.1), with appropriate reinterpretation of y, T, and X, and can in principle form the basis for estimation.[1] Note that the error term ε incorporates both changes in unobservables U and changes in residual stochasticity e. Because it includes changes in unobservables, ε is likely to be correlated with ΔT. Moreover, we see that ε is likely to exhibit heteroskedasticity. As we explore further in Section 1.4, this "first difference" formulation has the advantage of differencing out any elements of U that are fixed over time, but has the potential disadvantage of increasing the variance of the error term.

There are a few important practical general implications of the exercise of deriving (1.2). First, first-differencing data is valuable as it allows the researcher to linearize nonlinear relationships, at least for small changes in y, T, and X. Second, it is really useful to have information from an initial period when the treatment variable is the same for all agents. Third, all but the simplest models deliver coefficients that are heterogeneous as functions of both observables and unobservables. If the model being estimated is sure to be the true data-generating process (which it never actually is), then coefficients in the linear (1.2) may allow for recovery of estimates of some or all of the model's parameters. Even if individual model parameters cannot be identified, $B(x, u)$ represents the causal effect of T on y for an agent with characteristics (x, u). Regardless of the true underlying data-generating process, this is an object which is often of inherent interest to researchers. Finally, the exact specification of the control set $\widetilde{X}$ depends crucially on the underlying economic model; thus, this object can very easily be misspecified. For this reason, there are distinct advantages to using estimators that permit elements of $\widetilde{X}$ to be dropped.

Our discussion of the recovery of treatment effects in this chapter primarily examines "total effects" of treatments on outcomes, or full derivatives $\frac{dy}{dT}$. Of course, the decomposition of these total effects into direct and indirect effects, in which causal links from the

[1] In some contexts, it may be appropriate to differentiate over space rather than time. We leave a more complete discussion of this issue to the Chapter 3 on spatial methods by Gibbons et al. and our discussion of the RD research design in Section 1.6.

treatment to the outcome operate both independently and through the treatment's influence on other predictor variables, is also interesting (Pearl, 2009). The distinction between total effects versus direct and indirect effects is a statistical restatement that the generic economic model with the equilibrium condition $y = f(T, X, U, e)$ used as a starting point above includes only exogenous variables on the right-hand side. Decomposition into direct and indirect effects of treatment is often recovered in economics applications by using some model structure, since indirect effects by definition operate through some endogenous channel. In Sections 1.4 and 1.5, we return to discussions of direct and indirect effects in the contexts of considerations of properties of particular estimators.

1.2.1 A binary treatment environment

Though urban and regional applications often involve more complicated environments, we begin by considering the case in which the treatment is binary. Analysis of this simple case is a straightforward point of departure as it is well understood in the statistics literature going back to the classic treatment of Rubin (1974), and discussed extensively in Holland (1986), and in the economics literature going back to Roy (1951). Because the recovery of causal relationships in environments with binary treatment environments is also discussed at length by DiNardo and Lee (2011), we leave the development of many details to them. Indeed, much of our mission in this chapter is to extend their discussion of various empirical identification strategies to environments in which the treatment is continuous and the data are spatially indexed. The simplicity of the binary treatment environment is important, however, as properties of the various estimators we discuss in this chapter are well known for the binary treatment case.

On the basis of the setup in (1.1), a binary treatment variable yields the following equation for each treatment level, in which treated observations receive $T = 1$ and untreated (control) observations receive $T = 0$:

$$y_i^0 = X_i\delta_i + U_i + e_i,$$
$$y_i^1 = \beta_i + X_i\delta_i + U_i + e_i.$$

These two equations describe the potential outcome for each agent i if that agent were not treated and if that agent were treated, respectively. The resulting causal effect of treatment for agent i is thus β_i. When all agents in the population are considered, the result is two separate distributions of outcomes y, one for each treatment status. In evaluating the effects of the treatment, we typically aim to characterize differences between elements of these two distributions.

It should be immediately evident from this example with binary treatments that it is impossible to recover each particular β_i without further assumptions on the data-generating process, even with ideal data. This is the fundamental problem of causal inference: no agent can simultaneously be in both the treated group and the untreated group.

That is, there is no counterfactual available for individual members of any population or sample, since each agent is either treated or not treated. In the language of Holland (1986), there is not "unit homogeneity" if each observation has its own treatment effect. Even if we had panel data such that we could observe individuals before and after treatment, the contextual environment of "before treatment" versus "after treatment" is collinear with the treatment itself. That is, the context can be thought of as an element of X (or U if not accounted for). Each individual and time period combination would have its own observation index, and therefore its own treatment effect.[2]

To make progress on recovering information about causal effects of treatment, we need to limit ourselves to considering how to identify elements of the distribution of treatment effects over the population. This recognition brings up the fundamental issue that we address in this chapter: how to identify groups of agents that are similar on both observables and unobservables but who have received different levels of treatment. If the treatment effect is different for each agent, then the agents are so fundamentally different by definition that recovering any information about the distribution of β_is is a hopeless endeavor. To make progress on identification of treatment effects, we must put restrictions on the coefficients in the above equations such that they are not unique across individuals, but instead may be unique only across individuals with different observables and unobservables. One general formulation for doing so is the following:

$$y_i^0 = X_i D(U_i) + U_i + e_i,$$
$$y_i^1 = B(X_i, U_i) + X_i D(U_i) + U_i + e_i.$$

Because the distribution of treatment effects captured in the $B(\cdot)$ function depends on the characteristics of the treated agent only and not on the identity of each agent itself, we can imagine finding another agent with the same observable and unobservable characteristics with whom the treated agent can be compared. In practice, since we do not by definition know the unobservable characteristics of any agent, we do not have any way to recover the "marginal" treatment effect (MTE) for any particular unobserved type U without the imposition of an economic model, as in Heckman and Vytlacil (2005). Instead, depending on how the treatment is assigned, we are potentially able to recover various model-agnostic statistics about the distribution of $B(X, U)$ over the population. Note that we restrict the coefficients on observables X to be functions only of U. To account for potential nonlinear impacts of X (that interact with U), one can define X to include polynomial terms and interactions.

[2] In a few cases, researchers have assumed that unboservables do not differ over time and have attempted to estimate individual treatment effects by causing individual fixed effects to interact with a treatment variable. The work of De La Roca and Puga (2014) is an example in the context of estimating causal effects of city sizes in labor market histories on individuals' wage profiles. Section 1.3 discusses in detail the assumptions needed for fixed effects identification strategies like this to deliver credible estimates of causal effects.

1.2.2 A taxonomy of treatment effects

Before returning to an empirical model with continuous treatments, it is useful to consider the various treatment effects that may be of interest in the context of the binary treatment environment. These treatment effect definitions generalize with minor modifications to the continuous treatment case, as explained below. In the following sections, we carefully consider which treatment effects can be identified with each of the estimators that we consider.

One way of conceptualizing the binary treatment environment is as the existence of two counterfactual distributions in the population y^0 and y^1 which differ only because of treatment status. The restrictions on the empirical model formulated above force the difference between these two distributions for agents of a given type (x, u) to be $B(x, u)$.

The most closely related causal effect is the MTE. As in Heckman and Vytlacil (2005), we define MTE(x, u) as the causal effect of treating an individual with characteristics $X = x$ and $U = u$:

$$\text{MTE}(x,u) \equiv E[y^1 - y^0 | X = x, U = u] = B(x,u).$$

While the MTE is a useful construct, it is only possible to recover any particular MTE within the context of a specified economic model. This is because the MTE is indexed by unobservable U, which is an object that the researcher can never know directly, but can only assign to individuals through the structure of a model. Heckman and Vytlacil (2005) consider a simple generalized Roy-type sorting model (Roy, 1951) on the basis of which they identify the full distribution of MTEs. All other treatment effects can be viewed as weighted averages of various combinations of MTEs.

Unconditional quantile treatment effects (QTEs) of Abadie et al. (2002) provide information about the distribution of treatment effects, as indexed by the realization of outcome variables. The QTE for quantile τ is the difference in the τth quantile of the y^1 and y^0 distributions, which in this case is the τth quantile of the distribution $f(B(X, U))$. QTEs are informative about whether the treatment differentially influences different parts of the distribution of the outcome of interest. Athey and Imbens (2006) show how to estimate the full counterfactual distributions y^1 and y^0 without any functional form assumptions assuming treatment randomization, thereby allowing for calculation of all QTEs. The difficulty with QTEs is that their recovery typically requires randomization to apply very broadly to the distribution of potential outcomes, which rarely occurs. QTEs do not provide information about the unobserved characteristics of agents to whom they apply, though one can similarly define QTEs over the conditional distributions of unobservables only $f_x(B(x, U))$ given $X = x$.

Perhaps the commonest treatment effect of interest is the average treatment effect (ATE). The ATE describes the mean treatment effect averaged over all members of the population with a particular set of observed characteristics x and is represented as follows:

$$\mathrm{ATE}(x) \equiv E(y^1 - y^0 | X = x) = \int B(x, U) \mathrm{d}F(U | X = x).$$

Often, rather than being interested in the ATE for a particular subpopulation, researchers may be interested in the ATE for the full population:

$$\mathrm{ATE} \equiv E(y^1 - y^0) = \int B(X, U) \mathrm{d}F(X, U).$$

As with QTEs, it is important to recognize that the ATE is not easily recovered in most empirical contexts without strong model assumptions. The reason is that in the absence of widespread randomization, there are some groups which either always receive the treatment or never receive the treatment. Since calculation of the ATE requires knowing the MTE for the full joint distribution of (X, U), the portions of the support of $f(X, U)$ which are in only the treated state or the untreated state must have their MTE distributions inferred by model assumption. Depending on the approach, the model used to recover these MTE distributions may be statistical or economic.

The local average treatment effect (LATE), first defined by Imbens and Angrist (1994) and also discussed by Bjorklund and Moffitt (1987), is the average effect of treating the subset of the joint distribution of X and U that has been induced into (or out of) treatment through explicit or pseudo randomization. Suppose that an "instrument" Z allows the researcher to manipulate the probability that agents end up in the treatment group or the control group. Imagine manipulating Z from values z to z', where $\Pr(D = 1 | Z = z) > \Pr(D = 1 | Z = z')$ for all combinations of X and U.[3] The resulting LATE is defined as

$$\mathrm{LATE}(z, z') \equiv \frac{E[y | Z = z] - E[y | Z = z']}{\Pr(D = 1 | Z = z) - \Pr(D = 1 | Z = z')}. \tag{1.3}$$

That is, the LATE captures the change among those newly treated in the mean of y for a change in the fraction treated. This definition can be interpreted as a simple weighted average of all MTEs:

$$\mathrm{LATE}(z, z') = \frac{\int B(X, U)[\Pr(D = 1 | Z = z, X, U) - \Pr(D = 1 | Z = z', X, U)] \mathrm{d}F(X, U)}{\Pr(D = 1 | Z = z) - \Pr(D = 1 | Z = z')}$$

Here we see that the weights depend on the relative probability of being induced into the treatment group rather than the control group by the change in the instrument Z. In principle, this manipulation of the instrument could cause some increase in the

[3] It is also possible to define the LATE for cases in which the variation in Z induces movement into treatment for some types and out of treatment for other types. However, to the extent that such bidirectional flows are unobserved, the resulting object is very difficult to interpret as it conflates positive treatment effects for some agents with negative treatment effects for others.

probability of treatment for all observed and unobserved types. Heckman and Vytlacil (2005) consider LATE's interpretation in the context of a structural model in which each value of U explicitly determines the choice into or out of treatment. That is, the range of U for which there is identification is the range over which the manipulation of the instrument Z induces membership in the treated group that would not otherwise have occurred.

Unlike the MTE, QTE, and ATE, the LATE is defined on the basis of the empirical context because the empirical context determines (z, z'). The LATE is an important concept because it is often the only treatment effect that can be identified when there exists randomization over only some subset of the support of the joint distribution of X and U.[4]

The intention to treat (ITT) is the average effect of offering the treatment. This is a policy-relevant treatment effect for many program evaluations since many of those offered the opportunity to participate in government programs do not accept it. Suppose that agents in the group offered treatment have $Z = 1$ and those in the group not offered treatment (the "control" group) have $Z = 0$. Those who would accept the offer of treatment if available have $D = 1$ and others have $D = 0$. We assume that those in the control group cannot under any circumstances procure the treatment. That is, if $Z = 0$, D necessarily equals 0. However, those in the treatment group may refuse treatment, such that $Z = 1$ and $D = 0$ for some agents. Given this environment and assuming that membership in the group offered treatment is randomized, we have

$$\begin{aligned}\text{ITT} &\equiv E(y|Z=1) - E(y|Z=0)\\ &= E(y^1|Z=1, D=1)\Pr(D=1|Z=1) - E(y^0|Z=0, D=1)\Pr(D=1|Z=0)\\ &= E(y^1 - y^0|D=1)\Pr(D=1)\\ &= \int B(X,U)\Pr(D=1|X,U)\mathrm{d}F(X,U).\end{aligned}$$

This simple expression for ITT assumes that because of treatment randomization, $E(y^0|Z=1, D=0) = E(y^0|Z=0, D=0)$. Like other treatment effects considered above, ITT can be conditioned on X.

The treatment on the treated (TT) is the average effect of the treatment for those who would choose to accept an offer for treatment. This can be expressed as

$$\begin{aligned}\text{TT} &\equiv E(y^1 - y^0|D=1)\\ &= \frac{\int B(X,U)\Pr(D=1|X,U)\mathrm{d}F(X,U)}{\int \Pr(D=1|X,U)\mathrm{d}F(X,U)}.\end{aligned}$$

Notice that TT is typically greater in magnitude than ITT, because it is defined only for those with $D = 1$. In the above expression TT is written as the MTE weighted by the probability of treatment for each combination of X and U, with high values of U

[4] LATE can also be conditioned on values of X provided that there is some variation in Z for $X = x$.

presumably being more likely to select agents into treatment, normalized by the mass of the portion of the distribution $f(X, U)$ that selects agents into treatment. The closely related treatment on the untreated is the average effect of the treatment for those who choose not to accept the treatment offer. Notice that if every agent were to accept the offer of treatment, ITT = TT = ATE.

To be more concrete about the differences between these various treatment effects, we compare them in the context of the Moving to Opportunity (MTO) experiment, which randomized Section 8 housing vouchers to two treatment groups of public housing residents in five cities in the mid 1990s. Data on a control group that was not offered vouchers were also collected. Households in the "Section 8" treatment group received only a housing voucher, which subsidized rent in any apartment whose landlord would accept the voucher. The "experimental" treatment group was additionally provided with counseling and was required to move to a neighborhood with a poverty rate below 10% for at least 1 year. Baseline information about households in the treatment and control groups was collected prior to randomization and in various posttreatment periods. Let us consider labor market earnings as an example outcome for the Section 8 treatment group.

Each household in the population of public housing residents has some particular observed and unobserved characteristics (x, u). MTE(x, u) is the causal effect on earnings of moving a household with characteristics (x, u) out of public housing into a Section 8 apartment of its choice. Because the MTE is conceptualized such that a different value of U is assigned to each household with a different treatment effect, there is only one possible MTE per (x, u) combination. The QTE for quantile τ is the comparison of earnings quantile τ in the treatment group relative to the control group in an environment in which all treated households comply with the treatment. ATE(x) is the average difference in earnings for the treatment group versus the control group for those households with characteristics x assuming all treated households comply. ITT is the average difference in earnings between treatment and control groups, whether or not those in the treatment group accepted the voucher. TT is the average difference in earnings between those in the treatment group that accepted the offer of the voucher and those in the control group who would have accepted the voucher if it had been offered. In the binary treatment context, LATE is identical to TT, since the housing voucher offer manipulates the probability of leaving public housing for a Section 8 subsidized apartment. As we discuss further in Section 1.5, LATE terminology is most commonly invoked when IV estimation is used to recover causal links from a continuous treatment to an outcome. For example, since the offer of the housing voucher caused treated households to move to lower-poverty neighborhoods at a higher rate than control households, one can conceptualize the LATE of neighborhood poverty on household earnings. This LATE applies only to the types of households induced by the treatment to move to lower-poverty neighborhoods.

1.2.3 Continuous treatments

With continuous treatments, instead of imagining two counterfactual states for each agent in the population, y_i^0 and y_i^1, we imagine a continuum of counterfactual states, which we denote y_i^T. To be consistent with the literature and allow parameters of the data-generating process to be tractably estimated using standard techniques, we restrict our attention to the following linear model which puts only a few additional restrictions on (1.1):

$$y_i = T_i B(X_i, U_i) + X_i D(U_i) + U_i + e_i. \tag{1.4}$$

While it is commonly implemented as a linear equation, there is no need to interpret (1.4) as strictly linear since T could be formulated as a vector of treatments which are a polynomial in one continuous treatment variable, just as X can incorporate higher-order terms. Note that we typically do not consider the possibility that $B(X_i, U_i)$ and $D(U_i)$ can be functions of the treatments themselves.

Each of the treatment effects discussed above applies to the continuous case as well with only slight modification (Heckman et al., 2006). In general, treatment effects for a continuous treatment must also be indexed by the specific values of the treatment variables to which they refer. For example, the prior subsection defines the ATE for moving from treatment value 0 to treatment value 1, which could be written as $\text{ATE}_{0,1}(x)$. Because of the linearity assumption in (1.4), (or that $B(\cdot)$ is not itself a function of T), any treatment effects in the continuous case are identical regardless of which unit iteration of the treatment variable is considered. That is, $\text{ATE}_{0,1}(x) = \text{ATE}_{q,q+1}(x)$ for all q. Therefore, each of the treatment effects defined above maintains its definition in the continuous case with minimal adjustment for any arbitrary unit iteration in T, understanding, of course, that this comes by assumption and may not hold beyond the support of T.

It is important to emphasize that while we sometimes consider the case $B(X_i, U_i) = \beta$, most empirical research should recognize the possibility that there exists some "essential" heterogeneity across agents in the causal effects of treatment variables of interest. If that is the case, the assumption of a homogeneous treatment effect can lead to invalid interpretations of estimation results. In the course of this chapter, we lay out which elements of the distribution of β can be recovered with various estimators commonly applicable to recovering causal relationships of interest to urban and regional economists.

1.2.4 Randomization

One difficulty that comes out of this section's motivation for using an economic model of behavior as a starting point for empirical investigation is that as researchers we can never be sure what the "correct" empirical specification is for an estimating equation because we never know the true data-generating process for y. Even if we did know what variables belong in X and W, it is often the case that different particular economic models

have the same such exogenous variables as inputs into the data-generating process. Structural parameters are informative only in the context of the structural model within which they are defined. Therefore, rather than concerning ourselves with recovering structural parameters, we often find it fruitful to concentrate empirical work on recovery of particular treatment effects, which then may also have interpretations in the context of specific structural models. The main challenge in doing so is that there are almost always unobservables that influence y yet may be correlated with the treatment variables of interest. This is the classic econometric identification problem.

One path toward a solution to this identification problem is to recognize that if there is randomization in treatment variables T, it is unnecessary to observe X and U to recover some information about $B(X, U)$. The role of randomization is that it assigns different values of T to agents with the same X and U. That is, it creates comparable treated and untreated populations. Of course, the reason that we need randomization to achieve this, rather than simply some assignment rule based on observables, is that U is unobserved. By its very nature, pure randomization of T over the population balances the joint distribution of X and U for all treatment levels.

With pure randomization of T over the population and a data-generating process described by (1.4), it is straightforward to see that the OLS estimate of β in a simple regression of y on T yields the ATE. In particular,

$$p\lim(\hat{\beta}_{\mathrm{OLS}}) = E[B(X,U)] = \mathrm{ATE},$$

which is simply a difference in means between treatment and control groups. Intuitively, this result comes about because randomization ensures that the full distribution of individuals in the population receives each level of treatment. One may wish to control for X in this regression in order to reduce the variance of the error term, and as a result, the standard error of $\hat{\beta}_{\mathrm{OLS}}$. By extension, it is also straightforward to estimate a series of specific ATEs ATE(x) by regressing y on T interacting with dummy variables capturing various portions of the support of X. For example, if a researcher is interested in knowing the ATE among those with observable attributes in sets A and B, which partition the full support of X, the researcher could estimate the following regression equation by OLS:

$$y = T1(X \in A)\beta_A + T1(X \in B)\beta_B + X\delta + \varepsilon.$$

In this equation, $1(\cdot)$ is the indicator function. The result is that $p\lim(\hat{\beta}_{\mathrm{AOLS}}) = E[B(X,U)|X \in A]$. That is, $\hat{\beta}_A$ as estimated by OLS captures the ATE for the portion of the X distribution in set A. It is important to recognize here that the distributions of unobservables in sets A and B may be quite different. There is no way to know whether the reason that OLS estimates of β_A and β_B may be different is because set A contains individuals with a distribution of observables (on which they have been partitioned) or unobservables correlated with these observables different from those in set B. One can extend this procedure to estimate a broader set of ATEs.

Recovery of treatment effects with simple OLS regression typically requires explicit treatment randomization. However, implementation of randomized controlled trials (RCTs) can be quite challenging and expensive. Duflo et al. (2008) provide a practical guide and toolkit for researchers wishing to introduce randomization as a part of the research design in a field experiment.[5] A general issue with all experiments is that it is rarely possible or practical to randomize a treatment over the full population. Small sample sizes often make inference about treatment effects which apply to subpopulations difficult. For this reason, estimation of treatment effect heterogeneity is often limited to simple interactions of T and X in a regression model.[6]

Individual participation in randomized trials is rarely mandatory. This means that those participating in an experiment may differ on unobservables from other populations of interest. Randomization of treatment thus often occurs over only a subset of the population of interest. For example, in the MTO experiment, housing vouchers were offered only to those who had the motivation and initiative to show up to an initial meeting at which the program was described. While it is possible to see whether these MTO subjects differ on some observables from remaining public housing residents, they may differ more markedly on unobserved attributes that also influence well-being measures and labor market outcomes of interest. That is, because the sample over which the treatment is randomized is almost always self-selected on some unobservables, any results necessarily only apply to this self-selected group. As a result, there is likely to be some portion of the support of the distribution of U for which treatment effects cannot be recovered without extrapolation. Equally important is that it is common for many agents offered treatment not to accept it. That is, even though the treatment and control groups have the same distribution of unobservables, those who do and those who do not actually get treated do not. In these contexts, it is typically infeasible to recover the full distribution of treatment effects, and researchers focus on estimating ITT and TT.

Ludwig et al. (2013) summarize estimated treatment effects of MTO using data from 10–15 years after program implementation. They find that the program had no detectable effect on economic outcomes, youth schooling, or physical health. However, they do find some positive effects on mental health and measures of subjective well-being. This evidence follows up the study of Kling et al. (2007), which reports positive effects of MTO on behavioral outcomes for girls but negative effects for boys 5–8 years after implementation. Galiani et al. (2012) leverage the MTO randomization to estimate a structural model of neighborhood choice. They use their estimates to recover counterfactual

[5] Most RCTs conducted by American researchers can be found at the AEA RCT Registry website. Even though this is a voluntary registry, the AEA encourages the registration of all new RCTs.

[6] When researchers are interested in recovering treatment effects for certain subpopulations, these groups are typically oversampled relative to their share of the full population. When using data for these groups to recover other treatment effects or parameters,one should apply sampling weights to ensure that these oversampled groups do not contribute disproportionately to the estimates.

distributions of poverty rates in neighborhoods chosen by voucher recipients given alternative voucher assignment policies that were never actually implemented. They find that take-up of the voucher offer is severely reduced by restricting destination neighborhoods to the point of being counterproductive if such restrictions limit destination choice too much. This is a good example of a study that uses clean identification to recover parameters of a structural model, and ultimately a broader set of treatment effects than could be recovered using atheoretical methods alone.

There are many potential concerns about extrapolating the causal impacts of the MTO experiment from program effects to neighborhood effects. Indeed, the neighborhood improvements caused by housing voucher randomization are conflated with the disruption of moving, changes in neighborhood quality may not have been sufficiently large to generate statistically measurable effects, voucher recipients select particular destination neighborhoods of their choice, and MTO results may not generalize to other settings. Moreover, the MTO experiment reveals little about the effects of moving the approximately 50% of households who chose not to leave public housing despite receiving the offer of a housing voucher. Despite those caveats, the MTO experiment has produced among the most convincing estimates of the impacts of changes in neighborhood quality on individual outcomes. In particular, these results have weakened the "spatial mismatch hypothesis" view that low neighborhood quality and poor job access promote high rates of unemployment in poor neighborhoods (Kain, 1992).

Explicit treatment randomization has also generated data that are informative about the internal and external effects of improved housing conditions. Galiani et al. (2013) examine effects of the randomized provision of prefabricated homes for slum dwellers in El Salvador, Mexico, and Uruguay. They find that beneficiaries exhibited no improvement in labor market outcomes but improved general well-being and housing conditions relative to a control group. Freedman (2014) finds that tax credits for home improvements that were allocated to applicants by lottery in St Louis, Missouri slightly increase the value of neighboring homes.

As with treatment effect estimation in most settings, one important general consideration about using data with treatments allocated by lottery is the potential existence of general equilibrium effects. Interpretation of average differences in outcomes between treatment and control groups as treatment effects requires that the stable unit treatment value assumption (SUTVA) (Cox, 1958) of no direct or indirect influence of the treatment of one observation on outcomes of control observations must hold. For example, if in the MTO environment some control group households were to hear about neighborhood relocation options from experimental group households and act on this information, the SUTVA would be violated. To avoid this problem, many RCTs in development economics randomize treatment at the village level rather than the household level. However, since many questions of interest to urban and regional economists are fundamentally about the operation of cities rather than villages, this strategy may be of limited use in our field.

Nonetheless, RCTs for answering urban and regional questions will likely become commoner as evaluating the impacts of urban policy interventions becomes more important in developing countries, where urbanization is rapidly occurring.

One additional setting in which explicit randomization has been used to learn about causal effects is in analysis of peer effects. Without randomization, it is very difficult to get around the problem that people very likely sort into peer groups, including classes in school and friendship networks, on correlated unobservables. Sacerdote (2001) uses the random assignment of freshman roommates at Dartmouth College to recover estimates of peer effects in college performance. Bayer et al. (2009) use the random allocation of juvenile prisoners to cells to recover information about peer effects in recidivism. However, using data collected about experimentally manipulated peer groups among freshmen at the Air Force Academy, Carrell et al. (2013) find negative peer effects on the lowest-ability group members, perhaps partly because of endogenous subgroup formation which separated them from their highest-ability peers. The randomization of students into classrooms in the first year of the Project Star program in Tennessee has also been used to recover estimates of peer effects; see Graham (2008), for example.

Much of the remainder of this chapter considers strategies for recovering treatment effects for settings in which explicit treatment randomization is not available. Section 1.4 essentially considers various strategies for indirectly controlling for unobservables U. Section 1.5 considers strategies for identifying and effectively making use of pseudorandom variation in treatments. Section 1.6 considers how best to make use of discontinuities in treatment intensity. As a general principle, we reiterate that whatever the empirical strategy used, it is critical for the researcher to understand the source of variation that is providing identification for parameters of interest. Thinking through such identification arguments often reveals the existence of potential endogeneity problems in which the treatment variable may be correlated with elements in W and/or the extent to which the treatment effects being estimated apply only to certain narrow subpopulations.

While perhaps not ideal, there are many contexts in which neither randomization nor credible strategies for controlling for unobservables are available to recover treatment effects of interest. The main alternative viable strategy is to explicitly model the heterogeneity and sorting equilibrium and recover treatment effects through model simulation. Holmes and Sieg discuss such structural options at length in Chapter 2. It should be emphasized that making use of model structure requires much stronger assumptions than are needed for a randomized treatment to yield credible treatment effects. Moreover, because no model completely describes the data-generating process, the credibility of model-derived results still requires careful consideration of the sources of variation in the data that are identifying estimates, and whether these sources of variation are random (unlikely), or at least plausibly uncorrelated with mechanisms that could be important but are not explicitly modeled.

1.3. SPATIAL AGGREGATION

Before delving into the specifics of various identification strategies and econometric estimators, we briefly explore the implications of having a data structure that is spatially aggregated above the individual, household, or firm level. Such a data structure may be imposed by a data provider, be chosen by the researcher because the treatment is administered to regions rather than individual agents, or be chosen by the researcher in order to strengthen the empirical strategy. When imposed by the researcher, spatial aggregation of data is often carried out to alleviate concerns about SUTVA violations, in which spillovers occur between spatially proximate geographic units with different levels of treatment. Researchers often aggregate data to the local labor market or metropolitan area level in order to avoid this potential problem.

Suppose that the treatment and outcomes are observed at some level of spatial aggregation such as census tracts or zip codes, indexed by j. In the case of a binary treatment that is applied to the same fraction of the measure of each (x, u) in each location, a strong assumption, the equation of the data-generating process becomes

$$\widetilde{y}_j = S_j \widetilde{B}(X_j, U_j) + \frac{1}{N_j}\sum_{i(j)} X_i D(U_i) + \widetilde{U}_j + \widetilde{e}_j.$$

In this equation, tildes ($\sim$) indicate sample means over all observations in j. N_j is the total number of observations in j, S_j is the fraction of observations in region j that were treated, and $\widetilde{B}(X_j, U_j) = \int B(X, U)\mathrm{d}F_j(X, U)$, where $F_j(X, U)$ is the joint cumulative distribution function of X and U in unit j. Notice that because of the heterogeneous coefficients $D(U_i)$, $\frac{1}{N_j}\sum_{i(j)} X_i D(U_i)$ cannot in general be simplified into some simple function of means $\widetilde{X}_j$. Therefore, controlling for mean values of each element of X does not appropriately control for observables about individual agents unless $D(U_i) = \delta$. Instead, the full distribution of X within each j shows up in the aggregate equation. Therefore, in this sort of aggregation environment it makes sense to control not just for the mean but also for the full distribution of each observable characteristic if possible. Therefore, if regional means of X are all that is observed about control variables, we can think of other elements of the within-j distributions of X as being part of $\widetilde{U}_j$.[7]

In the case of a more general continuous set of treatments and heterogeneous treatment effects, aggregation gives rise to the nonseparable treatment terms $\frac{1}{N_j}\sum_{i(j)} T_i B(X_i, U_i)$ replacing $S_j \widetilde{B}(X_j, U_j)$ above. Estimation of statistics about $B(X, U)$ is

[7] If the goal is to recover the treatment effect averaged across individuals (rather than regions j), one should weight any estimation by N_j. Doing so allows the more populous regions to influence the estimates more than the regions that have few agents. If, however, the goal is to recover the treatment effect averaged across regions, one should not weight such an estimation.

thus quite difficult without further assumptions about the underlying data-generating process. One common simplifying assumption is that of perfect sorting across regions. This assumption can be justified to an approximation as the equilibrium in a Tiebout (1956) sorting model like that specified by Epple and Platt (1998). With this structural assumption, which applies more accurately to finer levels of spatial aggregation, we have a resulting data-generating process given by

$$\widetilde{\gamma}_j = T_j B(X_j, U_j) + X_j D(U_j) + U_j + \widetilde{u}_j.$$

Because of homogeneity within each region j in X and U, we need only index these elements by j to represent any and all quantiles of their distributions in j. Without this sort of homogeneity assumption, it becomes clear that while perhaps some progress can be made with spatially aggregate data in recovering information about $B(X, U)$, making use of micro data or the structure of a sorting model would be preferable for recovering treatment effects, even in a context with explicit treatment randomization.

Rather than having an underlying data-generating process described by (1.4), in some contexts we determine the treatment itself at the local area level. For example, the federal Empowerment Zone (EZ) program treated certain census tracts with various forms of government subsidies, and the Clean Air Act treated certain counties with pollution reductions. Often with these sorts of policies, we are interested in the effects on local residents or firms. At the local area (e.g., census tract) level, the data-generating process is thus

$$\widetilde{\gamma}_j = T_j \widetilde{B}(X_j, U_j) + \frac{1}{N_j}\sum_{i(j)} X_i D(U_i) + \widetilde{U}_j + \widetilde{u}_j. \tag{1.5}$$

As above, in this equation, $\widetilde{B}(X_j, U_j)$ denotes the average effect of the treatment in each region j given the distribution of X and U in unit j. In this case we do not need assumptions about homogeneity of populations in local areas or homogeneity of treatment effects to make some progress in recovering information about $B(X_j, U_j)$. In particular, given global randomization in T_j and no changes in location that is related to receiving the treatment, an OLS regression of mean outcomes on the treatment dummy weighted by the population of each region j yields a coefficient on the treatment with a probability limit of the ATE, by the law of iterated expectations.

One key assumption here is that the composition of the population of each region j does not respond to the treatment. This assumption is a strong one. If the treatment changes the amenity value of certain locations, we may expect certain types of people to move out of untreated locations into treated locations, thereby changing the joint distribution of the population in each location $f_j(X, U)$ and breaking the orthogonality between T and $\widetilde{U}$ needed to identify $E[\widetilde{B}(X_j, U_j)]$, even with initial treatment randomization across space. While one can look in the data for such resorting on observables X, including such intermediate outcomes as controls may bias treatment effect estimates since such intermediate outcomes are now endogenous. Cellini et al. (2010) provide

an alternative strategy to deal with such situations in the context of a dynamic model. Once again, making use of an economic model of behavior that takes sorting into account would aid econometric identification.

The final aggregation structure that we consider here is one in which each metropolitan area or other large spatial aggregation is an observation, potentially at different points in time. The sorts of questions that lend themselves to be answered with such highly aggregated data are those for which the full data-generating process must be described at the local labor market level and subsumes a set of complicated micro level interactions. One can conceptualize this by aggregating (1.4) to the local labor market level while recognizing that (1.4) incorporates the simultaneous existence of heterogeneous treatment effects, heterogeneous treatments across agents within each local labor market, and spatial lags. For example, measuring the size of agglomeration within local labor markets (Glaeser et al., 1992; Henderson et al., 1995) and measuring the effects of highways on urban decentralization (Baum-Snow, 2007) or urban growth (Duranton and Turner, 2012) lend themselves to be considered using aggregate data structures. Sorting difficulties or other general equilibrium effects that would make econometric identification difficult when examining micro data are aggregated away in these examples. For these types of applications, we typically think of the treatment as occurring at the metropolitan area level because even those metropolitan area subregions that were not explicitly treated are indirectly influenced by the treatment through general equilibrium effects. For this sort of empirical strategy to be successful, it is essential that the data be at a sufficient level of spatial aggregation that there are minimal links across observations. If the data are not sufficiently aggregated, the endogeneity problem caused by spillovers across spatial units of observation may be very difficult to handle.

The following equation captures the data-generating process for some local labor market aggregate statistic y such as population or GDP:

$$y_k = T_k B(X_k, U_k) + X_k D(U_k) + U_k + u_k. \tag{1.6}$$

In this equation, k indexes local labor markets or other highly aggregated spatial units such as states, which are spatial aggregates of j. Depending on the context, the coefficients may be heterogeneous as a function of the distribution of household or firm characteristics in k or other summary attributes of k, either of which we denote as the couple (X_k, U_k). If the treatment effect of interest concerns effects on individuals, this equation is analogous to (1.5), and one thus may need to consider any potential resorting of the population across k in response to the treatment. If instead the goal is to recover treatment effects on metropolitan area aggregate measures, this equation is perfectly analogous to (1.4), and exhibits all of the same challenges with respect to econometric identification and the interpretation of estimates, though the mechanisms may be subtle owing to sorting. One difference from more micro analyses which in practice is often important is that typically the number of observations is quite small. For example, historical data on

metropolitan areas in the United States sometimes include information for only 100 regions nationwide. With such a limited number of observations, statistical power becomes weak very quickly if treatment variables are defined too nonparametrically. Therefore, little statistical power may be available to recover a lot of information about the $B(\cdot)$ function in (1.6).

One word of general caution about estimation of empirical models with spatially indexed data is that standard errors are likely to be understated without implementation of an appropriate correction. This is because common elements of unobservables U in nearby observations manifest themselves as correlated errors. Spatially and/or temporally correlated unobservables W (or, equivalently, unexplained components of γ) is why such spatially correlated errors ensue. Bertrand et al. (2004) discuss block bootstrap (Efron and Tibishirani, 1994) and clustering (Moulton 1990, 1986) methods to account for these problems in environments in which there is a fixed number of observations per cluster and the number of clusters increases toward infinity. Cameron et al. (2008) compare various procedures for calculating standard errors with a small number of clusters using Monte Carlo simulation. Their results indicate that the "clustered wild bootstrap-t" procedure generates the most accurate test statistics when clusters are independent and the number of clusters is small. Bester et al. (2011) discuss estimation of heteroskedasticity-autocorrelation consistent standard errors and generalized cluster procedures for conducting inference with spatially correlated errors when clusters are not independent and the number of clusters is fixed but the number of observations within each cluster goes to infinity.

Now that we have specified the possibilities for the types of data-generating processes that show up most often in urban and regional empirical applications, we consider various empirical strategies for recovering treatment effects.

1.4. SELECTION ON OBSERVABLES

While having a source of explicit or pseudo randomization is typically the preferred way to recover the most credible causal relationships in data, there are many important questions that do not lend themselves easily to this sort of empirical strategy. As such, in this section we consider options for recovering causal parameters of interest in the absence of such randomization. It should be clear that estimating (1.4) by simple OLS recovers only the ATE, $E[B(X, U)]$, in the unlikely event that T is uncorrelated with U, or that T is fully randomized. This section thus explores alternatives to simple OLS that do not involve explicit or implicit randomization, and therefore may not account for the influence of unobserved variables in the economic relationship of interest. These other methods are fixed effects, DD, and matching estimators. We emphasize that these methods can sometimes most successfully be used in tandem with each other and/or with other empirical strategies discussed elsewhere in this chapter. Key decisions in implementing nonexperimental estimators in

many contexts are the choices of treatment and particularly control groups. The primary goal in choosing a control group is to choose a set of observations for which the distribution of unobservables is likely to be similar to that in the treatment group. Below we present some formal options for doing this by examining the distribution of observables, though it is standard to assign all untreated observations to the control group in a robustness check while explicitly accounting for differences in observables. For example, the final subsection discusses estimators that reweight observations in the control group to match its distribution of observables with that in the treatment group.

We emphasize that it is almost as much an art as a science to determine the most convincing identification strategy. This determination depends crucially on the setting and the structure of the available data. For example, if the available data include an individual level panel, fixed effects methods are feasible. If the data are structured as two repeated cross sections, DD may be most feasible. Even within the identification strategies that we explore, the details of implementation require many decisions. As such, we hope this section provides a general guide to the available options, along with their advantages and pitfalls and examples of their use in published research, rather than specific recipes for carrying out empirical work.

1.4.1 Fixed effects methods

Fixed effects and panel methods can be used when there are multiple observations per agent or spatial unit. Inclusion of fixed effects in a regression is intended to remove all unobserved characteristics that are fixed over time, or across space if multiple agents are observed in the same spatial unit, from the error term. This means that any components of unobservables that are fixed over time are controlled for through inclusion of fixed effects. DD, whose discussion we delay to the following subsection, is a particular identification strategy which typically incorporates fixed effects. We consider the use of panel data on individuals or firms, homes, and spatial units at various levels of aggregation, respectively.

A generic fixed effects regression specification, for individuals i at times t, is as follows:

$$y_{it} = T_{it}\beta + X_{it}\delta + \alpha_i + \varepsilon_{it}. \tag{1.7}$$

In the absence of the fixed effects α_i, β is identified by comparing outcomes at different levels of T both between and within agents i. Inclusion of fixed effects is equivalent to differencing y, T, and X relative to sample means within each i. Therefore, β in a fixed effects regression such as (1.7) is identified by comparing changes in y for different changes in T (or first derivatives) within agents. Variation in T between agents is not used to recover information about β. With more than two time periods, one can also estimate (1.7) on first-differenced data, which identifies β by comparing DD (or second derivatives) within agents.

Because β in the above regression is identified from variation in T over time within agents, those agents with more variation in T influence the estimate of β more.

Therefore, if treatment effects are heterogeneous at β_i across agents, $\hat{\beta}_{\text{FE}}$ does not capture the ATE, but rather captures some combination of individual treatment effects weighted by each individual's contribution to econometric identification. Indeed, Gibbons et al. (2013) derive that the fixed effects estimator for β is

$$\hat{\beta}_{\text{FE}} = \sum_{i=1}^{I} \left(\frac{N_i}{N} \hat{\beta}_i \frac{\widehat{\text{Var}}(\widetilde{T}_i)}{\widehat{\text{Var}}(\widetilde{T})} \right).$$

In this expression, $\widetilde{T}$ is the residual after projecting T onto other covariates, including fixed effects. $\text{Var}(\widetilde{T}_i)$ is the variance of this object within i, while $\text{Var}(\widetilde{T})$ is its variance overall in the data. Commensurate with the intuition given above, this coefficient is a particularly weighted combination of individual treatment effects. If such treatment effect heterogeneity is important, one can instead estimate individual treatment effects β_i in the following interacted regression equation, in which $\widetilde{\alpha}_i$ are fixed effects that are distinct from α_i in (1.7):

$$y_{it} = T_{it}\beta_i + X_{it}\delta + \widetilde{\alpha}_i + \varepsilon_{it}.$$

Then, these individual treatment effects can be averaged at will. For example, Wooldridge (2005) suggests the "sample-weighted" treatment effect, which is identical to the ATE if each agent is sampled the same number of times, as $\sum_{i=1}^{I} \left(\frac{N_i}{N} \hat{\beta}_i \right)$. Unfortunately, in many applications there is no variation in T across time for some agents, making it impossible to identify their individual treatment effects, nor the sample-weighted treatment effect nor the ATE.

In the urban economics literature, regression models with individual fixed effects have been extensively employed to try to understand the effects of city size or density on wages, and by extension productivity, through agglomeration economies. Glaeser and Maré (2001), Combes et al. (2008), Baum-Snow and Pavan (2012), and De La Roca and Puga (2014) among others estimate Mincerian regressions of log wages on experience, some measure of city size, and individual fixed effects that resemble the following formulation:

$$\ln w_{it} = \beta[\text{citysize}]_{it} + X_{it}\delta + \alpha_i + \varepsilon_{it}. \tag{1.8}$$

Identification of the city size coefficient β comes from individuals' moves across cities of different sizes. Note that citysize can be specified as a vector of treatment dummy variables or as a more continuous measure of city size or density. In the context of the data-generating process (1.4), the role of the individual fixed effects α_i is to control for the time-invariant component of U_i. As a consequence, one interpretation of α_i is as indicators of time-invariant ability or skill. These studies consistently find strong relationships between wages and city size even after controlling for individual fixed effects, though inclusion of individual fixed effects typically reduces the coefficient on city size or density

by about one-third to one-half. The prima facie implication of this result is that while there is a causal effect of city size or density on wages, there is also important positive sorting of high fixed effect (unobserved ability) individuals into larger cities that must be accounted for in any evaluation of agglomeration economies through wages.

The greatest threat to identification in such studies is that some unobservable that may predict wages and labor market attachment is correlated with decisions to move across cities of different sizes. Individuals with positive unobserved personal productivity shocks may be more likely to move to larger cities. Potential omitted variables could be marital status, home foreclosure, winning the lottery, moving to care for a sick relative, losing one's job, or moving to start a better job. These unobserved variables are time-varying components of U_i, though one could argue that variation in job offer or separation rates across cities should be counted as part of the variation in city productivity.[8] If such endogeneity of the move decision is important, making use of only the within-individual variation in city size may actually introduce more bias to the estimate of β than not including fixed effects and making use of comparisons between individuals. Fixed effects models make no use whatsoever of any potential information in the "control" group of individuals who never moved but who may have unobservables similar to those of individuals who are located in cities of different sizes.[9]

Heterogeneous treatment effects are also of first-order importance for consideration for two reasons. First, those who move more frequently are weighted more heavily in the calculation of the city size effect β. If more able people with higher wage growth potential move more often, they receive higher weight in the estimation of β. If this is the case, their types U are oversampled from the MTE distribution $B(X, U)$, and β may thus highly overstate the ATE. Moreover, the fact that moves are more prevalent soon after labor force entry means that the fixed effect estimator recovers the causal effect of city size primarily for those early in their working lives and not for the average stage in one's career. In the language of Section 1.2, we can think of labor market experience as an element of X and the MTE $B(X, U)$ as being larger at certain values of X than at others. Therefore, even without an omitted variables problem, the fixed effects estimator in this case recovers a particular LATE which may overstate the ATE because of both oversampling of high-ability individuals and moves early in the life cycle. Failure to incorporate this treatment effect heterogeneity into the empirical specification can bias the fixed effects estimates, in which case

[8] While differences across cities of different sizes in the arrival rate of job offers and separations are typically considered one mechanism for agglomeration economies, this data-generating process is inherently dynamic with the job match as an important state variable. Therefore, in the context of an estimation equation such as (1.8) which could never capture such a data-generating process, it is more straightforward to treat search and matching as showing up in U_i rather than as part of the coefficient on citysize. Baum-Snow and Pavan (2012) consider how to recover estimates of the importance of search and matching in agglomeration economies using a dynamic structural model.

[9] Observations about individuals that remain in the same location during the sample period do help increase the precision of the estimates of δ.

they would not be good measures of individual ability. These observations are made by De La Roca and Puga (2014) using Spanish data and Baum-Snow and Pavan (2012) using US data in their assessments of the effects of city size on wages.

Absent some source of randomization in treatment, the literature has heretofore been only partially successful at handling the potential endogeneity of moves without the use of a structural model, as in Baum-Snow and Pavan (2012). De La Roca and Puga (2014) have made some progress in recovering information about heterogeneity in treatment effects and in the amount of selective migration by allowing β and δ to differ by individual fixed effects α_i. They estimate their empirical model iteratively by first capturing fixed effects and then interactions until a stable set of fixed effects is estimated. They find that returns to experience are larger for higher-ability individuals in larger cities, but wage level differences do not depend much on ability. By examining the distributions of fixed effects in different locations, Combes et al. (2012) argue that selective migration is not a big enough phenomenon in French data to drive a large wedge between the true ATE and OLS estimates of city size coefficients, a conclusion that Baum-Snow and Pavan (2012) and De La Roca and Puga (2014) share.

Another context in which fixed effects methods are standard is in hedonic models. With use of data on home prices from transactions and home characteristics, fixed effects remove time-invariant unobserved home characteristics that contribute to home value. Repeat sales hedonic models (which originally excluded observable home characteristics) are the basis of housing price indices going back to Bailey et al. (1963), including the S&P Case–Schiller index (Case and Shiller 1987, 1989). Repeat sales indices are constructed using a regression model such as the following, typically with some adjustment for potential heteroskedasticity in the errors:

$$\ln p_{ijt} = \beta_{jt} + X_{ijt}\delta + \alpha_i + \varepsilon_{ijt}.$$

In this equation, $\ln p_{ijt}$ is the log transaction price of home i in market j at time t. The fixed effects α_i account for unobserved fixed home characteristics, β_{jt} captures the home price index for market j at time t, and X_{ijt} includes time-variant home characteristics. Rosenthal (2014) uses a similar specification with homeowner's log income on the left-hand side to account for fixed unobserved home characteristics in his investigation of filtering.

This repeat sales specification also forms the basis for several studies which evaluate the willingness to pay for various local public goods and services, including various aspects of actual and perceived school quality. For example, Figlio and Lucas (2004) examine how housing prices and mobility changed when new school report cards in Florida provided the public with condensed information about local public school quality. To achieve this, they partition $\beta_{jt} = \mu_{jt} + T_{jt}\beta + X^s_{jt}\gamma$. In this expression, T_{jt} is a vector of dummy variables for the locally zoned elementary school's state-assigned grades in attendance zone j and X^s_{jt} is a vector of school characteristics that go into construction of the grade. The estimated treatment effect β reflects a causal effect of school grades on local housing values.

Econometric identification comes from the assertion that reported grades were a surprise and involve considerable random noise, and therefore are unlikely to be correlated with neighborhood unobservables. Moreover, all time-varying observable attributes about local schools are controlled for in X^s and there is no possible correlation between better school grades and time-invariant influences on home prices because of controls for home fixed effects α_i. The interpretation of the β vector is thus the average effects of changing neighborhood school grades on local home prices. It is important to recognize that the hedonic valuation of an A grade is likely identified mostly from variation in homes in quite wealthy neighborhoods with a strong taste for school quality, because these are the locations in which schools had variation in the A grade dummy, whereas the hedonic valuation of an F grade is identified primarily from poor neighborhoods. Therefore, these are local treatment effects which apply only for the subset of the full distribution of homes that experienced variation in relevant grades.

Beyond the local nature of such β estimates, clear interpretation of hedonic regression results requires careful consideration of the data-generating process for home prices. Hedonic models starting with that of Rosen (1974) indicate that shifts in the quality of one attribute of a product may induce a shift in the composition of buyers of that product. In addition, the elasticity of housing supply determines the extent to which such quality increases may be reflected in prices versus quantities. In this context, an increase in perceived local school quality and the resulting outward shift in local housing demand may be driven by wealthier residents looking to move into the neighborhood. These wealthier residents may seek higher quantities of housing services, and the demand shift may spur developers to increase the housing stock. Therefore, even if a regression such as that specified above is well identified and β is a causal effect of school grades on home prices, it is not straightforward to interpret it as the marginal willingness to pay by any particular potential buyer for this increase in local public goods. Indeed, Figlio and Lucas (2004) demonstrate that A grades induced sorting of higher-achieving students into the schools' attendance zones—students whose parents are likely willing to pay more for school quality than the families they replaced. Greenstone and Gallagher (2008) consider how to recover estimates of welfare consequences of toxic waste cleanups using home price data aggregated to the census tract level. In general, because neighborhoods with different attributes have different household compositions, β in the standard hedonic equation above recovers only the marginal willingness to pay under the strong assumption that all households have homogeneous preferences over neighborhood attributes.[10]

[10] Recovery of heterogeneity in marginal willingness to pay for neighborhood attributes typically requires additional economic modeling. The article by Bayer et al. (2007), which we discuss in Section 1.6, shows how to recover the distribution of willingness to pay for school quality and sociodemographic characteristics of neighborhoods using a structural model married with an RD identification strategy to control for unobserved neighborhood characteristics. Kuminoff et al. (2013) present a review of the many structural models of supply and demand equilibrium in housing markets that can be used to recover willingness to pay for public goods.

Another setting in which fixed effects have been effectively used is to control for unobserved neighborhood characteristics in cross-sectional or repeated cross-sectional data with geographic identifiers. A typical specification is as follows, in which j indexes local units such as census tracts or block groups:

$$y_{ijt} = b_{jt} + T_{ijt}\beta + X_{ijt}\delta + \varepsilon_{ijt}.$$

Campbell et al. (2011) use this sort of specification to examine the effects of forced sales, through foreclosure or resident death, for example, on home prices. In their context, the treatment is a dummy that equals 1 if a home transaction was a forced sale or 0 otherwise. Census tract-period fixed effects b_{jt} control for the possibility that homes may be more likely to be force sold in lower socioeconomic status neighborhoods. Autor et al. (2014) use a similar specification to measure the effects of rent decontrol in Cambridge, Massachusetts, on housing values and Ellen et al. (2013) do so for examining the effects of foreclosures on crime. Bayer et al. (2008) use census block group fixed effects to control for sorting and unobserved job options in their evaluation of job referral networks in which each observation is set up as a worker pair. Their basic identifying assumption is that those looking for a home can at best find one in a particular block group rather than a particular block, yet they find that living on the same block is strongly related to working on the same block conditional on individual and block fixed effects.

One somewhat arbitrary feature of the standard use of spatial unit fixed effects is the assignment of each observation to only one particular spatial region fixed effect, even though observations typically differ in their centrality in such regions. That is, those observations on the edge of a census tract or block group may receive some spillover from neighboring tracts' unobserved characteristics and not all locations within spatial unit j are likely to have exactly the same set of unobservables. To the extent that the treatment differs as a function of location (e.g., because of spillovers from nearby regions) in a way that is correlated with subregion level unobservables, estimates of β would be biased and inconsistent. One way of accounting for microgeographic fixed effects that alleviates this problem is by using a spatial moving average specification. We replace b_{jt} in the above regression equation with

$$b_{ijt} = \sum_k \left[W[\text{dist}(i,k)]\widetilde{b}_{kt} \right].$$

Assuming knowledge of the exact location of each i and indexing spatial units by k, one can take a weighted average of nearby fixed effects. In this expression, $W(\cdot)$ is a weighting function that equals 1 when the distance between observations is 0 and declines with distance or adjacency. This weighting function could have one estimated parameter ρ and could take a standard form with exponential or linear decay, as in $W(d) = e^{-\rho d}$ or $W(d) = \max[1 - \frac{d}{\rho}, 0]$. Estimation of the fixed effects and $\widetilde{b}_{kt}$ and decay parameter ρ could be implemented by nonlinear least squares or the generalized method of moments (GMM). One could also generalize this specification to incorporate a separate

individual fixed effect for smaller spatial aggregations. This is a particular case of the spatial moving average model which is discussed at greater length in Chapter 3 by Gibbons et al. and in which the endogenous portion of the error term is controlled for.

We delay our discussion of fixed effects estimators applied to data aggregated to the local labor market level to the following subsection.

1.4.2 Difference in differences methods

The DD identification strategy is a particularly common application of fixed effects. To be viable, it typically requires a data structure in which "treatment" and "control" groups are observed in multiple treatment environments, at least one of which is the same for the two groups. Typically, one difference is over time such that in initial periods the treatment has not yet been implemented, though in some studies treatment and control groups are instead compared in different locations or contexts other than time periods. Differencing over time (or across contexts), often implemented by including group or subgroup fixed effects, purges from the error term any time-invariant unobservables U that may differ between treatment and control groups. Differencing across groups, typically implemented by including time fixed effects, purges from the error term time-varying elements of unobservables U that are the same in the treatment and control groups. The primary identification assumption in DD estimators is that there are no time-varying differences in unobservables that are correlated with the treatment. The DD strategy can be generalized to the case in which the treatment is given to different observations at different points in time and/or to incorporate additional "differences."

Implementation of the DD identification strategy is straightforward. With data in levels, one can think of the coefficient of interest as that on the interaction between the treatment group and a posttreatment dummy. One can equivalently calculate a simple DD in mean outcomes for the treatment group versus the control group in the posttreatment period versus the pretreatment period. The following regression equation, which can be estimated by OLS, incorporates the standard DD specification for panel data, in which β is the coefficient of interest. It includes period fixed effects ρ_t, individual fixed effects κ_i (which can be constrained to be the same within entire treatment and control groups, or subsets thereof), and the treatment variable of interest T_{it}, which is only nonzero for the posttreatment period:

$$y_{it} = \rho_t + \kappa_i + T_{it}\beta + X_{it}\delta + \varepsilon_{it}. \tag{1.9}$$

One may also wish to control for X. However, if unobservables are differenced out by the DD estimator, observable controls X should be differenced out as well. Therefore, in most cases controlling for X will not matter for estimating β since X is orthogonal to T conditional on the fixed effects. Below we consider the consequences of controlling for X in cases in which X is correlated with T. At least one period of data in both the

pretreatment environment and the posttreatment environment is required in order to recover a DD estimate. To ease exposition, we denote period 0 as the pretreatment period and period 1 as the posttreatment period.

Depending on the context, the DD estimator may consistently recover different treatment effects or no treatment effect at all. In the context of the data-generating process described by (1.5), consistent estimation of any treatment effect requires that any shocks to U are not correlated with the treatment. Put another way, any differences in the composition of the treatment and control groups in period 1 versus period 0 must be random. In mathematical terms, the key identification assumption is

$$\begin{aligned}(E[U|T_1=1,t=1]-E[U|T_1=1,t=0])-(E[U|T_1=0,t=1]\\ -E[U|T_1=0,t=0])=0.\end{aligned} \tag{1.10}$$

This assumption is valid as long as there are no time-varying unobservables that differ across treatment and control groups and predict the outcome of interest. Differencing between treatment and control groups over time (or, equivalently, including group fixed effects κ_i) purges all fixed differences between the treatment and control groups, even if the distribution of unobservables is different in these two groups. Differencing across groups at each point in time (or, equivalently, including time fixed effects ρ_t) controls for differences in the pretreatment and posttreatment environments. The comparison between these two differences thus recovers a treatment effect averaged over the distribution of observables and unobservables in the treatment group provided that any differences in unobservables between the treatment and control groups are not time varying.

It is straightforward to derive that $\hat{\beta}_{\text{OLS}}$ only consistently estimates $\text{ATE} = E[B(X, U)]$ if all of those in the treatment group receive a full treatment, none in the control group do, and the treatment is fully randomized, meaning that the treatment and control groups have the same joint distribution of observables and unobservables. However, because the DD estimator is typically applied in settings in which some selection into treatment can occur, it is unlikely that an ATE is recovered. This selection into treatment can be conceptualized as existing for spatial units or for individuals within spatial units. Because spatial units cannot select out of treatment, a well-identified DD estimator recovers the TT for data-generating processes such as (1.6), in which the object of interest is at the level of spatial units rather than individual agents. If we think of the treatment as being applied to spatial units but individual agents to be the objects of interest as in (1.5), we can also think of the DD identification strategy as delivering TT for spatial units. However, if those for whom $T_{it} = 1$ can refuse treatment (as is typical) and the set of agents offered treatment is representative of the overall population, the DD estimator at best recovers ITT as defined at the individual agent level. If the researcher has information about the probability that agents who received the offer of treatment accept it, this ITT estimate can be rescaled to produce an agent-level estimate of TT.

It is common to use the DD identification strategy to analyze situations in which a treatment is applied to specific regions and outcomes of interest are at the individual level. Though the researcher may care about such individual-level outcomes, outcomes may only be reported at spatially aggregated levels such as census tracts or counties, as in (1.5). In this context, the treatment group is in practice identified as treated locations, in which individuals are presumably more likely to be treated. An important threat to identification in such a setting in which aggregate data are used is the potential resorting of individuals (on unobservables) between the treatment and control groups. If the treatment is valuable to some people in untreated areas, they may migrate to treated areas, thereby displacing some that do not benefit as much from the treatment. Such sorting on unobservables that is correlated with (and happens because of) the treatment would violate a version of the identification condition (1.10) with aggregate data (which looks exactly the same because of the law of iterated expectations), thereby invalidating the DD identification strategy.

One indicator pointing to a high likelihood of differing distributions in unobservables in the treatment and control groups existing before treatment versus after treatment is differing pretreatment trends in outcomes for the two groups. For example, if the control group experienced a positive shock in period 0 and is reverting toward its long-run mean between periods 0 and 1, that would cause the DD estimator to overstate the true effect of the treatment. Similarly, if the treatment group received a negative shock prior to treatment, this would similarly make it look like the treatment had a causal effect when all that is different is simply mean reversion. Indeed, in some settings agents are selected for treatment because of low values of observables, some of which may be transitory. This threat to identification is colloquially known as the "Ashenfelter dip" (Ashenfelter, 1978).

As empirical researchers, we often have access to a data set with some observables that are available to be included as controls. It is not clear that these variables should always be used. Indeed, one should think of most elements of X as analogous to the W variables that make up U, except that they are observed. Including these elements of X should thus not influence the estimate of β in (1.9) if the DD strategy is sound, though they may reduce its estimated standard error. However, in some settings there may be elements of X that describe attributes of agents on which they sort in response to the treatment. This phenomenon may arise, for example, in cases in which the treatment and control groups are defined as geographic units rather than individuals. If such sorting across treatment/control groups is fully predicted by attributes, then controlling for X is appropriate as it rebalances the treatment and control groups in both periods. That is, the two identification requirements on conditional expectations of U listed above may be true conditional on X even if not unconditionally. However, if inclusion of X in (1.9) influences the estimate of β for this reason, and sorting on observables exists, it is likely that sorting on unobservables also exists, thereby invalidating the identification assumptions listed above. Therefore, comparison of estimates of β including and

excluding controls for X is some indication as to whether sorting on unobservables may be biasing the coefficient of interest.

In some settings, it may be the case that some elements of X respond directly to the treatment. For example, it may be that incomes increased in areas that received federal EZ funding at the same time that income influences the outcome of interest γ such as the home ownership rate. In this example, controlling for income changes the estimate of β because absent controls for income and assuming $E(T\varepsilon) = 0$, β measures a full derivative, whereas controlling for income, β captures a partial derivative. However, controlling for an endogenous variable such as income runs the risk of violating the basic identification condition $E[X\varepsilon] = 0$, thereby rendering $\hat{\beta}_{\text{OLS}}$ inconsistent. This violation would occur if, in this example, income were a function of T and some unobservable in ε, thereby making T correlated with ε as well. Therefore, a less fraught approach for recovering the partial effect of T on γ holding income constant is to directly estimate the treatment's effect on income (by making it an outcome), and then separating out that effect directly to recover the residual effect of the treatment on the real outcome of interest γ, which does require knowledge of $\frac{\partial \gamma}{\partial X}$ from elsewhere. Note that a standard robustness check in DD estimators is to control for pretreatment X variables interacting with time. These are exogenous to the treatment because the treatment is 0 in all pretreatment observations.

Ham et al. (2011) use several flavors of the DD estimator to evaluate various impacts of several local economic development programs, including the federal EZ program. This program's first round started in 1994 and provided tax credits to businesses to hire local residents, reduced borrowing costs for community development, and committed billions of dollars in community development block grants to these communities. EZ status was awarded to a group of poor census tracts in each of 11 cities selected for the program. Ham et al. (2011) use census tract data to evaluate the effects of EZ status on poverty, labor earnings, and employment, and argue that EZs improved all of these outcomes. Their initial analysis uses data from the 1990 and 2000 censuses, with nearby tracts acting as a control group for EZ tracts. One may be concerned that tracts with negative economic shocks prior to 1990 were selected to be EZ tracts because of this, violating the assumption of common pretreatment trends. To handle this, the authors introduce a third difference—between 1980 and 1990—making this a differences in differences in differences (DDD) estimator. In practice, one can implement a DDD estimator by carrying out the DD estimator exactly as laid out above on first-differenced data for each of two time spans. The advantage of the DDD estimator in this context is that any common linear trends in unobservables in treatment and control groups are differenced out, eliminating any potential bias because of an "Ashenfelter dip." However, any higher-order (e.g., quadratic) trends are not accounted for, nor is the possibility that the treatment status itself changed tract compositions. That is, if the treated tracts and control tracts have a different composition of residents and firms in 1990 and 2000 that is partly unobserved, part of any estimate recovered may reflect this composition shift.

The evaluation of the EZ program by Busso et al. (2013) also employs DD and DDD strategies but instead uses census tracts in areas that were barely rejected for inclusion in EZs in other cities as the control group. As with the Ham et al. (2011) study, the disadvantage of using this control group is that these locations were likely rejected for inclusion in the first round of the EZ program because they were slightly less distressed than those that ended up being included. The advantage of the Busso et al. (2013) approach is that they use an estimator that reweights the control group on observables to be more comparable than the equal weighting given by standard OLS. This study is further discussed in the following subsection, along with the use by Kline and Moretti (2014) of the same estimator in tandem with a DD identification strategy to evaluate the effects of the Tennessee Valley Authority on long-run outcomes.

Greenstone et al. (2010) use a DD estimator to recover the effects of large new industrial plants on incumbent plants' total factor productivity. Their treatment group is the set of counties which received new industrial plants and their control group is the set of counties that were barely rejected for the siting of an industrial plant. The idea is that counties chosen for these new plants should be similar on unobservables to those barely rejected, and indeed the paper shows evidence that the treatment and control groups of counties have similar pretreatment observable characteristics and pretreatment trends. Incumbent plant outcomes in treatment and control counties are compared before and after the arrival of new industrial plants, as are differential posttreatment trends in these outcomes. Their results indicate that these large new industrial plants had significant spillovers of about 5% on average to incumbent plant total factor productivity, with larger effects in closely related industries. This is direct evidence of positive agglomeration spillovers.

Figure 1.1, taken from Greenstone et al. (2010), is an instructive illustration of how the DD strategy can be implemented. The top panel shows the average total factor productivity in incumbent manufacturing plants in treatment and control counties each year from 7 years before to 5 years after the arrival of the new large industrial plant in each treatment county, normalized to zero in the year prior to entry. This plot shows that pretreatment trends were very similar for treatment and control groups, with these trends diverging starting at period 0. The bottom panel shows the differences between treatment and control groups in each period, and a marked shift up in these differences after period 0. The simplest DD estimator, which could be estimated with a specification such as (1.9), is indicated in the lower panel as the gap in average gaps between treatment and control groups after treatment relative to before treatment. The authors extend the simplest DD specification (1.9) to recover information about dynamic responses to the treatment. Greenstone and Gallagher (2008) use a similar strategy to argue that cleaning up hazardous waste sites had negligible effects on housing prices, housing quantities, population, and population composition in nearby census tracts. These can be thought of as special cases of the RD estimator discussed in detail in Section 1.6.

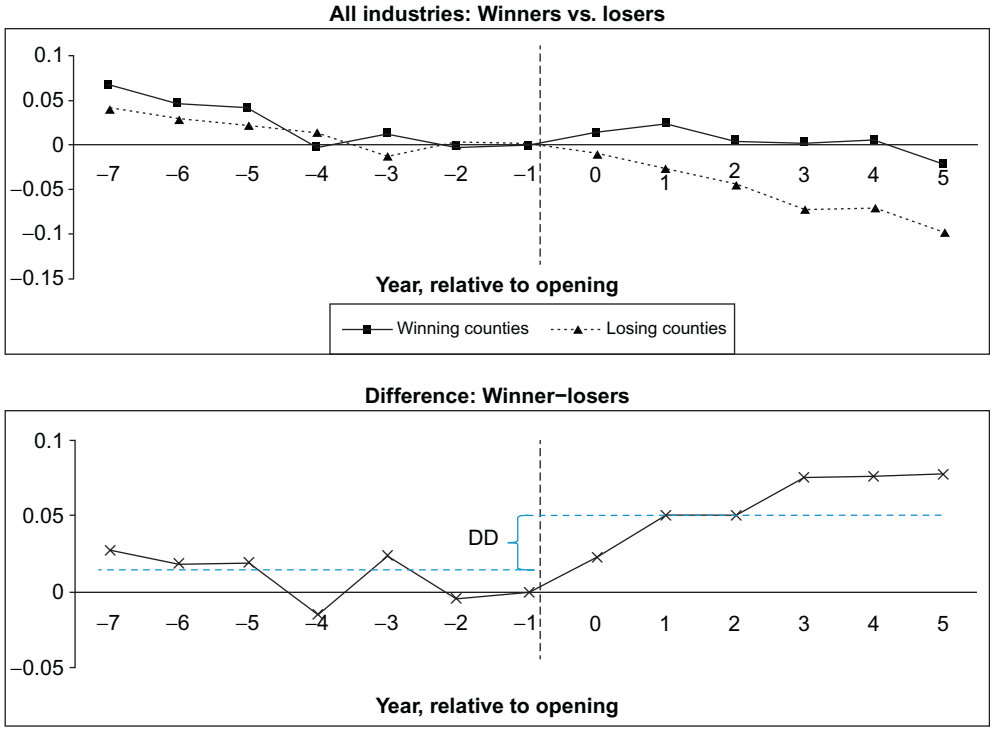

Figure 1.1 TFP of incumbent firms in "Winning" and "Losing" Counties from Greenstone et al. (2010).

A nonexhaustive list of other prominent empirical studies in urban and regional economics which make use of DD or DDD identification strategies follows. Field (2007) examines the labor supply effects of land titling in Peru by comparing squatters to those with land title in areas with recent title provision. Costa and Kahn (2000) examine the extent to which large cities better foster "power couple" location or formation by examining differences between large and small cities and various demographic groups who have more versus fewer constraints on forming a dual-worker couple over time. Linden and Rockoff (2008) show that home values declined nearer to the homes of sex offenders moving into neighborhoods relative to those further way. In a similar vein, Schwartz et al. (2006) demonstrate that new subsidized housing developments in New York City increased values of nearby homes more than those further away. These two spatial DD studies employ more flexible specifications than in (1.9) because they allow for full spatial variation in responses to treatment to be captured in the regression specification.

The DD identification strategy has also been applied in settings with data-generating processes that operate at the metropolitan area or county levels. For example, Redding and Sturm (2008) show that after the division of German, population growth rates in

areas near the West German border were less rapid, whereas after reunification they were more rapid than elsewhere in the country. This study uses differences over time and between border and nonborder regions. Baum-Snow and Lutz (2011) evaluate the effects of school desegregation on residential location patterns by race. Identification in this study comes from comparing metropolitan areas that had recently been treated with those that had been not by treated by 1970 or 1980. The years 1960 and 1990 bookend their study, in which all metropolitan areas in the sample were untreated and treated, respectively. This is implemented as regressions of the form of (1.9) in which i indexes metropolitan areas and T_{it} is a binary for whether the central school district in the metropolitan area is under court-ordered desegregation at time t. Because of variation in the timing of treatment, the compositions of the treatment and control groups depend on the year. Identification in this case depends on there not being unobservables that are correlated with the timing of treatment. Because all metropolitan areas go from being untreated to treated during the sample period exactly once, the resulting treatment effect estimates apply broadly within the sample used and can be interpreted as ATEs for the set of metropolitan areas considered.

Abadie et al. (2014) describe how to implement a method of "synthetic controls" as a way to construct the control group in DD-type estimation environments. This method is often applied when the treatment group is very small or consists of just one unit but there are many candidate control units. Instead of cherry-picking a few particular units for the control group that may or may not represent good counterfactuals for treated units, the authors show how to use a weighted combination of all available control observations, with weights set to represent how close they are to treated observations. The resulting treatment effect estimate is $\hat{\beta} = Y_{1t} - \sum_{j=2}^{J=1} w_j^* Y_{jt}$, where Y_{1t} is the outcome at time t for the treated unit (or an average among treated units), Y_{jt} are the outcomes for the control units, and w_j^* is a set of weights. These weights are chosen in a way that minimizes some distance criteria between predetermined characteristics of the treated units and the predetermined characteristics of the control units. For example, Abadie and Gardeazabal (2003) and Abadie et al. (2010) choose the vector $W^\star$ as the value of W that minimizes

$$\sum_{m=1}^{k} v_m \left(X_{1m} - X_{0m} W\right)^2.$$

In this expression, X_{1m} denotes the average value characteristic m for treated observations, while X_{0m} is the vector of the same characteristic for control observations, all calculated prior to treatment. Further, v_m is a measure of the importance of characteristic m, which can be chosen to be proportional to the predictive power of characteristic m for the outcome. The problem with the synthetic controls approach is that the choice of predetermined characteristics and distance criteria can be ad hoc, and one may end up giving too

much weight to control units that are not appropriate counterfactuals owing to differential pretrends or other unobserved components. But the interesting characteristic of this approach is that it allows for simple construction of generalized control groups. In the following subsection, we analyze matching methods that more directly use this idea.

1.4.3 Matching methods

The DD and fixed effects identification strategies discussed thus far are only credible if the treatment group is observed prior to treatment and there are no time-varying unobservables correlated with the treatment. However, there are may settings in which either a pretreatment period is not observed or time-varying unobservables that are different in the treatment and control groups and may influence outcomes are likely to exist. One potential solution to such problems is to use an estimator that makes use of information about observables to try to infer information about unobservables. We focus on cases in which the treatment is binary.

As a starting point, consider trying to recover information about the causal effect of treatment in the constant coefficient version of the data-generating process in (1.1) using cross-sectional data. That is, suppose the true data-generating process is as follows:

$$y_i = T_i\beta + X_i\delta + W_i\rho + u_i.$$

Note that because this is a constant coefficient model by assumption and if W and T are uncorrelated, the OLS estimate of β is the ATE. Trying to estimate this equation by OLS leads to biased estimates of β if some unobservables W are correlated with the treatment. One common heuristic method for addressing such potential bias is to estimate this equation by varying the set variables in the control set X. The idea is that as variables are moved from unobservables W to observables X, any reductions in estimates of β indicate omitted variables bias is influencing these estimates. If β is stable with inclusion of additional controls, there is more confidence that omitted variables bias is not a problem. Crucial for this method to be informative is for the R^2 of the model to increase as variables are moved from W to X. If R^2 does not increase, these are irrelevant variables with true coefficients of 0. As crucial is that the set of controls in X is in some sense representative of the full set of possible control variables $[X\ W]$. At the end of this subsection, we consider how examples in the literature have attempted to correct the bias using a proportional selection bias assumption, formalizing this intuition.

Standard practice for attempting to estimate causal effects in the absence of implicit randomization is to employ a propensity score matching estimator. The idea of such estimators, originally proposed by Rosenbaum and Rubin (1983), is to compare outcomes of individuals with the same propensity to be treated, some of whom receive treatment and others of whom do not. The underlying "propensity score" $P(X)$ is the probability of being treated, and depends on observables only. This score can be estimated by a probit or logit with a flexible specification.

The main difficulty with matching estimators is that they assume that selection into or out of treatment is fully predicted either by observables or by unobservables that do not predict the outcome of interest. If unobservables influence both outcomes and whether agents receive treatment, treated and untreated observations are not comparable for any given propensity score, and matching estimators are not informative about any treatment effect. If unobservables influence outcomes but not the probability of treatment, matching estimators are still informative about treatment effects. This intuition is the same intuition about potential threats to identification in OLS regression, so it is not surprising that OLS is a particular form of a propensity score matching estimator. Heckman and Navarro-Lozano (2004) demonstrate that matching estimators can be quite sensitive to the conditioning sets used and argue that control function methods in which choices are more explicitly modeled are more robust. We briefly consider such methods at the beginning of the following section.

Formally, the following conditions must hold in order for a propensity score estimator to produce consistent treatment effect estimates (Wooldridge, 2002):

$$E(y^0|X,T) = E(y^0|X), E(y^1|X,T) = E(y^1|X). \tag{1.11}$$

These conditions say that those receiving the treatment have the same mean outcomes whether they are treated or not as those who do not receive the treatment. That is, actually receiving treatment cannot predict outcomes in either the treated or untreated counterfactual states of the world. These assumptions are sometimes called "selection on observables" because they allow selection into treatment to be fully predicted by X, but not by U. This assumption implies $\mathrm{TT}(x) = \mathrm{ATE}(x)$, but not necessarily that $\mathrm{TT} = \mathrm{ATE}$.

Provided that the data set being used is rich with observables, there is information in the propensity score coupled with treatment status about whether unobservables correlated with the treatment may be an important source of bias. If there is very little overlap in the range of the propensity score in which both treated and untreated observations exist, this indicates that since treatment and control groups differ on observables, they may be more likely to differ on unobservables as well. Consequently, the range of the propensity score for which there is overlap is the region of the data for which the propensity score matching estimator is providing more convincing identification. As a result, it is often informative to graph the density of treated and untreated observations against the propensity score, plus the implied treatment effect at each level of the propensity score, to get a sense of the treatment effect over the range of the propensity score for which unobservables are less likely to be driving selection into treatment. To calculate such a treatment effect, one can nonparametrically estimate the conditional expectations $E(y|P(X),\ T = 1)$ and $E(y|P(X),\ T = 0)$ and then take the difference for every value of $P(X)$. This uses the argument that unobservables act in some sense like observables.

Figure 1.2 provides two schematic diagrams which match these suggested graphs. Panel (a) shows the density of treatment and control group observations as a function

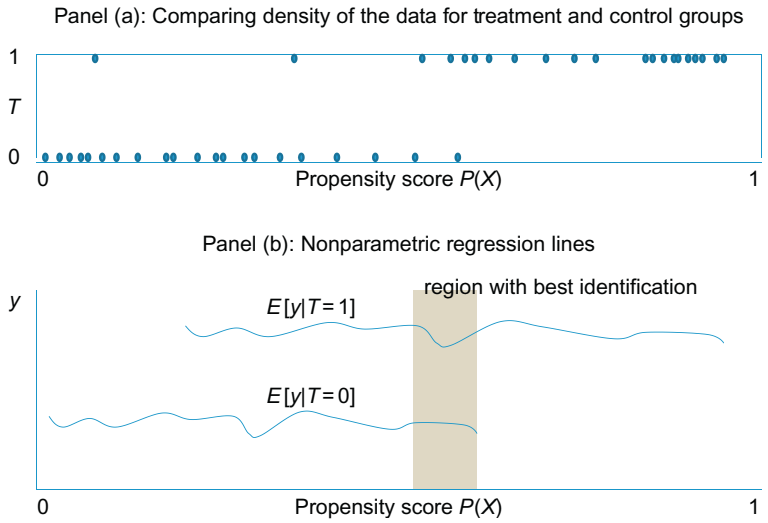

Figure 1.2 Schematic diagrams for matching estimators.

of the propensity score. In this example, there is very little overlap between the treatment and control groups. Indeed, just a few observations from both groups have similar propensity scores. Panel (b) presents nonparametric plots of some fictional outcome against the propensity score for treatment and control groups. Standard error bands are not included to make the figure less busy. However, it should be clear that standard error bands must be tighter at values of $P(X)$ near which there are more data. That is, even though it may be possible to calculate a nonparametric regression line for the treatment group at low values of the propensity score, it will be very imprecisely estimated because of the thin data in this region. The main message from Fig. 1.2 is that there are very few comparable observations across treatment and control groups at most propensity scores. Comparability between these two groups typically exists at propensity scores near 0.5, but may not exist for other regions. As a result, it may make sense to limit considerations of treatment effects to treated observations with control observations that have comparable propensity scores.[11]

As discussed by Dehejia and Wahba (2002), identifying "matched" observations in propensity score neighborhoods of treated observations is a fruitful way of identifying a reasonable control group if not many observations have been treated relative to the number of candidate controls. They suggest choosing a propensity score window and only making use of control observations within this window of each treated observation.

[11] While we would have liked to use an example from the urban economics literature to depict graphs such as those in in Fig. 1.2, this depiction has hardly ever been used in our field.

Given that the resulting control group observations are sufficiently close on observables to the treated observations, one can calculate TT as follows:

$$\widehat{\mathrm{TT}} = \frac{1}{N_{T=1}} \sum_{T_i=1} (y_i - \frac{1}{J_i} \sum_{j(i)} y_j).$$

In this expression, $N_{T=1}$ is the total number of treated observations and J_i is the number of "matched" control observations for treated observation i. Those control observations matched to i are indexed by $j(i)$. Treated observations without a match are discarded, with appropriate reinterpretation of TT to apply only to the remaining treated observations.

Standard implementation of the propensity score estimator, which strictly assumes the conditions in (1.11), uses all available data. Given first-step estimation of the propensity score $P(X)$, the following equation can be estimated in a second step by OLS regression:

$$y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 P(X_i) + \alpha_3 T_i (P(X_i) - E[P(X)]) + \varepsilon_i.$$

In this regression, α_1 is the ATE provided that $E[y^1|P(X)]$ and $E(y^0|P(X)]$ are both linear in $P(X)$. A related but more nonparametric procedure that allows for direct recovery of ATE(x) and TT(x) is to estimate a regression such as the following:

$$y_i = b_0 + b_1 T_i + X_i B_2 + T_i (X_i - \overline{X}) B_3 + u_i.$$

Here, $\mathrm{ATE}(x) = \mathrm{TT}(x) = b_1 + (x - \overline{x}) B_3$ and ATE $= b_1$. If there is no treatment effect heterogeneity and ATE(x) = ATE, then this equation reduces to a standard linear regression of y on T and X. Calculation of the propensity score using a linear probability model and no treatment effect heterogeneity reduces the first equation to standard OLS as well. Therefore, we can interpret the OLS as a propensity score matching estimator that incorporates no treatment effect heterogeneity.

Some prominent recent applications of matching estimators have adopted a variant due to Kline (2011) which can be implemented in two steps. First, estimate regressions of the form

$$y_i = c_0 + c_1 T_i + (1 - T_i) X_i C_2 + e_i.$$

Here, X is accounted for in the control group only and not in the treatment group. The purpose is to determine Oaxaca–Blinder-type weights C_2 which serve as inputs into the following treatment effect calculation:

$$\widehat{\mathrm{TT}} = \widehat{c}_1 - \frac{1}{N_{T=1}} \sum_{i=1}^{N} T_i X_i \widehat{C}_2.$$

This procedure compares the average outcome in treated observations with the average outcome in observations with the same distribution of X but that did not receive the treatment. Information from untreated observations in the first step is used to determine

the counterfactual mean for the treated set of observations absent treatment. Kline (2011) shows that this is equivalent to a propensity score reweighting estimator.

The best use of matching and propensity score methods is when there is a good reason to believe that conditional on X, treatment and control groups are similar on unobservables. In recent successful applications, this often involves marrying a matching estimator with a DD-type estimator, which is intended to make the treatment and control groups similar on unobservables. In addition, some observations in the untreated group are typically omitted from the control group in order to make the treatment and control groups as comparable as possible. Such use of propensity score matching estimators is a slightly more sophisticated version of the DD estimator, as they reweight control group observations to look like those in the treatment group on observables.

Busso et al. (2013) use the Oaxaca–Blinder estimator to compare outcomes in census tracts in federal EZs with those in areas that were rejected for inclusion in the program. They find that EZ tracts experienced 12–21% increases in total employment and 8–13% increases in weekly wages, but little change in rents or the composition of the population, though housing values and the percentage of residents with a college degree do increase. They carry out a placebo exercise that compares tracts that are similar on pretreatment observables but not assigned to EZs in EZ counties with the same control group and find no significant effects. Kline and Moretti (2014) use the same estimator in their evaluation of the Tennessee Valley Authority program, for which they trim counties adjacent to the Tennessee Valley Authority region and potential remaining control counties with propensity scores in the lowest 25% and from the control group. Their estimates indicate long-run significant positive effects on manufacturing employment, incomes, and land values and negative effects on agricultural employment.

Gobillon et al. (2012) employ a standard propensity score reweighting estimator to evaluate the effects of the French enterprise zone program, which provides wage subsidies for firms to hire local workers. They find that the program had a small significant effect on the rate at which unemployed workers find a job. McMillen and McDonald (2002) use such an estimator to examine how the type of zoning in Chicago influenced land values immediately after zoning was introduced in 1923. Using the propensity score to match prezoning characteristics between plots zoned for residential versus commercial use, they find that residential plots experienced greater price appreciation. As with the other studies discussed above, the propensity score estimator may be more defensible for this study since the treatment was presumably assigned on the basis of observables and so there is less opportunity for plots of land to sort in or out of treatment on the basis of unobservable characteristics. When individuals are analyzed such sorting concerns are more serious.

In addition to recovering treatment effects in cases of selection on observables, propensity scores can be useful to identify a control group of matched observations for cases in which a specific set of observations has been treated and a very large set of potential

control group observations must be pared down to include just close matches. Alesina et al. (2004) employ such an approach for evaluating the effects of racial heterogeneity on the number of jurisdictions. They identify "treatment" counties as those in northern states which experienced at least a 2 percentage point increase in the black population share during 1910–1920 (during World War I) or 1940–1950 (during World War II). Their challenge is to identify "control" counties that look as similar as possible on observables, and therefore (hopefully) unobservables. To achieve this goal, they first estimate a propensity score for all counties in affected states through a probit regression of treatment status on state fixed effects and various baseline county demographic characteristics and polynomials thereof. As in Dehejia and Wahba (2002), they identify propensity score windows around treated counties in which no significant difference in any observable exists. Then, these treatment and control groups were analyzed both descriptively and in a regression context. The results indicate that greater increases in racial heterogeneity were strong predictors of smaller declines in the number of school districts in the county.

Rather than using propensity score matches to identify a control group that look similar on observables to the treatment group, another strategy that also works with continuous treatments is to think of X as a representative set of potential control variables. Altonji et al. (2005) use this idea to evaluate the magnitude of omitted variables bias in the context of evaluating the causal effects of Catholic schools on high school graduation rates, college attendance, and test scores. Their basic assumption is that including an additional randomly chosen unobservable variable would have the same effect in reducing selection bias as including an additional available observable in X in an OLS regression. Oster (2013) reformulates this assumption as the following proportional selection relationship:

$$\nu\frac{\mathrm{Cov}(T,X\delta)}{\mathrm{Var}(X\delta)}=\frac{\mathrm{Cov}(T,W\rho)}{\mathrm{Var}(W\rho)}.$$

That is, the correlation between observables and the treatment is proportional to the correlation between the unobservables and the treatment.

To implement the resulting estimator, consider the following two regression equations, which can be estimated by OLS, yielding β' and β'' in addition to R^2 of R' and R'', respectively:

$$\begin{aligned}\gamma &= \alpha' + T\beta' + \varepsilon',\\ \gamma &= \alpha'' + T\beta'' + X\delta'' + \varepsilon''.\end{aligned}$$

Having estimated these regressions and capturing their coefficients and R^2, the only remaining required objects are the constant of proportionality ν and the maximum R^2 that would be recovered by estimating the full model, $R_{\max}$. These can be used in the following relationship, which incorporates the bias adjustment to the OLS regression from the full model:

$$\beta \xrightarrow{p} \beta'' - \nu \frac{(\beta' - \beta'')(R_{\max} - R'')}{(R'' - R')}.$$

Of course, the main difficulty is that ν and $R_{\max}$ are unknown. But one can get an idea of how large the bias could be by determining what ν would need to be for $\beta = 0$ given $R_{\max} = 1$. Standard errors need to be bootstrapped to conduct inference on the resulting bias-corrected coefficient.

The key obstacle to the use of matching, DD, and fixed effects estimators is the lack of any source of randomization. In some sense, all of these estimators end up in an environment in which we must assume that T is allocated in a way that is as good as random conditional on the other observed elements of the estimation equation. The following section's exploration of IV estimators instead focuses on environments in which there is some randomization in T, which is usually implicit.

1.5. IV ESTIMATORS

IV estimators are used to recover consistently estimated coefficients on treatment variables of interest when treatments are endogenous. One way of conceptualizing such an endogeneity problem is that a treatment variable is generated by a second linear equation which includes some unobservables that are correlated with unobservables which appear in the main estimation equation of interest. This makes the treatment T be correlated with the U part of the error term in the primary estimation equation, rendering the OLS estimate of the coefficient on the treatment biased and inconsistent. In the language of structural systems, there needs to be an "exclusion restriction" in which at least one observed variable must be excluded from one equation in order to identify coefficients of both equations without making ad hoc distributional assumptions. In the language of single-equation linear regression, there needs to be an "instrument" which isolates variation in T that is not correlated with any part of the error term in the main estimating equation. We sometimes label such variation "pseudorandom" because the role of the instrument is essentially to pick out random variation in T.

Consideration of how to estimate the classic Roy (1951) model by Gronau (1974) and Heckman (1979) is informative about the more structural background of the IV estimator. In this model, there is a binary treatment T into which individuals may self-select because it is presumably valuable for them. This self-selection generates a correlation between T and the error term in a linear regression of some outcome of interest on T and control variables X because of sorting on unobservables into the treatment. In particular, the underlying data-generating process is assumed to be

$$y^0 = X\delta_0 + U_0; \;\; y^1 = X\delta_1 + U_1.$$

Heckman (1979) shows that if U_0 and U_1 are jointly normal, one can identify δ_1 and evidence of selection into treatment. The key insight is that the choice of whether to accept

treatment can be recovered explicitly using the fact that only those for whom $y^1 > y^0$ select into treatment. Operationally, one way of estimating the model is by estimating the model as a "Heckman two-step." First, predict the probability of treatment as a function of X using a probit regression. Second, estimate the equation

$$y^1 = X\delta_1 + \rho\sigma_u\lambda(X\gamma) + \varepsilon.$$

In this equation, $\lambda(\cdot)$ is the inverse Mills ratio constructed from the first step, which controls for selection into treatment. Because y^0 was never observed in the original application, the standard treatment does not have a second step equation for y^0, though one could be constructed using analogous logic. The sign and magnitude of estimated ρ indicate the nature of selection into treatment on unobservables. One important insight of this work is thus that one can treat nonrandom selection into treatment as an omitted variables problem. The difficulty is that if the errors are not truly jointly normal, the model is misspecified and coefficients in the second step equation are inconsistenly estimated unless an exclusion restriction is imposed.

Altonji et al. (2005) also consider a two-equation structural system in their exploration of evaluating the effects of attending Catholic schools on college attendance. They consider a bivariate probit model in which a set of demographic characteristics predict both Catholic school attendance and college attendance, such that Catholic school attendance is an explicitly endogenous treatment variable. They demonstrate how the estimate of the coefficient on T (Catholic school attendance) depends crucially on the magnitude of the correlation between the errors in the two equations. Higher correlations between the error terms mean that there are more similar unobservables driving both Catholic school attendance and success in school. As a consequence, the causal effect on Catholic school attendance declines because this variable simply reflects more positively selected students as the error correlation increases.[12] In the context of a data-generating process such as (1.4), one way to make progress in breaking a potential correlation between T and U, which renders OLS or probit estimates inconsistent, is to find variables that predict T but are not correlated with U. These are instruments, or exclusion restrictions.

In summary, the IV estimator is used to break a potential correlation between T and U. This correlation could exist because individuals with high values of U are sorting into the treatment at higher rates than others, as in the classic two-equation structural selection model in which T is "endogenous" because it is generated by a second equation. Or this correlation could exist because, regardless of where T comes from, there are variables correlated with T for which the researcher cannot control that end up in U as a result.

[12] Neal (1997) considers a similar bivariate probit setup to address the same questions except that he excludes religious affiliation and local Catholic population density from the graduation equation. These exclusion restrictions allow for recovery of estimates of the covariance of the errors between the two equations and the coefficient on Catholic schooling in the estimation equation of primary interest.

This is an omitted variables problem. These two ways of thinking about why $E(TU) \neq 0$ have distinct intellectual histories but many of the same implications.

1.5.1 Foundations

To be mathematically precise, we can think of IV estimators as those that recover β in the following system of equations:

$$y_i = T_i\beta + X_i\delta + \varepsilon_i, \tag{1.12}$$

$$T_i = Z_i^1\zeta_1 + X\zeta_2 + \omega_i. \tag{1.13}$$

In the second equation, Z^1 is the set of excluded instruments, of which there must be at least one per treatment variable for this econometric model to be identified. These additional Z^1 variables are "excluded" from the first equation. In the first equation, recall that $\varepsilon_i = U_i + e_i$ from (1.4). Denote the set of exogenous variables as $Z = [Z^1 X]$. IV estimators recover consistent estimates of β if $E(Z\varepsilon) = 0$ and the coefficients on the excluded instruments ζ_1 in (1.13) are sufficiently different from 0 in a statistical sense. We sometimes use the "reduced form" of this two-equation system, which is as follows:

$$y_i = Z_i^1\phi_1 + X_i\phi_2 + \psi_i.$$

If there is just one excluded instrument per endogenous variable, one simple way to estimate β is through indirect least squares (ILS): $\widehat{\beta}_{\text{ILS}} = \frac{\widehat{\phi}_{\text{1OLS}}}{\widehat{\zeta}_{\text{1OLS}}}$. This is an intuitive object which shows how the first-stage coefficient rescales the reduced form effect of the instrument on the outcome.

Another simple intuitive way to estimate β is by substituting (1.13) into (1.4) and then explicitly including a proxy for ω_i in the estimation of the resulting (1.14):

$$y_i = T_i\beta + X_i\delta + \widehat{\omega}_i\zeta + e_i. \tag{1.14}$$

This proxy acts as a "control function" for unobservables correlated with T_i. In the linear case above, β can be properly estimated by using $\widehat{\omega}_i$ consistently recovered as residuals from OLS estimation of the first-stage (1.13). This method is closely related to the two-stage least squares (2SLS) estimator in which $\widehat{T}_i$ is predicted from the first stage and inserted in place of T_i in (1.12), which can then be estimated by OLS to recover $\widehat{\beta}_{\text{2SLS}}$.[13] However, as discussed in Imbens and Wooldridge (2007), the control function approach sometimes provides additional flexibility when dealing with nonlinear models. Moreover, the coefficient ζ has a useful economic interpretation. ω_i is positive for those observations which were treated more than expected as predicted by Z^1 and X. One could thus interpret those agents as having higher than predicted returns from receiving treatment. Therefore, the sign of ζ indicates whether the type of agent who had a higher

[13] For 2SLS estimation, it is important that the standard errors use estimates of ε_i calculated using the actual rather than the predicted T_i.

return from the treatment had better or worse outcomes y than the types of agents who had lower treatment returns. That is, ζ tells us about the nature of selection into treatment, much like the coefficient on the inverse Mills ratio does in Heckman (1979), as is fleshed out further in the development by Heckman and Honoré (1990) of the empirical content of Roy's model (Roy, 1951).

In addition to ILS, 2SLS, and control function methods, GMM, which makes use of the moment condition $E[Z^1\varepsilon] = 0$, and limited information maximum likelihood are options for estimating β in the two-equation econometric model specified in (1.12) and (1.13). All of the various estimators of β in (1.12) suffer from weak small sample properties, though limited information maximum likelihood has been found to be most robust in small samples. All of these estimators are identical if the model is just identified, meaning that there is the same number of excluded variables as there are endogenous variables. Recent work has found that 2SLS can be more robust in some instances with many instruments if they predict not only T but also an element of X (Kolesar et al., 2013).

Most important for successful implementation of IV is the choice of good excluded instruments. One fruitful way of conceptualizing an instrument is as a source of random variation in T conditional on X. That is, a good instrument generates variation in T conditional on X that is not correlated with any unobservables in U. However, each element of X must also be exogenous. Therefore, the best instruments are those that generate truly random variation in T and therefore require no conditioning on X in the first equation. With such ideal instruments, which typically are only found with explicit randomization, the prudent researcher can avoid having to control for any elements of X and facing the associated danger of introducing a potential endogeneity problems. We discuss using IV estimators as a means to make use of explicit randomization in the context of RD in the following section.

The more typical situation is that a researcher is concerned about the endogeneity of some treatment T and there is no explicit randomization available. The following is one strategy for selecting good candidate instruments: Consider all of the possible sources of variation in T. From this list, select the ones that are least likely to be correlated with variables that directly predict y or are correlated only with observables that predict y that are very likely exogenous. Coming up with this list typically requires both creativity and a detailed investigation into the process by which the treatment was assigned. There is no direct test for instrument exogeneity, only a set of exogeneity arguments that are unique to each setting, though there are various standard auxiliary tests, some of which are suggested below in the context of examples from the literature. The next step is to estimate the first stage, (1.13), and to evaluate whether the instruments are sufficiently strong predictors of T. If they are not, the researcher has to keep looking. If multiple strong instruments are identified, special care is needed, as is also discussed below.

If the partial F statistic from the test of whether coefficients on excluded instruments are each significantly different from 0 is above about 9, then the instruments are strong enough predictors of T such that the estimated standard errors on β can be used.[14] Otherwise, standard errors on β must be adjusted upward to account for a "weak instrument" problem. Stock and Yogo (2005) provide standard critical values for F tests for evaluating instrument strength. When implementing the primary specification of an IV estimator, one should control only for those predictors of γ that may be correlated with the instrument so as to avoid controlling for endogenous variables.

While the exposition thus far assumes a common coefficient β, in general we expect there to be heterogeneous coefficients on T of $B(X, U)$. Crucial to understanding IV estimates is to recognize that IV recovers a LATE, which is the average effect of the treatment for the subpopulation whose behavior was influenced by the excluded instrument, conditional on X (Imbens and Angrist, 1994). It typically requires further investigation to gather information about the particular LATE that is recovered from any given instrument. Continuous instruments and treatments in particular usually require some detective work to determine for whom the treatment effect being estimated by IV applies. With multiple instruments, it becomes even more complicated. Indeed, Heckman et al. (2006) lament that with many instruments it is often virtually impossible to determine which combination of MTEs is being estimated by IV.

Because of the fact that IV recovers a LATE, and that in typical urban economics applications it is difficult enough to find one valid instrument let alone many, it is prudent to stick to using only one excluded instrument at a time in most settings, with additional candidate instruments possibly used for robustness. The only reason to use multiple instruments at once is if one instrument by itself is too weak. Though it is possible to test for stability in β when switching between different instruments as a test of instrument validity, this process crucially assumes that the data are generated by a process with a constant coefficient. If instead there are heterogeneous coefficients, it may well be the case that multiple instruments generate different legitimate treatment effect estimates, all of which are different LATEs.

1.5.2 Examples of IV in urban economics

In the urban and regional economics literature, the IV empirical strategy has been most commonly used when the unit of observation is aggregated to the local labor market level. That is, the data-generating processes that have best lent themselves to IV estimation are either fully conceptualized at the aggregate level, as in (1.6), or are agent based but involve a treatment that operates at some aggregate geographic level, as in (1.5). Here we review examples of how IV has been used to successfully isolate exogenous components of local labor demand and labor supply shocks, construction of infrastructure, the

[14] This is equivalent to evaluating if the t statistic is above 3 if there is just one excluded instrument.

implementation of local economic development policies, and the prevalence of various drivers of local agglomeration spillovers.

The classic use of IV in economics is to isolate exogenous supply or demand shifters in some particular market. Since supply and demand functions are fundamentally theoretical constructs, use of IV to isolate demand or supply shocks is probably most effective when an economic model is incorporated into the analysis in some way in order to organize thoughts about the most important forces buttressing equilibrium prices and quantities. Given the centrality of the demand–supply paradigm in economics, use of IV to isolate exogenous variation in demand and supply has a strong tradition. For example, Angrist et al. (2000) use weather variables as a source of exogenous variation in supply shifts to recover demand system parameters using the well-known Fulton Street Fish Market data (Graddy, 1995).

Following in this tradition, one of the commonest uses of IV estimation in the urban and regional economics literature is to isolate sources of exogenous variation in local labor demand. The commonest instruments for doing so are attributed to Bartik (1991) and Blanchard and Katz (1992). The idea is to isolate shifts in local labor demand that come only from national shocks in each sector of the economy, thereby purging potentially endogenous local demand shocks driving variation in employment or wages. While this type of instrument has been used to help recover parameters of local labor supply functions, it has more often been used to isolate exogenous variation in metropolitan area wages or employment levels.

There are two ways that "Bartik" instruments are most commonly constructed. A quantity version of the instrument is constructed by fixing each local labor market's industry composition of employment at some base year and calculating the employment growth that would have occurred in each market had the industry composition not changed but employment in each industry had grown at the national rate for that industry. The price version of the instrument instead calculates the wage growth that would have occurred in each market had wages in each industry grown at the national rate for that industry, again holding the employment composition in each local labor market fixed to a base year. In order to allay potential concerns of a mechanical relationship between base year industry composition and unobservables driving an outcome of interest, researchers typically take industry composition from a year that predates measurements of any other variables used for estimation.[15]

A host of papers make use of such instruments for identification. Notowidigdo (2013) uses exogenous variation from Bartik instruments to demonstrate that positive local labor

[15] To allay the potential concern that any particular local labor market influences national employment or wage growth, many studies exclude the own local labor market or state in the calculation of national growth rates by sector. This means that this growth component of the instrument is slightly different for each observation.

demand shocks increase the population more than negative demand shocks reduce it, and that this asymmetry is more pronounced for less skilled workers. However, he finds that housing prices, wages, and rents do not exhibit the same asymmetric responses. Through the structure of a Roback (1982) style spatial equilibrium model, these results are interpreted as indicating low mobility costs for everyone and a concave local housing supply function. Leveraging the same exogenous variation in local labor demand for identification, GMM estimates of the full model reveal that less skilled workers are more highly compensated through various transfers for negative local labor demand shocks than highly skilled workers, which accounts for the different mobility rates of these two groups. In a precursor to Notowidigdo (2013), Bound and Holzer (2000) examine the general equilibrium population responses by skill to exogenous local labor demand shocks. Through GMM estimation of a spatial equilibrium model, Diamond (2013) uses the identifying variation available from Bartik instruments to recover how local labor demand shocks lead to knock-on shifts in local skill composition and skill-specific amenities. Boustan et al. (2013) use Bartik instruments to help demonstrate that jurisdictions with greater increases in income inequality collected more local government revenues and had higher expenditures. Luttmer (2005) uses Bartik instruments in a reduced form specification to control for changes in average area incomes in showing that people whose incomes fall behind those of their neighbors are less happy, even if everyone's incomes are increasing. Gould et al. (2002) use Bartik shocks as an instrument for income in examining the causal effects of income on local crime rates.

In an important study, Saiz (2010) uses Bartik instruments to isolate exogenous local housing demand shocks interacted with a measure of land unavailable for development and an index of housing market regulation to recover an estimate of the housing supply elasticity for each US metropolitan area. He estimates inverse housing supply regression equations of the form

$$\Delta \ln P_k = \alpha_0 + \alpha_1 \Delta \ln Q_k + \alpha_2 \text{unavailable\_land}_k \Delta \ln Q_k + \alpha_3 \text{WRI}_k \Delta \ln Q_k + u_k,$$

in which k indexes metropolitan area, P denotes housing price, Q denotes housing quantity, and WRI is an index of local housing market regulation. Differences are taken for the 1970–2000 period. Bartik quantity instruments provide exogenous variation in all terms which include $\Delta \ln Q_j$.[16] Housing supply elasticity estimates from this study have been widely used. In the work of Beaudry et al. (2014), such estimates interact with Bartik instruments to form a series of instruments in the estimation of a spatial equilibrium model which incorporates unemployment and wage bargaining frictions. The works

[16] Saiz (2010) also makes use of hours of January sun and immigration inflows as additional sources of exogenous variation in $\Delta \ln Q_k$ and the prevalence of evangelical Christians as a source of exogenous variation in WRI_k.

of Mian and Sufi (2009) and Chaney et al. (2012) are two prominent examples from the finance literature that use these Saiz (2010) housing elasticity measures.

The main source of identifying variation in Bartik instruments comes from differing base year industry compositions across local labor markets. Therefore, validity of these instruments relies on the assertion that neither industry composition nor unobserved variables correlated with it directly predict the outcome of interest conditional on controls. As with any IV, the credibility of this identification assumption depends on the context in which the IV is being applied. Generically, one may be concerned that base year industrial composition may be correlated with fundamentals related to trends in labor supply. For example, it may be the case that manufacturing-intensive cities have declined not only because the demand for skill has declined more in these locations, but also because they have deteriorated more in relative amenity values with the increasing blight and decay generated by obsolete manufacturing facilities. That is, negative labor supply shifts may be correlated with negative labor demand shifts. Indeed, when Bartik instruments are implemented using one-digit industry classifications, as is often done, the initial manufacturing share tends to drive a lot of the variation in the instrument. In these cases, one can conceptualize this IV as generating a comparison between manufacturing-heavy and nonmanufacturing-heavy local labor markets. Finally, depending on how it is implemented, the Bartik instrument may isolate variation in different components of labor demand depending on the skill composition of the workforce in the industry mix in the base year. For example, two local labor markets may be predicted to have similar employment growth because of the prevalence of retail and wholesale trade in one of them and the prevalence of business services in the other. In fact, the latter likely would have experienced a much greater outward shift in labor demand if measured in efficiency units terms, which may be the more appropriate quantity measure depending on the application.

Another common use of IV is to isolate exogenous variation in local labor supply. Following Card (2001), one common strategy for doing so is to make use of immigration shocks. As is discussed in more detail in Chapter 10 by Lewis and Peri, this variation has been used extensively in the immigration literature as an instrument for the flow of immigrants to domestic local labor markets. This instrument is typically constructed by multiplying the fraction of immigrants to the United States from various regions of origin worldwide that reside in each metropolitan area in a base year with the total flow of immigrants into the United States from each region over some subsequent time period, and then summing over all regions of origin.[17] As in Lewis (2011), an analogous exercise can be carried out by observed skill to generate variation across local labor markets in the relative supply of skill, though this exercise has a stronger first stage for less skilled groups.

[17] As with Bartik instruments, some studies leave out the own local labor market or state when calculating national immigrant flows from each world region of origin.

Boustan (2010) uses a similar historical pathways instrument for the size of the African American population in northern metropolitan areas after Word War II.

IV has also been widely used to isolate exogenous variation in infrastructure treatments. The commonest types of instruments used for transportation infrastructure variables are historical plans and networks. For example, Baum-Snow (2007) estimates the impacts of the construction of radial limited access highways serving central cities in US metropolitan areas on population decentralization. He finds that each radial highway emanating from a central city decentralized about 9% of the central city's population to the suburbs. He uses the highways laid out in a 1947 federal plan for a national highway system as a source of exogenous variation. The validity of this empirical strategy rests on the fact that the 1947 highway plan delineated routes that were chosen because they would facilitate military transportation and intercity trade. Local travel demand was not considered in making this highway plan. The 90% federal funding commitment for highway construction ensured that virtually all planned highways were built, with considerable additions to the interstate system to serve local travel demand. The primary analysis in Baum-Snow (2007) involves estimating 1950–1990 differenced regressions of the central city population on radial highways, controlling for metropolitan area population, in order to subsume the full time period during which the interstate system was constructed. Central to successful identification is to control for variables that may be correlated with planned highways and drive decentralization. Controls for central city size, 1950 metropolitan area population, and industrial structure in various specifications serve this purpose, though only the central city size control matters. Baum-Snow (2007) also reports estimates from a DD-type specification using data from decades between 1950 and 1990 and including metropolitan area and year fixed effects. For this empirical strategy, 1990 radial highways interacted with the fraction of federally funded mileage completed by the year of the observation enters as the highways instrument. Michaels (2008) uses a similar 1944 plan as an instrument for highways serving rural counties in his investigation of how better market integration changed the demand for skill. Though they turn out to be insufficiently strong, he also tries using the existence of nearby cities on the north–south or east–west axes relative to each county in question as instruments, since the interstate system is oriented in this way.

Duranton and Turner (2011, 2012) and Duranton et al. (2014) also use the 1947 plan as an instrument for highways, but supplement it with 1898 railroads and an index of continental exploration routes during the 1528–1850 period. These papers evaluate the effects of highways on the amount of intracity travel, urban growth, and the composition of interregional trade, respectively. Baum-Snow et al. (2014) similarly use aspects of historical urban road and railroad networks as an instrument for their modern counterparts in their investigation of changes in urban form in post-1990 Chinese cities. The idea of using historical infrastructure as instruments is that though such infrastructure is obsolete today, its rights of way are likely to be preserved, allowing for lower cost

modern construction. Dinkelman (2011) uses land gradient as an instrument for the prevalence of rural electrification in South Africa. She finds that much like new highways, electrification led to employment growth. As discussed further in Chapter 20 by Redding and Turner in this handbook, how to distinguish between the effects of infrastructure on growth versus redistribution is still very much an open question. Whatever their interpretation, however, well identified IV regressions can recover some causal effects of infrastructure.

Hoxby (2000) is one of the earlier users of IV estimation in the local public finance literature. This paper attempts to recover the effects of public school competition, as measured by the number of public school districts in metropolitan areas, on student test scores. To account for the potential endogeneity of the number of school districts, Hoxby uses the prevalence of rivers and streams in the metropolitan area as an instrument. The idea is that metropolitan areas with more rivers and streams had more school districts because historically it was difficult for students to cross rivers to get to school, but these natural features do not directly influence levels or accumulation of human capital today. Potentially crucial for identification, of course, is to control for factors that might be correlated with rivers and streams but predict test scores. For example, metropolitan areas with more rivers and streams may be more likely to be located in more productive parts of the country such as the Northeast and Midwest, so controlling for parents' education and outcomes may be important.[18] More recently, Serrato et al. (2014) have used city population revisions because of decennial censuses to isolate exogenous variation in federal transfers to recover that the local income multiplier is 1.57 per federal dollar and the fiscal cost per additional job is $30,000 per year.

One additional common type of instrument uses variation in political power and incentives. For example, Levitt (1997) uses mayoral election cycles as an instrument for the number of police deployed in cities in a given month in his investigation of the effects of police on crime. The idea is that mayors up for reelection expand the police force during this time in an attempt to reduce crime. Consistent with the intuition of ILS, this study essentially compares crime rates during election cycles with those at other times, scaling by the difference in the numbers of police in these two environments. Of course, isolating a causal effect of police requires controlling for other policy changes implemented during election cycles.[19] Hanson (2009) and Hanson and Rohlin (2011) use congressional representation on the Ways and Means Committee as an instrument for selection of proposed EZs for federal funding.

We hope that this incomplete survey of the use of IV in the urban and regional literature has shown that credible implementation of IV is far from a mechanical process. As with any empirical strategy, the successful use of IV requires careful thought about the

[18] Rothstein (2007) provides additional analysis of the question using additional data.
[19] See McCrary (2002) for a reanalysis of the same data set.

identifying variation at play. A convincing logical argument must be made for exogeneity of each instrument conditional on exogenous control variables, or equivalently that remaining variation in the instrument is uncorrelated with unobservables that drive the outcome of interest. In addition, ideally some idea should be given of which LATEs IV estimates using each instrument return.

One can use the mechanics of the IV estimator to recover TT in environments in which the treatment is explicitly randomized, as in the MTO studies discussed in Section 1.2.4. Katz et al. (2001) walk through this process in detail. In the MTO context, assign $Z = 1$ to households in the Section 8 treatment group and $Z = 0$ to households in the control group. $D = 1$ if a household moves out of public housing with a Section 8 voucher and $D = 0$ if the household does not. One can think of Z as being a valid instrument for D. Households receiving a voucher choose whether or not to use it, making D endogenous. Recall from Section 1.2.2 the definition of LATE, which in this binary treatment context becomes $\text{LATE} \equiv \frac{E[y|Z=1]-E[y|Z=0]}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}$. The numerator is the coefficient on Z in a "reduced form" regression of y on Z. The denominator is the coefficient on Z in a "first-stage" regression of D on Z. That is, we see in this simple context how LATE is a restatement of the ILS IV estimator. Additionally, recall from Section 1.2.2 the definition $\text{TT} \equiv E(y^1 - y^0|D = 1) = \frac{E[y|Z=1]-E[y|Z=0]}{\Pr(D=1|Z=1)}$. Therefore, TT = LATE if $\Pr(D = 1|Z = 0) = 0$, or no members of the control group use a Section 8 voucher to move out of public housing.

It is also typical to use the IV estimator to implement the RD empirical strategy. The following section details how this is done.

1.6. REGRESSION DISCONTINUITY

Use of the RD research design in economics has dramatically increased in the past decade, as attested in recent reviews by Lee and Lemieux (2010) and Imbens and Lemieux (2008). Our interpretation of RD estimates has also changed in this period. Initially thought of as another method to deal with selection on observables, RD was subsequently motivated as a type of local IV, and then finally defined as a creative way of implementing random assignment in a nonexperimental setting. In this section, we discuss the different interpretations of the RD framework, the relevant details on how to implement the approach, and some of its notable uses in urban and regional economics. Even though RD designs have been quite rare in urban economics papers until recently,[20] the approach shows much promise for future research, and we expect its use in urban economics to grow over time in the same way experienced by other applied economics fields. This section can be thought of as a first gateway to the approach; more detailed discussions are presented in Lee and Lemieux (2010) and Imbens and Lemieux (2008).

[20] For example, zero papers used the RD design as recently as 2010 in the *Journal of Urban Economics*.

1.6.1 Basic framework and interpretation

There are two main prerequisites for RD to apply as a potential empirical strategy. First, the researcher needs to know the selection into treatment rule, and there should be a discontinuity in how the treatment is assigned. For example, US cities often promote referenda that ask local citizens if they would approve raising extra funds through bond issuances that will be used to invest in local infrastructure. The selection rule in this case is based on the vote share needed to approve the bond issue, let us say two-thirds of the local vote. The discontinuity in treatment is obvious: cities whose referenda got less than two-thirds of the votes will not raise the funds, while cities whose referenda achieved the two-thirds mark will be able to issue the bonds and subsequently invest the proceeds in local infrastructure. The second prerequisite is that agents are not able to sort across the selection threshold. Such "selection" would by definition invalidate the ability to compare similar individuals in the control and treatment groups on either side of the threshold. In the referenda example, this no endogenous sorting condition means that cities are not able to manipulate the referendum in order to influence their ability to get one additional vote to reach the two-thirds threshold. At the end of the section we will discuss how researchers can potentially deal with violations of this condition, such as in boundary-type applications in which sorting is expected to happen over time.

If both conditions above are met, the RD estimate will provide a comparison of individuals in treatment and control groups that were "matched" on a single index—that is, the selection rule. This single index is usually referred to as the running variable or the assignment variable.

To formalize those concepts, define γ_i as the outcome of interest and T_i as the relevant binary treatment status, and assume $\beta_i = \beta$ and X_i is a vector of covariates:

$$\gamma_i = \alpha + T_i\beta + X_i\delta + U_i + e_i, \tag{1.15}$$

where $T_i = 1(Z_i \geq z_0)$. Z_i is the single index for selection into treatment, and z_0 is the discontinuity threshold. Individuals with $Z_i \geq z_0$ are assigned to the treatment group, while the remaining individuals are assigned to the control group. Such a setup is usually referred to as the "sharp" RD design because there is no ambiguity about treatment status given the known and deterministic selection rule. In this setting, the ATE of T_i on γ_i around the threshold is

$$\begin{aligned} E[\gamma_i|Z_i = z_0 + \Delta] - E[\gamma_i|Z_i = z_0 - \Delta] = \beta + \{E[X_i\delta|Z_i = z_0 + \Delta] - E[X_i\delta|Z_i = z_0 - \Delta]\} \\ + \{E[U_i + e_i|Z_i = z_0 + \Delta] - E[U_i + e_i|Z_i = z_0 - \Delta]\}. \end{aligned}$$

Note that this ATE applies only to the agents with characteristics of those near the threshold. Two key assumptions allow for the identification of ATE. First, continuity of the joint distribution of X_i and Z_i. This assumption makes the term $\{E[X_i\delta|Z_i = z_0 + \Delta] - E[X_i\delta|Z_i = z_0 - \Delta]\}$ in the equation above negligible, and guarantees that both the control group and the treatment group will have similar observed characteristics

around the discontinuity threshold. This assumption is easily tested in the data, and it is one of the reasons for interpreting RD as a selection on observables type of framework. The second assumption is that the joint distribution of the unobserved component ($U_i + e_i$) and Z_i is continuous, which makes the term $\{E[U_i + e_i|Z_i = z_0 + \Delta] - E[U_i + e_i|Z_i = z_0 - \Delta]\}$ also negligible. This assumption can never be tested. This type of sharp RD is analogous to random assignment in the sense that, around the threshold, the assignment of individuals to control and treatment groups is exogenous given the two assumptions above.

In some circumstances, however, the selection rule may not be deterministic. For example, even when local citizens approve a bond issue, overall market conditions may prevent the municipality from raising the funds. Or US cities in which a bond referendum failed today may try to pass other bond measures in the near future. Those events may turn the selection rule into a probabilistic equation, leading to the so-called fuzzy RD design. Formally, the treatment status T_i can be rewritten as

$$T_i = \theta_0 + \theta_1 G_i + u_i,$$

where $G_i = 1(Z_i \geq z_0)$, and u_i corresponds to the other unobserved components that determine treatment status. Plugging in the new equations for T_i and G_i in the outcome equation generates

$$y_i = \alpha + \beta\theta_0 + G_i\beta\theta_1 + u_i\beta + X_i\delta + U_i + e_i,$$

and the new treatment effect around the threshold becomes

$$\begin{aligned} E[y_i|Z_i = z_0 + \Delta] - E[y_i|Z_i = z_0 - \Delta] = \beta\theta_1 + \beta\{E[u_i|Z_i = z_0 + \Delta] - E[u_i|Z_i = z_0 - \Delta]\} \\ + \{E[X_i\delta|Z_i = z_0 + \Delta] - E[X_i\delta|Z_i = z_0 - \Delta]\} + \{E[U_i + e_i|Z_i = z_0 + \Delta] \\ - E[U_i + e_i|Z_i = z_0 - \Delta]\}. \end{aligned}$$

In order to estimate the parameter β we first need to back out the parameter θ_1, which establishes the relationship between G_i and T_i,

$$E[T_i|Z_i = z_0 + \Delta] - E[T_i|Z_i = z_0 - \Delta] = \theta_1 + \{E[u_i|Z_i = z_0 + \Delta] - E[u_i|Z_i = z_0 - \Delta]\},$$

and a LATE can be recovered using the ratio of the reduced form impact of the single index Z_i on outcome y_i, and of the first stage described above:

$$\beta = \frac{E[y_i|Z_i = z_0 + \Delta] - E[y_i|Z_i = z_0 - \Delta]}{E[T_i|Z_i = z_0 + \Delta] - E[T_i|Z_i = z_0 - \Delta]}. \tag{1.16}$$

This expression closely resembles the definition of LATE in (1.3). The reason the fuzzy RD design can be thought of as delivering a LATE is that the treatment effect is recovered only for some agents. If the set of agents induced into treatment by having an assignment variable value that is beyond the critical threshold is random, then this coincides with the same ATE estimated in the sharp RD environment. However, if the fuzzy RD occurs

because a group of agents do not comply with the "treatment" of being beyond the threshold, presumably because they differ from compliers on some observables or unobservables, then the fuzzy RD design allows the researcher to recover only a LATE, which can also be thought of as a particular version of treatment on the treated (TT).

The validity of the fuzzy RD design relies on the following assumptions: (1) there is random assignment of control and treatment groups around the threshold; (2) there is a strong first stage, allowing the estimation of θ_1; (3) there is an exclusion restriction, so that the term $\{E[u_i|Z_i = z_0 + \Delta] - E[u_i|Z_i = z_0 - \Delta]\}$ also becomes negligible.[21] This setup is very similar to the IV approach covered in the previous section, and the fuzzy RD is sometimes interpreted as a local IV.

As emphasized in DiNardo and Lee (2011), the simplistic IV interpretation misses the most important characteristic of the RD design: the random assignment of treatment and control groups. Even though the fuzzy design resembles the mechanics of an IV approach, the key characteristic of the design is the ability of mimicking random assignment in a nonexperimental setting. In fact, the fuzzy RD design could be more properly designated as a locally randomized IV.

An important issue in RD designs is external validity, as one potential interpretation of the approach is that "it only estimates treatment effects for those individuals close to the threshold." DiNardo and Lee (2011) clarify the interpretation of those estimates by using the idea that individuals do not get to choose where they locate with respect to the RD threshold. If that is the case, RD estimates can be viewed as a weighted average effect, where the weights are proportional to the ex ante likelihood that the value of the individual's assignment variable would lie in a neighborhood of the threshold.

Independent of using a sharp or fuzzy design, the RD approach provides a method of approximating the empirical estimation to a randomization setting. As discussed in earlier sections, randomization is the Holy Grail of empirical work, and any method that allows nonexperimental approaches to replicate the characteristics of a experimental design is bound to be welcomed by researchers.

1.6.2 Implementation

The popularity of the RD approach is explained not only by its relationship with randomized experiments, but also because of the transparency of the framework. RD estimation can be transparently shown in a graphical format. The standard RD figure plots conditional or unconditional means of the treatment and/or outcome of interest by bins of the assignment variable. Following the bond issue example, Cellini et al. (2010) show average expenditures and average capital outlays per pupil by the vote share in a bond referendum (see Fig. 1.3). This simple figure first shows that a treatment

[21] This approach also relies on a monotonicity assumption, similar to the one used to cleanly interpret LATE in an IV setting. It means that as one moves across the assignment variable threshold, the probability of treatment for every combination of observables X and unobservables U increases.

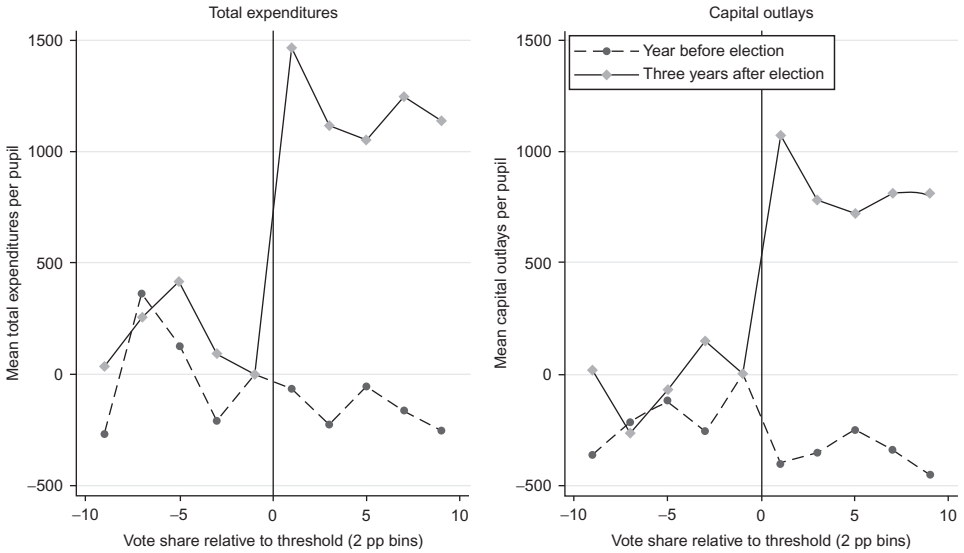

Figure 1.3 Total spending and capital outlays per Pupil, by vote share, 1 year before and 3 years after Election (Cellini et al., 2010). Graph shows average total expenditures (left panel) and capital outlays (right panel) per pupil, by the vote share in the focal bond election. Focal elections are grouped into bins 2 percentage points wide: measures that passed by between 0.001% and 2% are assigned to the 1 bin; those that failed by similar margins are assigned to the −1 bin. Averages are conditional on year fixed effects and the −1 bin is normalized to zero.

exists: total expenditures and capital outlays increased for school districts that had vote shares above the threshold, and only in the 3 years after the bond measure was approved. It also tests the sharpness of the research design: school districts whose referenda had vote shares below the threshold had similar expenditures and capital outlays in the year before and in the 3 years after the referendum. The combination of these results for treatment and control groups is a clear discontinuity of a given magnitude around the threshold.

A similar graphical approach should be used to test the validity of the research design. All relevant covariates should be displayed in unconditional plots by bins of the assignment variable, and the statistical test of a discontinuity for each covariate should be presented. This is the main test of the assumption that control and treatment groups have balanced characteristics around the discontinuity threshold. An additional test of sorting around the discontinuity can be performed by plotting the total number of observations in each bin against the running variable. That will test whether there is a disproportional number of individuals on each side of the threshold, which could potentially indicate the ability of individuals to manipulate their treatment status and therefore invalidate the research design—see McCrary (2008). In practice though, such sorting would usually show up as differences in other covariates as well. Finally, other common robustness tests, including testing for a discontinuity in predetermined covariates (in the case of a

treatment that has a time component), testing if the outcome variable presents a discontinuity at a fake discontinuity threshold, meaning that a discontinuity only happens at the true threshold, and testing whether other unrelated outcomes, have a similarly discontinuous relationship with the running variable, which would indicate that the treatment may not be the only mechanism impacting outcomes.

Many RD applications also plot parametric or nonparametric estimates of the ATE along the unconditional means of the assignment variable. When a parametric estimate is used, the graphical analysis can also help with the choice of the functional form for the RD single index. As mentioned earlier, the assignment variable Z_i can be interpreted as a single index of the sources of observed bias in the relationship between outcome and treatment status. If the single index is smooth at the RD threshold z_0, that indicates that any discontinuity in y_i would be due to T_i. In the easiest case, there is no correlation between the outcome y_i conditional on treatment status and the running variable Z_i, and a simple regression such as $y_i = \alpha_0 + T_i\beta + \epsilon_i$ would generate proper estimates of the ATE. A commoner situation is where y_i is also some function of Z_i, with similar slopes on either side of the threshold. A more general empirical model that allows for different functions of Z_i above and below z_0 which is commonly used to implement sharp RD estimation is

$$y_i = \alpha_0 + T_i\alpha_1 + f_1(z_0 - Z_i)1(Z_i < z_0) + f_2(Z_i - z_0)1(Z_i \geq z_0) + X_i\delta + \epsilon_i, \qquad (1.17)$$

where $T_i = 1(Z_i \geq z_0)$ in the sharp RD case. Many researchers implement $f_1(\cdot)$ and $f_2(\cdot)$ as cubic or quadratic polynomials with estimated coefficients, imposing the constraints that $f_1(0) = f_2(0) = 0$ by excluding intercept terms from the polynomials. The inclusion of α_0 in (1.17) allows the level of y_0 at $Z = z_0 - \Delta$ to be nonzero. This equation can be estimated by OLS. The underlying idea, again, is to compare treatment and control units near the threshold z_0. The role of the $f_1(\cdot)$ and $f_2(\cdot)$ control functions in (1.17) is to control for (continuous) trends in observables and unobservables moving away from the assignment variable threshold. Though not necessary if the RD empirical strategy is sound, it is common to additionally control for observables X in order to reduce the variance of the error term and more precisely estimate α_1. As with our discussion of including observables in the DD estimators, it is important not to include any observables that may respond to the treatment, meaning they are endogenous. Moreover, it is common not to utilize data beyond a certain distance from the threshold z_0 for estimation because such observations do not contribute to identification yet they can influence parametric estimates of the control functions.

The empirical model in (1.17) can also be used as a basis for estimating a LATE in environments that lend themselves to using a fuzzy RD research design. Here, however, the researcher must also consider the following auxiliary treatment equation:

$$T_i = \gamma_0 + D_i\rho + g_1(z_0 - Z_i)1(Z_i < z_0) + g_2(Z_i - z_0)1(Z_i \geq z_0) + X_i\nu + u_i,$$

where $D_i = 1(Z_i \geq z_0)$, and T_i in (1.17) is simply a treatment indicator. As this is now a simultaneous equations model, the fuzzy RD LATE can thus be estimated using any IV estimator. Commensurate with (1.16), the ILS estimate of the fuzzy RD LATE is $\frac{\hat{\alpha}_1}{\hat{\rho}}$.

Nonparametric estimation can also be used to recover the ATE at the discontinuity threshold—see Hahn et al. (2001). The randomization nature of the RD design implies that most estimation methods should lead to similar conclusions. If ATE estimates from different methods diverge, that is usually a symptom of a more fundamental problem, such as a small number of observations near z_0. In fact, the main practical limitation of nonparametric methods is that they require a large number of observations near the threshold, especially since nonparametric estimators are quite sensitive to bandwidth choice at boundaries.

To this point, we have assumed that we know the critical value z_0 of the assignment variable at which there is a discontinuous change in treatment probability. In some contexts, that critical value is unknown. It is possible to estimate the "structural break" z_0 jointly with the treatment effect at z_0. This can be done by estimating (1.17) by OLS for every candidate z_0, and then choosing the $\hat{z_0}$ that maximizes R^2. The work of Card David and Rothstein (2008) is one notable example in the urban economics literature that carries out this procedure. This paper recovers estimates of the critical fraction of the population that is black in neighborhoods at which they "tip," meaning they lose a large number of white residents. Jointly estimated with these tipping points are the magnitudes of this tipping.

1.6.3 Examples of RD in urban economics

There are various examples of RD applications in urban economics. Ferreira and Gyourko (2009) study the impacts of local politics on fiscal outcomes of US cities. Chay and Greenstone (2005) recover hedonic estimates of willingness to pay for air quality improvements in US counties. Baum-Snow and Marion (2009) estimate the impacts of low income housing subsidies on surrounding neighborhoods. Ferreira (2010) studies the impact of property taxes on residential mobility, and Pence (2006) studies the impact of mortgage credit laws on loan size. In this subsection we first discuss the bond referenda example that was mentioned above in detail. We then discuss the use of the "boundary discontinuity" research design, which is a particular application of RD that comes with its own challenges.

Cellini et al. (2010) investigate the importance of capital spending in education. There are two central barriers to identification in this setting. First, resources may be endogenous to local outcomes. Spending is usually correlated with the socioeconomic status of students. Second, even causal estimates of the impact of school investments may not be able to capture all measured benefits to students, such as nonacademic benefits. To deal with this second issue, they look at housing markets. Given standard theory (Oates, 1969), if home buyers value a local project more than they value the taxes they

pay to finance it, spending increases should lead to higher housing prices—also implying that the initial tax rate was inefficiently low.

In order to isolate exogenous variation in school investments, they create control and treatment groups based on school districts in California that had very close bond referenda. The logic is that a district where the proposal for a bond passes by one vote is likely to be similar to one where the proposal fails by the same margin. They test and confirm this assumption using three methods: they show that control and treatment groups have balanced covariates around the margin of victory threshold, they show that the prebond outcomes and trends of those outcomes are also balanced, and they show that the distribution of bond measures by vote share is not discontinuous around the threshold.

They also test whether the design is sharp or fuzzy by looking at the future behavior of districts after a bond referendum. Districts in which a bond referendum failed were more likely to pass and approve another bond measure within the next 5 years. The authors deal with the dynamic nature of bond referenda by developing two estimators of ITT and TT. The estimates indicate that passage of a bond measure causes house prices to rise by about 6%, with this effect appearing gradually over 2–3 years following the referendum, and the effect persists for about a decade. Finally, the authors convert their preferred TT estimates of the impact of bond passage on investments and prices into the willingness to pay for marginal home buyers. They find a marginal willingness to pay of $1.50 or more for each $1 of capital spending. Even though several papers in the public choice literature emphasize the potential for "Leviathan" governments, those estimates suggest the opposite for this California case.

We now consider the boundary discontinuity research design. Many researchers have used geographic boundaries to construct more comparable treatment and control groups that are likely to mitigate omitted variable biases. Holmes (1998), for example, aspires to disentangle the effects of state policies from other state-specific characteristics. As discussed in Section 1.4.2, a DD approach is often less than ideal when applied to large geographic areas such as states. Holmes's strategy is to zoom in on state borders at which one state has right-to-work laws and the other state does not. Geography, climate, fertility of soil, access to raw materials, and access to rivers, ports, etc., may be the same for cities on either side of the border. Such a design thus mitigates potential biases arising from differences in omitted factors. Looking across these borders, Holmes (1998) finds that manufacturing activity is much higher on the "probusiness" sides of the borders.

But borders are usually not randomly assigned. They may follow certain geographic features, such as rivers, or they may be the result of a political process, such as when states choose boundaries for congressional districts. The lack of randomization implies that there might be more than one factor that is not similar across geographic areas separated by boundaries. For example, some boundaries may be used to separate multiple jurisdictions, such as cities, school districts, counties, states, and perhaps countries. Even if

borders were randomly assigned, there is ample opportunity for sorting of agents or policies across borders on unobservable characteristics.

These issues can be illustrated in the example of valuation of school quality. Black (1999) compares house prices on either side of school attendance boundaries in order to estimate valuation of school quality on the high-quality side versus the low-quality side. Attendance zones rather than school district boundaries are used because no other local service provision is different on either side of these boundaries. School district boundaries would have two problems: they may also be city or county boundaries, and different districts may have very different systems of school financing. School attendance zones, on the other hand, have similar financing systems, and are unlikely to be used to separate other types of jurisdictions. Black also shows that the distance to the boundary matters. Only small distances, within 0.2 miles, are likely to guarantee similarity in local features.

However, even those precise local attendance zones may not deal with the issue of endogenous sorting of families. Given a discontinuity in local school quality at the boundary, one might expect that residential sorting would lead to discontinuities in the characteristics of the households living on opposite sides of the same boundary—even when the housing stock was initially identical on both sides. Bayer et al. (2007) empirically report those discontinuities for the case of the San Francisco Bay Area. High income, high education level, and white households are more likely to be concentrated on the high school quality side of the attendance zone boundaries. Those differences are noticeable even within very small distances to the boundary. Given these sorting patterns, it becomes important to control for neighborhood demographic characteristics when estimating the value of school quality, since the house price differences may reflect the discontinuities in school quality and also the discontinuities in sociodemographics. As in Black (1999), Bayer et al. (2007) find that including boundary fixed effects in standard hedonic regressions reduces the estimated valuation of school quality. But they also find that such valuation is reduced even further, by approximately 50%, when precise sociodemographic characteristics are added.

Additional caveats are that even the best data sets will not have all of the sociodemographic characteristics that may influence house prices. Also, most data sets have limited information about detailed characteristics of houses, such as type of floor and views. Biases may arise if such unobserved housing features or unobserved demographic characteristics differ across boundaries used for identification. These problems could be mitigated in settings where boundaries were recently randomly assigned, and therefore families or firms still did not have enough time to re-sort.

In another use of the boundary discontinuity empirical setup, Turner et al. (2014) examine land prices across municipal borders to decompose the welfare consequences of land use regulation into own lot, external, and supply components. The idea is that as long as land use regulation is enforced evenly over space up to municipal borders, one can recover the direct costs of regulation by comparing across borders. Indirect

(spillover) costs of regulation can be found with a spatial differencing type estimator within jurisdictions adjacent to those with regulatory changes. Supply effects of regulation are reflected in differences across municipal borders in the share of land that is developed. Results indicate strong negative effects of land use regulations on the value of land and welfare that operate through all three channels.

Recent developments in labor economics and public finance have also uncovered many discontinuities in slopes, using the so-called regression kink (RK) design (Card David and Weber, 2012). These kinks are a common feature of many policy rules, such as the formulas that establish the value of unemployment insurance benefits as a function of previous earnings. Card et al. explain that the basic intuition of the RK design is similar to that of the RD design and is based on a comparison of the relationship between the outcome variable (e.g., duration of unemployment) and the treatment variable (e.g., unemployment benefit levels) at the point of the policy kink. However, in contrast to an RD design, which compares the levels of the outcome and treatment variables, the estimated causal effect in an RK design is given by the ratio of the changes in the slope of the outcome and treatment variables at the kink point. As with RD, one threat to identification is sorting at the kink. This type of sorting often results in visible bunching in the distribution of the running variable at the kink point and invalidates the assumptions underlying the RK design. However, though such bunching may invalidate RD and RK designs, many researchers in public economics—such as Saez (2010) and Chetty et al. (2011)—have been able to leverage this type of bunching to recover estimates of the behavioral responses to various public policies such as income taxes. The idea in such "bunching designs" is to compare the actual bunching observed in the data with the predictions from a behavioral model that does not have the policy kink. Assuming everything else is constant, any differences between the amount of bunching observed in the data and the amount that would be implied by the model in the absence of the policy kink can be attributed directly to the policy variation around the kink. Recent applications of this approach to housing markets include Best and Kleven (2014), Kopczuk and Munroe (2014), and De Fusco and Paciorek (2014).

Finally, in some situations one may observe both an RD and an RK at the same threshold—see Turner (2012). New developments in these areas of research may arise in the coming years, as researchers thrive to understand the underlying sources of variation in the data that allow for identification of treatment effects that are difficult to credibly estimate with nonexperimental data.

1.7. CONCLUSION

This chapter has laid out some best practices for recovering causal empirical relationships in urban and regional economics contexts. We hope that we have successfully conveyed the idea that carrying out quality empirical work requires creativity and careful thought.

Beyond basic decisions about the general empirical strategy to be used are always many smaller decisions that are inherently particular to the question at hand and available data. In general, however, two central considerations should permeate all empirical work that aspires to recover causal relationships in data. The first is to consider the sources of variation in treatment variables that identify these relationships of interest. The second is to recognize which treatment effect, if any, is being estimated.

We see a bright future for empirical research in urban and regional economics. The wide integration of tractable economic theory and empirical inquiry among those working on urban and regional questions in economics positions our field well to make convincing progress on important questions. The wide range of detailed spatially indexed data available to us provides many opportunities for the beginnings of serious investigations of new topics. Indeed, while recovery of treatment effects is important, a descriptive understanding of important patterns in the data is perhaps more important for new questions. Particularly in our field, which is finding itself overwhelmed with newly available data, the first step should always be to get a handle on the facts. Doing so often leads to ideas about convincing identification strategies that can be used to recover causal relationships of interest.

REFERENCES

Abadie, A., Angrist, J., Imbens, G., 2002. Instrumental variables estimatesof the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70, 91–117.

Abadie, A., Diamond, A., Hainmueller, J., 2010. Synthetic control methods for comparative case studies: estimating the effect of california's tobacco control program. J. Am. Stat. Assoc. 105, 493–505.

Abadie, A., Diamond, A., Hainmueller, J., 2014. Comparative politics and the synthetic control method. Am. J. Polit. Sci. (Online, forthcoming).

Abadie, A., Gardeazabal, J., 2003. The economic costs of conflict: a case study of the basque country. Am. Econ. Rev. 93, 113–132.

Alesina, A., Baqir, R., Hoxby, C., 2004. Political jurisdictions in heterogeneous communities. J. Polit. Econ. 112, 348–396.

Altonji, J., Elder, T., Taber, C., 2005. Selection on observed andunobserved variables: assessing the effectiveness of catholic schools. J. Polit. Econ. 113, 151–184.

Angrist, J., Graddy, K., Imbens, G., 2000. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. Rev. Econ. Stud. 67, 499–527.

Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. Rev. Econ. Stat. 60, 47–57.

Athey, S., Imbens, G., 2006. Identification and inference in nonlinear difference-in-differences models. Econometrica 74, 431–497.

Autor, D., Palmer, C., Pathak, P., 2014. Housing market spillovers: evidence from the end of rent control in Cambridge Massachusetts. J. Polit. Econ. 122, 661–717.

Bailey, M., Muth, R., Nourse, H., 1963. A regression method for real estate price index construction. J. Am. Stat. Assoc. 58, 933–942.

Bartik, T., 1991. Who Benefits from State and Local Economic Development Policies? Upjohn Institute, Kalamzoo, MI.

Baum-Snow, N., 2007. Did highways cause suburbanization? Q. J. Econ. 122, 775–805.

Baum-Snow, N., Brandt, L., Henderson, J.V., Turner, M., Zhang, Q., 2014. Roads, Railroads and Decentralization of Chinese Cities (manuscript).

Baum-Snow, N., Lutz, B., 2011. School desegregation, school choice and changes in residential location patterns by race. Am. Econ. Rev. 101, 3019–3046.
Baum-Snow, N., Marion, J., 2009. The effects of low income housing tax credit developments on neighborhoods. J. Publ. Econ. 93, 654–666.
Baum-Snow, N., Pavan, R., 2012. Understanding the city size wage gap. Rev. Econ. Stud. 79, 88–127.
Bayer, P., Ferreira, F., McMillan, R., 2007. A unified framework for measuring preferences for schools and neighborhoods. J. Polit. Econ. 115, 588–638.
Bayer, P., Hjalmarsson, R., Pozen, D., 2009. Building criminal capital behind bars: peer effects in juvenile corrections. Q. J. Econ. 124, 105–147.
Bayer, P., Ross, S., Topa, G., 2008. Place of work and place of residence: informal hiring networks and labor market outcomes. J. Polit. Econ. 116, 1150–1196.
Beaudry, P., Green, D., Sand, B., 2014. Spatial equilibrium with unemployment and wage bargaining: theory and estimation. J. Urban Econ. 79, 2–19.
Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? Q. J. Econ. 119, 249–275.
Best, M.C., Kleven, H.J., 2014. Housing Market Responses to Transaction Taxes: Evidence from Notches and Stimulus in the UK. Mimeo.
Bester, A., Conley, T., Hansen, C., 2011. Inference with dependent data using cluster covariance estimators. J. Econometr. 165, 137–151.
Bjorklund, A., Moffitt, R., 1987. The estimation of wage gains and welfare gains in self-selection models. Rev. Econ. Stat. 69, 42–49.
Black, S., 1999. Do better schools matter? Parental valuation of elementary education. Q. J. Econ. 114, 577–599.
Blanchard, O.J., Katz, L.F., 1992. Regional evolutions. Brook. Pap. Econ. Act. 1, 1–69.
Bound, J., Holzer, H.J., 2000. Demand shifts, population adjustments and labor market outcomes during the 1980's. J. Labor Econ. 18, 20–54.
Boustan, L., Ferreira, F., Winkler, H., Zolt, E.M., 2013. The effect of income inequality on taxation and public expenditures: evidence from U.S. municipalities and school districts, 1970–2000. Rev. Econ. Stat. 95, 1291–1302.
Boustan, L.P., 2010. Was postwar suburbanization "white flight"? Evidence-from the black migration. Q. J. Econ. 125, 417–443.
Busso, M., Gregory, J., Kline, P., 2013. Assessing the incidence and efficiency of a prominentplace based policy. Am. Econ. Rev. 103, 897–947.
Cameron, A.C., Gelbach, J.B., Miller, D.L., 2008. Bootstrap-based improvements for inference with clustered errors. Rev. Econ. Stat. 90, 414–427.
Campbell, J., Giglio, S., Pathak, P., 2011. Forced sales and house prices. Am. Econ. Rev. 101, 2108–2131.
Card, D., 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. J. Labor Econ. 19, 22–64.
Card David, A.M., Rothstein, J., 2008. Tipping and the dynamics of segregation. Q. J. Econ. 123, 177–218.
Card David, David Lee, Z.P., Weber, A., 2012. Nonlinear policy rules and the identification and estimation of causal effects in a generalized regression kink design, NBER Working paper No. 18564.
Carrell, S., Sacerdote, B., West, J., 2013. From natural variation to optimal policy? The importance of endogenous peer group formation. Econometrica 81, 855–882.
Case, K., Shiller, R., 1987. Prices of Single Family Homes Since 1970: New Indexes for Four Cities. New England Economic Review, Boston, MA September/October.
Case, K., Shiller, R., 1989. The efficiency of the market for single-family homes. Am. Econ. Rev. 79, 125–137.
Cellini, S., Ferreira, F., Rothstein, J., 2010. The value of school facility investments: evidence from a dynamic regression discontinuity design. Q. J. Econ. 125, 215–261.
Chaney, T., Sraer, D., Thesmar, D., 2012. The collateral channel: how real estate shocks affect corporate investment. Am. Econ. Rev. 102, 2381–2409.
Chay, K., Greenstone, M., 2005. Does air quality matter? Evidencefrom the housing market. J. Polit. Econ. 113, 376–424.

Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D., Yagan, D., 2011. How does your kindergarten classroom affect your earnings? Evidence from project STAR. Q. J. Econ. 126, 1593–1660.
Combes, P.P., Duranton, G., Gobillon, L., 2008. Spatial wage disparities: sorting matters! J. Urban Econ. 63, 723–742.
Combes, P.P., Duranton, G., Gobillon, L., Roux, S., 2012. Sorting and local wage and skill distributions in france. Reg. Sci. Urban Econ. 42, 913–930.
Costa, D., Kahn, M., 2000. Power couples: changes in the locational choice of the college educated, 1940–1990. Q. J. Econ. 115, 1287–1315.
Cox, D.R., 1958. Some problems connected with statistical inference. Ann. Math. Stat. 29, 357–372.
De La Roca, J., Puga, D., 2014. Learning by Working in Big Cities (manuscript).
Dehejia, R., Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. Rev. Econ. Stat. 84, 151–161.
Diamond, R., 2013. The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000 (manuscript).
DiNardo, J., Lee, D., 2011. Program evaluation and research designs. In: Orley, A., David, C. (Eds.), Handbook of Labor Economics. Part A, Vol 4. Elsevier, Amsterdam, pp. 463–536.
Dinkelman, T., 2011. The effects of rural electrification on employment: new evidence from South Africa. Am. Econ. Rev. 101, 3078–3108.
Duflo, E., Glennerster, R., Kremer, M., 2008. Using randomization in development economics research: A toolkit. In: Srinivasan, T.N., Behrman, J. (Eds.), Handbook of Development Economics. Volume 4. Elsevier, Amsterdam, pp. 3895–3962.
Duranton, G., Morrow, P., Turner, M.A., 2014. Roads and trade: evidence from the U.S. Rev. Econ. Stud. 81, 681–724.
Duranton, G., Turner, M., 2011. The fundamental law of road congestion: evidence from the US. Am. Econ. Rev. 101, 2616–2652.
Duranton, G., Turner, M., 2012. Urban growth and transportation. Rev. Econ. Stud. 79, 1407–1440.
Efron, B., Tibishirani, R., 1994. An Introduction to the Bootstrap. Monograph in Applied Statistics and Probability, No 57, Chapman & Hall, New York, NY.
Ellen, I., Lacoe, J., Sharygin, C., 2013. Do foreclosures causecrime? J. Urban Econ. 74, 59–70.
Epple, D., Platt, G., 1998. Equilibrium and local redistribution in an urban economy when households differ in both preferences and incomes. J. Urban Econ. 43, 23–51.
Ferreira, F., 2010. You can take it with you: proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities. J. Publ. Econ. 94, 661–673.
Ferreira, F., Gyourko, J., 2009. Do political parties matter? Evidence from U.S. cities. Q. J. Econ. 124, 399–422.
Field, E., 2007. Entitled to work: urban property rights and labor supply in Peru. Q. J. Econ. 122, 1561–1602.
Figlio, D., Lucas, M., 2004. What's in a grade? School report cards and the housing market. Am. Econ. Rev. 94, 591–605.
Freedman, M., 2014. Tax Incentives and Housing Investment in Low Income Neighborhoods (manuscript).
Fusco, De, Anthony, A., Paciorek, A., 2014. The interest rate elasticity of mortgage demand: evidence from bunching at the conforming loan limit. Fin. Econ. Disc. Ser. 2014-11.
Galiani, S., Gertler, P., Cooper, R., Martinez, S., Ross, A., Undurraga, R., 2013. Shelter from the Storm: Upgrading Housing Infrastructure in Latin American Slums. NBER Working paper 19322.
Galiani, S., Murphy, A., Pantano, J., 2012. Estimating Neighborhood Choice Models: Lessons from a Housing Assistance Experiment (manuscript).
Gibbons, C., Serrato, J.C.S., Urbancic, M., 2013. Broken or Fixed Effects? Working paper.
Glaeser, E., Hedi Kallal, J.S., Shleifer, A., 1992. Growth in cities. J. Polit. Econ. 100, 1126–1152.
Glaeser, E., Maré, D., 2001. Cities and skills. J. Labor Econ. 19, 316–342.
Gobillon, L., Magnac, T., Selod, H., 2012. Do unemployed workers benefit from enterprise zones? The french experience. J. Publ. Econ. 96, 881–892.
Gould, E., Weinberg, B., Mustard, D., 2002. Crime rates and local labor market opportunities in the United States: 1979–1997. Rev. Econ. Stat. 84, 45–61.

Graddy, K., 1995. Testing for imperfect competition at the fulton fish market. Rand J. Econ. 26, 75–92.
Graham, B., 2008. Identifying social interactions through conditional variance restrictions. Econometrica 76, 643–660.
Greenstone, M., Gallagher, J., 2008. Does hazardous waste matter? Evidence from the housing market and the superfund program. Q. J. Econ. 123, 951–1003.
Greenstone, M., Hornbeck, R., Moretti, E., 2010. Identifying agglomeration spillovers: evidence from winners and losers of large plant openings. J. Polit. Econ. 118, 536–598.
Gronau, R., 1974. Wage comparisons. a selectivity bias. J. Polit. Econ. 82, 1119–1143.
Hahn, J., Todd, P., van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69, 201–209.
Ham, J., Swenson, C., Imbroglu, A., Song, H., 2011. Government programs can improve local labor markets: evidence from state enterprise zones, federal empowerment zones and federal enterprise community. J. Publ. Econ. 95, 779–797.
Hanson, A., 2009. Local employment, poverty, and property value effects of geographically-targeted tax incentives: an instrumental variables approach. Reg. Sci. Urban Econ. 39, 721–731.
Hanson, A., Rohlin, S., 2011. The effect of location based tax incentives on establishment location and employment across industry sectors. Publ. Financ. Rev. 39, 195–225.
Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–162.
Heckman, J., Honoré, B., 1990. The empirical content of the roy model. Econometrica 58, 1121–1149.
Heckman, J., Navarro-Lozano, S., 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. Rev. Econ. Stat. 86, 30–57.
Heckman, J., Urzua, S., Vytlacil, E., 2006. Understanding instrumental variables in models with essential heterogeneity. Rev. Econ. Stat. 88, 389–432.
Heckman, J., Vytlacil, E., 2005. Structural equations, treatment effects, and econometric policy evaluation. Econometrica 73, 669–738.
Henderson, V., Kuncoro, A., Turner, M., 1995. Industrial development in cities. J. Polit. Econ 103, 1067–1090.
Holland, P., 1986. Statistics and causal inference. J. Am. Stat. Assoc. 81, 945–960.
Holmes, T., 1998. The effects of state policies on the location of industry: evidence from state borders. J. Polit. Econ. 106, 667–705.
Hoxby, C., 2000. Does competition among public schools benefit students and taxpayers? Am. Econ. Rev. 90, 1209–1238.
Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–475.
Imbens, G., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. J. Econometr. 142, 615–635.
Imbens, G., Wooldridge, J., 2007. Control function and related methods. In: What's New In Econometrics? NBER Lecture Note 6.
Kain, J.F., 1992. The spatial mismatch hypothesis: three decades later. Hous. Pol. Debate 3, 371–462.
Katz, L.F., Kling, J.R., Liebman, J.B., 2001. Moving to opportunity in Boston: early results of a randomized mobility experiment. Q. J. Econ. 116, 607–654.
Kline, P., 2011. Oaxaca-blinder as a reweighting estimator. Am. Econ. Rev. 101, 532–537.
Kline, P., Moretti, E., 2014. Local economic development, agglomeration economies, and the big push: 100 years of evidence from the Tennessee valley authority. Q. J. Econ. 129, 275–331.
Kling, J., Liebman, J., Katz, L., 2007. Experimental analysis of neighborhood effects. Econometrica 75, 83–119.
Kolesar, M., Chetty, R., Friedman, J., E.G., 2013. Identification and Inference with Many Invalid Instruments (manuscript).
Kopczuk, W., Munroe, D.J., 2014. Mansion tax: the effect of transfer taxes on the residential real estate market. Am. Econ. J. Econ. Pol. (forthcoming).
Kuminoff, N.V., Smith, V.K., Timmins, C., 2013. The new economics of equilibrium sorting and policy evaluation using housing markets. J. Econ. Liter. 51, 1007–1062.
Lee, D., Lemieux, T., 2010. Regression discontinuity designs in economics. J. Econ. Liter. 48, 281–355.

Levitt, S., 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. Am. Econ. Rev. 87, 270–290.
Lewis, E., 2011. Immigration, skill mix, and capital skill complementarity. Q. J. Econ. 126, 1029–1069.
Linden, L., Rockoff, J., 2008. Estimates of the impact of crime risk onproperty values from megan's laws. Am. Econ. Rev. 98, 1103–1127.
Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., Sanbonmatsu, L., 2013. Long-term neighborhood effects on low-income families: evidence from moving to opportunity. Am. Econ. Rev. 103, 226–231.
Luttmer, E., 2005. Neighbors as negatives: relative earnings and well-being. Q. J. Econ. 130, 963–1002.
McCrary, J., 2002. Using electoral cycles in police hiring to estimate the effect of police on crime: comment. Am. Econ. Rev. 92, 1236–1243.
McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: a density test. J. Econometr. 142, 698–714.
McMillen, D., McDonald, J., 2002. Land values in a newly zoned city. Rev. Econ. Stat. 84, 62–72.
Mian, A., Sufi, A., 2009. The consequences of mortgage credit expansion: evidence from the U.S. mortgage default crisis. Q. J. Econ. 124, 1449–1496.
Michaels, G., 2008. The effect of trade on the demand for skill—evidence from the interstate highway system. Rev. Econ. Stat. 90, 683–701.
Moulton, B., 1986. Random group effects and the precision of regressionestimates. J. Econometr. 32, 385–397.
Moulton, B., 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. Rev. Econ. Stat. 72, 334–338.
Neal, D., 1997. The effects of catholic secondary schooling on educational achievement. J. Labor Econ. 15, 98–123.
Notowidigdo, 2013. The Incidence of Local Labor Demand Shocks (manuscript).
Oates, W.E., 1969. The effects of property taxes and local public spending on property values: an empirical study of tax capitalization and the tiebout hypothesis. J. Polit. Econ. 77, 957–971.
Oster, E., 2013. Unobservable Selection and Coefficient Stability: Theory and Validation. Working paper.
Pearl, J., 2009. Causal inference in statistics: an overview. Stat. Surv. 3, 96–146.
Pence, K.M., 2006. Foreclosing on opportunity: state laws and mortgage credit. Rev. Econ. Stat. 88, 177–182.
Redding, S., Sturm, D., 2008. The costs of remoteness: evidence from german division and reunification. Am. Econ. Rev. 98, 1766–1797.
Roback, J., 1982. Wages, rents and the quality of life. J. Polit. Econ. 90, 1257–1278.
Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. J. Polit. Econ. 82, 34–55.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rosenthal, S., 2014. Are private markets and filtering a viable source of low-income housing? Estimates from a "repeat income" model. Am. Econ. Rev. 104, 687–706.
Rothstein, J., 2007. Does competition among public schools benefit students and taxpayers? A comment on hoxby (2000). Am. Econ. Rev. 97, 2026–2037.
Roy, A.D., 1951. Some thoughts on the distribution of earnings. Oxf. Econ. Pap. New Ser. 3, 135–146.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66, 688–701.
Sacerdote, B., 2001. Peer effects with random assignment: results for Dartmouth roommates. Q. J. Econ. 116, 681–704.
Saez, E., 2010. Do taxpayers bunch at kink points? Am. Econ. J. Econ. Pol. 2, 180–212.
Saiz, A., 2010. The geographic determinants of housing supply. Q. J. Econ. 125, 1253–1296.
Schwartz, A.E., Ellen, I.G., Voicu, I., Schill, M., 2006. The external effects of place-based subsidized housing. Reg. Sci. Urban Econ. 36, 679–707.
Serrato, S., Carlos, J., Wingender, P., 2014. Estimating Local Fiscal Multipliers (manuscript).

Stock, J., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In: Stock, J., Andrews, D. (Eds.), Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg. Cambridge University Press, Cambridge, pp. 109–120.
Tiebout, C., 1956. A pure theory of local expenditures. J. Polit. Econ. 64, 416–424.
Turner, M.A., Haughwout, A., van der Klaauw, W., 2014. Land use regulation and welfare. Econometrica 82, 1341–1403.
Turner, N., 2012. Who benefits from student aid? The economic incidence of tax based federal student aid. Econ. Educ. Rev. 31, 463–481.
Wooldridge, J., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.
Wooldridge, J., 2005. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. Port. Econ. J. 1, 117–139.