Three Essays on Semiparametric Methods for the Evaluation of Social Programs

by

Matias Busso

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2008

Doctoral Committee:

Professor Jeffrey A. Smith, Chair
Professor John E. DiNardo
Professor Serena Ng
Professor Jan Svejnar
Assistant Professor Justin R. McCrary, University of California, Berkeley

To Mariana

Acknowledgements

The chapters in this dissertation benefited greatly from multiple conversations with faculty and students at the University of Michigan. I owe an enormous intellectual debt to many people. Jeff Smith inspired me to only work on questions I cared about. I thank him for his feedback, advice, encouragement, and humor. Justin McCrary and John DiNardo are outstanding teachers, mentors and faithful coauthors. They played a fundamental role in shaping my vocation and understanding of applied microeconometrics, and were always very generous with their time and ideas. I spent many hours working alongside Justin whose patience, energy and generosity are, I believe, limitless. John's unorthodox office hours always turned up being a fruitful exchange of ideas of all sorts. I thank him for his encouragement and advice, and for teaching that amazing class when I was in my second year. I am also very grateful to Serena Ng. First, for teaching me econometrics. Second, for her support, care and willingness to make of me a better scholar. I also thank Jan Svejnar and Lutz Kilian for their backing and support over all these years. John Bound and Charlie Brown provided comments that greatly improved the second chapter of this dissertation. I am also indebted to fellow graduate student, friend and coauthor Patrick Kline. I was very lucky to have crossed paths with Pat at Michigan. I learned a lot by working with him. Not only is he an excellent economist, he is also unstoppable.

Leonardo Gasparini, Huberto Ennis, Guido Porto, Walter Sosa-Escudero and Jose Wynne at the Universidad Nacional de La Plata in Argentina, were very influential and helpful with my plans to pursue a graduate education in the US. Federico Cerimedo, long-time friend and colleague, helped me to become a much better economist when we were both giving the first steps in the profession.

My family, many old friends in Argentina, and many new friends in Ann Arbor provided unconditional support during all these years. My greatest debt is to my wife, Mariana, who put up patiently with the many late nights and weekends of my work, and whose love and care at home allowed me to finish this project, and made our life in Ann Arbor much happier.

Table of Contents

Chapter

List of Tables

Table

List of Figures

Figure

Chapter 1

Introduction

This dissertation studies semiparametric methods for the evaluation of social programs. The first essay, with Patrick Kline, evaluates Round I of the federal urban Empowerment Zone (EZ) program, which constitutes one of the largest standardized federal interventions in impoverished urban American neighborhoods since President Johnson's Model Cities program. The EZ program is a series of spatially targeted tax incentives and block grants designed to encourage economic, physical, and social investment in the neediest urban and rural areas in the United States. We use four decades of Census data on urban neighborhoods in conjunction with information on the proposed boundaries of rejected EZs to assess the impact of Round I EZ designation on local labor and housing market outcomes over the period 1994-2000. Utilizing a semiparametric difference-in-differences estimator we find that neighborhoods receiving EZ designation experienced substantial improvements in labor market conditions and moderate increases in rents relative to rejected and future Empowerment Zones. These effects were accompanied by small changes in the demographic composition of the neighborhoods, though evidence from disaggregate Census tabulations suggests that these changes account for little of the observed improvements. No evidence exists of large scale gentrification, indicating that many of the benefits (and costs) of the program have been captured by pre-existing residents.

The second essay, with John DiNardo and Justin McCrary, explores the finite sample properties of several semiparametric estimators of average treatment effects, including propensity score inverse probability weighting (IPW), matching, and double robust estimators. When there is good overlap in the distribution of propensity scores for treatment and control units, IPW estimators are preferred on bias grounds and attain the semiparametric efficiency bound even for samples of size $n = 100$. Pair matching exhibits similarly good performance in terms of bias, but has notably higher variance. Local linear and ridge matching are competitive with reweighting in terms of bias and variance, but only once $n = 500$. Nearest-neighbor, kernel, and blocking matching are not competitive. When overlap is close to failing, none of the estimators examined perform well and $\sqrt{n}$ -asymptotics may be a poor guide to finite sample performance. Trimming rules, commonly used in the face of problems with overlap, are effective only in settings with homogeneous treatment effects.

In the third essay I propose a sequential method of moments variance estimator of IPW estimators of average treatment effects. IPW estimators are becoming increasingly popular to compute average treatment effects. Obtaining valid standard errors for these estimators, however, can be difficult because of the 2-step nature of the estimation procedure. In this essay, I note that IPW is a sequential method of moments (SQMM) estimator which, in cases in which a parametric propensity score model is assumed, has a simple expression of the asymptotic variance. This variance estimator can be used to test not only hypotheses about treatment effects for a given outcome but also hypotheses involving multiple outcomes and/or different estimands. Using Monte Carlo simulations I find that tests based on the proposed SQMM variance estimator have good finite sample size and power compared to competing inference strategies. Tests that ignore the fact that the weights are estimated tend to severely overreject. Tests based on the percentile-t bootstrap method using a bootstrap SQMM variance have very similar size and power properties as the ones obtained using the asymptotic SQMM variance. I interpret this as evidence that the bootstrap percentile-t method is not providing any refinement to the asymptotic variance, which indicates that the SQMM variance estimator is a good enough approximation to the true variance of the treatment effect estimator.

Chapter 2

Do Local Economic Development Programs Work? Evidence from the Federal Empowerment Zone Program[1]

Local economic development programs are an important, yet understudied, feature of the U.S. tax and expenditure system. Timothy Bartik (2002) estimates that state and local governments spend \$20-30 billion per year on economic development programs with an additional \$6 billion per annum coming from the federal government. However, little academic work has been done examining the impact of these expenditures on local communities, largely because of the small scale and general diversity of most such programs.[2] This paper evaluates the federal urban Empowerment Zone (EZ) program, which constitutes one of the largest standardized federal interventions in impoverished urban American neighborhoods since President Johnson's Model Cities program.

With a mandate to revitalize distressed urban communities, the EZ program represents a nexus between social welfare policy and economic development efforts. Unlike conventional anti-poverty programs, Empowerment Zones aim to help the poor by subsidizing demand for their services at local firms, which has made them one of the few social welfare programs popular on both sides of the congressional aisle. In an era where non-entitlement spending on social welfare programs has been scaled back dramatically, the federal Empowerment Zone program has enjoyed rapid growth. After the initial funding of six first round EZs and two "supplemental" EZs in 1994, fifteen more cities were awarded zones in 1999, followed by another eight in 2001. An additional forty-nine urban areas were concurrently granted smaller Enterprise Communities (ECs) which entailed a reduced package of benefits. The enthusiasm for spatially targeted tax credits has led to the birth of a

[1] This paper was written with Patrick Kline.

The authors would like to thank Soren Anderson, Timothy Bartik, John Bound, Charlie Brown, Kerwin Charles, John DiNardo, Taryn Dinkelman, Jesse Gregory, Jim Hines, Ben Keys, Justin McCrary, Gary Solon, Joel Slemrod, and Jeff Smith for encouragement and advice on this project. We would also like to thank participants of the University of Michigan Labor Seminar, the Michigan Public Finance Brownbag Lunch, and the Upjohn Institute Seminar for useful comments.

[2] See Bartik (1991) and the volume by Nolan and Wong (2004) for a review.

variety of new zones, each modifying the original EZ concept in different ways.[3] Most recently, the justification for tax abatement zones has been expanded to include disaster relief. For example, in the wake of the September 11th attacks, parts of New York city were designated "Liberty Zones" and granted a variety of localized tax credits; while, in 2006, Congress passed legislation authorizing a set of "Gulf Opportunity Zones" for areas stricken by Hurricane Katrina.

These recent forays of the IRS into the business of local economic development should merit the attention of economists. The GAO (1999) estimates that the first round Empowerment Zones will cost $2.5 billion over the course of the ten year program. Given that EZ neighborhoods have a total population of under a million people, subsidies of this magnitude, when directed to such relatively small urban areas, might be expected to have important effects upon the behavior of firms and workers. Measuring the nature and magnitude of these behavioral responses is crucial for understanding the equity-efficiency tradeoffs inherent in geographically targeted transfers.[4]

The EZ program was pre-dated by a series of state initiated "enterprise zones" which varied dramatically in scale, purpose, and implementation.[5] A modest literature evaluating the state level programs reaches mixed conclusions reflecting, in part, the enormous diversity of the programs under examination.[6] Some programs only provide for investment subsidies while others include employment tax credits; some state zones cover hundreds of square miles, while others are focused on particular neighborhoods within a few cities. Besides differences in the structure of the programs themselves, a number of methodological problems hinder clear interpretation of the enterprise zone literature. Many of the early studies faced difficulties obtaining data corresponding to the boundaries of the state zones, relying instead upon evaluations at higher levels of aggregation such as the zip code or city which likely reduced the statistical power of the estimates. Furthermore, most studies rely upon simple variants of the differences in differences research design without examining in any detail the suitability of the control groups being used to proxy the counterfactual behavior of the zones (a notable exception being Boarnet and Bogart (1996)). Finally, all of the studies of which we are aware save for Papke (1994) calculate standard errors ignoring issues of spatial and temporal dependence in the data making it difficult to assess exactly how precise previous studies have been and whether the differences in results are attributable to chance.

[3] In addition to urban EZs and ECs, there are a series of rural EZs and ECs, Enhanced Enterprise Communities (EECs), and 28 urban and 12 rural "Renewal Communities" entitled to benefits similar in magnitude to EZs.

[4] See Nichols and Zeckhauser (1982) for an introduction to the economics of targeting.

[5] See Papke (1993) and Hebert et al. (2001) for a history of the Empowerment and Enterprise Zone ideas.

[6] See Papke (1993, 1994), Boarnet and Bogart (1996), Bondonio (2003), Bondonio and Engberg (2000), Elvery (2003), and Engberg and Greenbaum (1999). Peters and Peters and Fisher (2002) provide a review.

The federal EZ program is much larger in scope and scale than its state level precursors and involves a standardized package of fiscal benefits applied to neighborhoods defined in terms of 1990 census tracts. Unlike most state level zones, the EZ program ties business tax credits to the employment of local residents and includes a series of large block grants aimed at reducing poverty and improving local infrastructure. The only large scale study of the impact of EZ designation is an interim evaluation (Hebert et al., 2001) performed for HUD by Abt Associates in conjunction with the Urban Institute, which finds that EZs had large effects on job creation, with increases in local payrolls on the order of 10%.

The Abt study suffers from a number of important weaknesses. First, it relies upon within city comparisons of census tracts which are likely to overstate the effect of the program if EZ designation merely reallocates jobs between neighborhoods. Second, the matching algorithm used to find controls for the EZ tracts is poorly documented and standard errors are not provided making it difficult to draw strong conclusions regarding the results. Moreover, important questions exist about the quality and representativeness of the Dunn and Bradstreet data used in the analysis.[7] Third, since local governments designed Empowerment Zone boundaries, it is possible that census tracts awarded EZs would have improved relative to other tracts in the same city even in the absence of EZ designation if the boundaries were drawn based upon trends emerging at the beginning of the 1990s. Finally, the study provides no guidance as to whether the jobs being created in EZs were staffed by local residents, whether the neighborhood composition of EZ residents changed, and whether poverty, unemployment, or the local housing market responded to the treatment–questions that are key to evaluating the success or failure of the program.

This paper uses four decades of census data on local neighborhoods in conjunction with proprietary EZ application data obtained from HUD to assess the impact of Round I EZ designation on residential sorting behavior and local labor and housing market outcomes over the period 1994-2000.[8] Unlike previous studies we use census tracts in rejected and future Empowerment Zones as controls for first round EZs. Since these tracts were nominated for EZ designation by their local governments, they are likely to share unobserved traits and trends in common with first round EZs which also underwent a local nomination phase. We present an extensive body of evidence indicating that these controls serve as good proxies for the counterfactual behavior of EZ tracts over the 1990s. Moreover, because most of our control tracts are in different cities than those winning EZs, they are substantially less susceptible to contamination by spillover or general equilibrium effects than those of previous studies. We use a variety of semiparametric methods to adjust for

[7] See Heeringa and Haeussler (1993) and Appendix A of the Abt report.

[8] The outcomes are: poverty, employment, unemployment, owner occupied housing values, rents, mean earnings, population, the fraction of houses that are vacant, the fraction of the neighborhood that is black, the fraction of residents who live in the same house as five years ago, and the fraction of residents who hold a college degree.

the small observable differences that do exist between our control tracts and EZs and to increase the statistical power of our analysis.

We find that neighborhoods receiving EZ designation experienced substantial improvements in the labor market outcomes of zone residents and moderate increases in housing values and rents relative to observationally equivalent tracts in rejected and future zones. These effects were accompanied by small changes in the demographic composition of the neighborhoods. We provide evidence from disaggregate census tabulations that the observed improvements in the local labor market conditions of EZ neighborhoods are unlikely to have resulted from these demographic changes alone. Employment rates, for example, seem to have increased even among young high school dropouts. However, given the high rates of turnover in EZ neighborhoods we cannot determine whether the benefits of EZ designation were captured by pre-existing residents or new arrivals with similar demographic characteristics.

An impact analysis is performed indicating that the EZ program created approximately $1 billion of additional wage and salary earnings in EZ neighborhoods and another $1 billion in property wealth. A comparison of IRS data with our impact estimates suggests that the tax credits associated with EZ designation are unlikely to have been the only source of the observed employment gains. Rather, we conclude that the block grants and outside funds leveraged by EZ designation, perhaps in conjunction with changes in expectations associated with EZ status, are likely to have contributed substantially to the changes in the local labor market.

The remainder of the paper is structured as follows: Section I provides background on the EZ program, Section II discusses the expected impact of EZ benefits, and Section III describes the data used. Section IV introduces the identification strategy and details the methodology used, Section V discusses results and tests for violations of the assumptions underlying our identification strategy. Section VI provides an impact analysis and Section VII concludes.

I A Crash Course in Empowerment

The federal Empowerment Zone program is a series of spatially targeted tax incentives and block grants designed to encourage economic, physical, and social investment in the neediest urban and rural areas in the United States. Talk of a federal program caught on early in President Clinton's first term following the 1992 Los Angeles riots. In 1993, Congress authorized the creation of a series of Empowerment Zones and smaller Enterprise Communities (ECs) that were to be administered by the Department of Housing and Urban Development (HUD) and awarded via a competitive application process.

Communities were invited to create their own plans for an EZ and submit them to HUD for consideration. Plans included the boundaries of the proposed zone, how community development funds would be used, and how state and local governments and community organizations would take actions to complement the federal assistance. In addition to providing a guidebook to communities hoping to apply, HUD held a series of regional workshops to explain the EZ initiative and the requisite application process. Nominating local governments were required to draw up EZ boundaries in terms of census tracts, list key demographic characteristics of each proposed tract including the 1990 poverty rate as measured in the Decennial Census, and specify whether the tracts were contiguous or located in the central business district.[9]

HUD initially awarded EZs to six urban communities: Atlanta, Baltimore, Chicago, Detroit, New York City, and Philadelphia/Camden. Two additional cities, Los Angeles and Cleveland, received "supplemental" EZ (SEZ) designation but were awarded full EZ designation two years later. Forty-nine rejected cities were awarded ECs. Table 2.1 shows summary statistics of EZ neighborhoods by city. The average Round I EZ spanned 10.6 square miles, contained 117,399 people, and had a 1990 poverty rate of 45%. Most zones are contiguous groupings of census tracts, although some EZs, such as the one in Chicago pictured in Figure 2.1, cover multiple disjoint groupings of tracts.

EZ designation brought with it a host of fiscal and procedural benefits, which we briefly summarize here:[10]

1. Employment Tax Credits —Starting in 1994, firms operating in the six original EZs became eligible for a credit of up to 20 percent of the first $15,000 in wages earned in that year by each employee who lived and worked in the community.[11] Tax credits for each such employee were available to a business for as long as ten years, with the maximum annual credit per employee declining over time. This was a substantial subsidy given that, in 1990, the average EZ worker only earned approximately $16,000 in wage and salary income.

2. Title XX Social Services Block Grant (SSBG) Funds —Each EZ became eligible for $100 million in SSBG funds, while each SEZ was eligible for $3 million in SSBG funds. These funds could be used for such purposes as: training programs, youth services, promotion of home ownership, and emergency housing assistance.

3. Section 108 Loan Guarantees/Economic Development Initiative (EDI) Grants —EDI funds are large flexible grants which are meant to be used in conjunction with other sources of HUD funding to facilitate large scale physical development projects. The

[9] For example, the application asked "Does any tract that includes the central business district have a poverty rate of less than 35%?" and "Do all census tracts of the nominated zone have 20% or more poverty rate?"

[10] See IRS (2004) for more details.

[11] Firms located in the two supplemental Empowerment Zones did not become eligible for the tax credit until 1999.

two SEZ's, Los Angeles and Cleveland, received EDI grants of $125 and $87 million respectively. The six original EZs were not eligible for these grants. Section 108 Loan Guarantees allow local governments to obtain loans for economic development projects. Los Angeles received $325 million in 108 loan guarantees and Cleveland received $87 million.

4. Enterprise Zone Facility Bonds —State and local governments can issue tax-exempt bonds to provide loans to qualified businesses to finance certain property. A business cannot receive more than $3 million in bond financing per zone or $20 million across all zones nationwide.

5. Increased Section 179 Expensing —Section 179 of the Internal Revenue Code provides write-offs for depreciable, tangible property owned by businesses in designated zones. Qualified target area business taxpayers could write off $20,000 more than the usual first-year maximum (which in 1994 was $18,000).

6. Regulatory Waivers/Priority in Other Federal Programs —Qualified EZ/EC areas were given priority in other Federal assistance programs. Furthermore, as part of their applications, EZ/EC applicants were encouraged to request any waivers in Federal program requirements or restrictions that were felt to be necessary for the successful implementation of their local revitalization strategy.

The subsidies available to zone businesses increased substantially over the first four years of the program with the surprise introduction of two additional wage credits (the Work Opportunity Tax credit and the Welfare to Work Tax Credit),[12] an expansion of the EZ Facility Bonds program, and changes in the treatment of capital gains realized from the sale of EZ assets. By all accounts, the degree of *potential* fiscal intervention in EZ neighborhoods was substantial.[13]

Nevertheless, it is difficult to assess exactly how extensive participation in the program has been. GAO (1999) estimated that the EZ program would cost $2.5 billion over its ten year life with 95 percent of the costs coming from the employment credit.[14] IRS data show that, in the year 2000, close to five hundred corporations, and over five thousand individuals, claimed EZ Employment Credits worth a total of approximately $23.5 and $22 million, respectively.[15] Roughly $200 million in employment credits were claimed over the period 1994 to 2000, with the amount claimed each year trending up steadily over time.

[12] Work Opportunity Tax Credits enabled businesses to claim up to $2,400 per worker in tax credits for first year wages paid to qualifying employees such as ex-felons, and youth ages 18-24 who are zone residents. Welfare to Work Tax Credits allow businesses to claim credits for up to $3,500 of first year and $5,000 of second year wages paid to workers who are long-term recipients of family assistance.

[13] While the SSBG and EDI funds were fungible, the wage credits and capital write offs were relatively narrowly targeted. Wages paid to workers employed for less than ninety days or relatives were not eligible for the wage credits nor were payments to unofficial workers not on the payroll. Similarly, for a business to be eligible for the tax exempt bond financing or the increased Section 179 expensing it must be able to demonstrate that the majority of its income is earned within the zone and that 35% of its employees are zone residents.

[14] The EZ program has subsequently been extended to expire in 2009.

[15] These figures come from GAO (2004).

So despite the slow ratcheting up of participation, reasonably large tax subsidies have been dispensed to EZ neighborhoods in the form of wage subsidies. In contrast, only 17 EZ facility bonds were issued before 2000 totalling approximately $50 million, so the impact of the tax exempt bond financing is probably minimal.

Survey data provide information about who participated in the tax incentives and why. A 1997 survey of zone businesses conducted by HUD found that most firms were unaware of the existence of the EZ program, that only 11% claimed to be using the wage tax credit, and only 4% claimed to be using the Section 179 deductions.[16] Such figures mask heterogeneity in participation rates by firm size. The HUD survey found that large firms used the tax credits more intensively with 63% and 30% utilization rates for the wage subsidies and capital write-offs respectively.[17] Another survey conducted by the GAO (1999) found that 55% of large urban businesses using the employment credits were manufacturing firms. The most commonly cited reasons for not using the wage credits were that firms were either unaware of the benefits or did not qualify for them because their employees lived outside of the zone. However, even among large firms, 27% responded that they were not aware of the credit. The low rates of participation in the Section 179 write-off program were most often attributed to lack of knowledge about the program and ineligibility due to lack of profits or qualifying investments. Since tax credits can only be claimed against a company's taxable profits, many small firms (15%), appear to have been unable to take advantage of the program due to insufficient taxable income.

Although the tax benefits accompanying EZ designation were somewhat underutilized by firms, the General Accounting Office (2004) estimates that state agencies had drawn down approximately 60% of Round I SSBG funds by 2003 and were on target to fully expend their allocations by the expiration of the program in 2010. More difficult to measure is the degree of outside investment leveraged by EZ designation. While the first round EZs were allocated roughly $800 million dollars in SSBG and EDI funds, the annual reports of the various EZs suggest that massive amounts of outside capital have accompanied the grant spending. HUD (2003) claims that $12 billion in public and private investment have been raised from Federal "seed" money accompanying the broader EZ/EC program. Our own analysis of HUD data suggests that the amount spent on first round EZs over the period 1994-2000 is substantially less than this, but still much greater than the initial amount of block grant funding allocated.

Table 2.2 summarizes information from HUD's internal performance monitoring system on the amount of money spent on various program activities by source. Audits by HUD's

[16] These figures come from Hebert et al. (2001).

[17] See tables 3-13, 3-14 and 3-15 in Hebert et al. (2001). The sample sizes used in the survey are not large enough to make strong inferences regarding the relationship between size and participation.

Office of Inspector General[18] and the GAO (2006)[19] have called the accuracy of these data into question, so the figures reported should be interpreted with caution. The six original EZs reported spending roughly \$2 billion by 2000, with more than four dollars of outside money accompanying every dollar of SSBG funds. The most commonly reported use of funds was enhancing access to capital. One-stop capital shops providing loans to EZ businesses and entrepreneurs were a component of the plans of most EZs. In Detroit, a consortium of lenders provided \$1.2 billion to be used in a local loan pool. Although these funds are listed as being spent, it is difficult to know what fraction were actually loaned out. Analysis of the HUD data in Hebert et al. (2001) indicates that the total size of all loan pools across the six original EZs was only \$79 million. The second most common use of the funds was business development which included technical and financial assistance. Third and fourth most common respectively were expenditures on housing development and public safety.

Compiling the tax and expenditure information together and allowing for biases in the reporting behavior of EZs, we estimate that the EZ program resulted in expenditures over the period 1994-2000 of between one and three billion dollars. While this amount of expenditure is below what was originally envisaged at the inception of the program, it is still quite substantial considering that together the EZs constitute a 92 square mile area containing less than a million residents.

II Expected Impact

The benefits accompanying EZ designation might be expected to impact a number of features of local communities.[20] Here we consider the aggregate variables most likely to respond to the treatment and the economic interpretation of those responses.

The wage subsidies should have two effects on local labor markets, both militating towards increased employment of zone residents. First, there should be a scale effect in that the average cost of labor should fall and production should expand. Second, there should be a substitution effect as outside workers are replaced by cheaper zone workers. If outside workers are relatively unwilling to relocate to EZ neighborhoods and zone residents vary substantially in their disutility of work, then we might expect any employment increases to be accompanied by corresponding increases in local wages.

[18] See Chouteau (1999) and Wolfe (2003).

[19] While the GAO could not find suitable documentation corroborating the dollar amount spent on each program, they were able to verify HUD data on the number of activities undertaken. Their analysis of this data indicated that "community development" projects which include "workforce development, human services, education, and assistance to businesses" accounted for more than 50 percent of the activities implemented in the 6 original urban EZs.

[20] See Papke (1993) for a general equilibrium model of the effects of localized tax incentives.

If firms are only willing to hire the most qualified workers from a neighborhood, then employment gains need not be accompanied by reductions in poverty as the relatively high skilled workers will merely shift from one job to another. Likewise, if EZ neighborhoods lack residents with the sorts of skills desired by firms then the wage subsidies may not be successful in increasing neighborhood employment as firms will not find it profitable to hire unproductive workers even at a substantial discount.

To the extent that block grants and other subsidies increase the profitability of local businesses, such as by alleviating capital constraints, providing technical assistance, or reducing crime, a scale effect should ensue, leading to an increase in the number of jobs inside EZs.[21] Moreover, if, as suggested by HUD's administrative data, a substantial portion of funds are being invested in workforce development and the matching of workers to local employers, we should expect local employment of zone residents to increase. Funds spent on improvement of infrastructure and physical redevelopment might also be expected to temporarily increase local employment in the form of construction jobs.

Housing markets should respond in tandem with zone labor markets. Firms and residential developers[22] may bid up the price of zone land in pursuit of EZ benefits if those benefits are deemed valuable. Likewise, block grants and outside investments in physical development and community safety are likely to improve the amenities associated with EZs, possibly stimulating residential demand in the area.[23] The asset values of land and owner occupied housing may rise quickly if expectations of future market conditions are influenced by EZ designation and there are obstacles in the short run to increasing housing supply. Rental rates, by contrast, will reflect supply and demand conditions in the spot market for housing. However if zone amenities improve, or if outside workers seek to migrate to the zone in anticipation of future neighborhood improvements, quality adjusted rents will rise.[24] Over longer time horizons the supply of housing may increase or the quality of the housing stock may adjust, both of which should moderate any price effects.

Since most zone residents are renters, large increases in rents may lead to gentrification and neighborhood churning as more affluent newcomers displace prior zone residents. To the extent that gentrification does occur, it should be reflected in changes in the demographic composition of zone neighborhoods. Increases in the price of land might also be

[21] Reductions in the price of capital should also bring with them a substitution effect as capital is substituted for labor. In theory this effect could outweigh the scale effect and yield negative employment effects if capital and low skilled labor are gross substitutes. We consider such extreme cases implausible. However, the substitutability of capital and low-skill labor may be expected to result in fairly small net impacts on employment.

[22] EDI and SSBG funds are targeted towards the development of affordable housing and the promotion of home ownership. In practice, these funds, in conjunction with the Low-Income Housing Tax Credit, are often spent in public-private physical development projects.

[23] According to Hebert et al. (2001) the majority of EZ businesses reported in 2000 that neighborhood conditions were "much improved" or "somewhat improved" since 1997.

[24] In some of the zone cities rents are regulated meaning that housing will be rationed.

expected to bring with them reductions in the fraction of units in a neighborhood that are vacant. However, local landlords may postpone the sale of vacant units to developers if property values are expected to rise faster than the interest rate. Therefore the expected impact of EZ designation on the fraction of units vacant is ambiguous.

III Data

To perform the analysis we constructed a detailed panel dataset combining information from the Decennial Census, the County/City Databook, and HUD. The primary data source utilized is the Neighborhood Change Database (NCDB) which is a panel of census tracts spanning the period 1970-2000 constructed by Geolytics and the Urban Institute. Appendix I provides more detailed information about this dataset and how it was constructed. Tract level Decennial Census information from the NCDB was merged with relevant editions of the County/City Databook to yield a hierarchical longitudinal dataset with four decades worth of information on cities and tracts.[25]

In order to construct a suitable control group for EZs, we obtained 73 of the 78 first round EZ applications submitted to HUD by nominating jurisdictions via a Freedom of Information Act request.[26] These applications contain the tract composition of rejected zones, along with information regarding the number of political stakeholders involved in each proposed zone.[27] We merged this information with data from HUD's web site detailing the tract composition of future zones to create a composite set of rejected and future zones to serve as controls for EZs in our empirical work. Appendix Table 2.A1 details the composition of the cities in our evaluation sample, whether they applied for a Round I EZ, and the treatments (if any) they received.

IV Methodology

A Identification Strategy

The credibility of any non-experimental evaluation hinges critically upon the nature of the treatment assignment mechanism. In order to receive EZ designation, tracts had to pass two stages of selection. First, they had to be nominated by local officials for inclusion in an EZ. Second, the EZ proposal of which they were a part had to be chosen by HUD. While little is known about the initial nomination process, HUD's decision making process has been fairly well documented. EZ applications were ranked and scored according to

[25] Tracts that crossed city boundaries were assigned to the city containing the highest fraction of their population.

[26] The scoring information is not in the public domain and was not released to us by HUD.

[27] Since the applications proposed EZs in terms of 1990 census tracts and the NCDB uses 2000 census tract definitions we use the Census Tract Relationship Files of the U.S. Census Bureau to map the former into the latter.

their ability to meet four criteria: economic opportunity, community-based partnership, sustainable community development, and a strategic vision for change. Explicit eligibility criteria specified minimum rates of poverty and unemployment and maximum population thresholds for groups of proposed census tracts as measured in the 1990 Census.[28] The authorizing legislation also reserved designations for nominees with certain characteristics.[29] Scores were assigned to each application by an interagency review team consisting of approximately 90 individuals. HUD's Department of Community Planning and Development oversaw the review team. After the HUD committee submitted its scores and recommendations the selection decisions were made by HUD Secretary Cisneros in consultation with a 26 member oversight organization known as the Community Empowerment Board. The CEB was chaired by Vice President Gore and staffed by cabinet secretaries and other high ranking officials. After designations were made the CEB was used to coordinate support for EZs and ECs from other agencies.

Following allegations of impropriety in the popular press an investigation was conducted by the HUD inspector general finding some irregularities in the scoring process including that some of the lower ranked EC applications were considered for awards.[30] However, the audit indicated that all six of the first round EZs were chosen from a list of 22 applications designated as "strong" by the HUD selection committee. Wallace (2003) analyzes the assignment process, finding that political variables are poor predictors of EZ designation. Rather, variables such as community participation, size of the empowerment zone, and poverty were the best predictors of receipt of treatment.

We will compare the experience over the 1990s of Round I EZs to tracts in rejected and later round zones with similar historical Census characteristics.[31] Since much of the data used by HUD to select zones came from the 1990 Census it seems reasonable to believe that rejected and future zones with similar census covariates can serve as suitable controls for winning zones. We present a variety of evidence including a series of "false experiments" suggesting that this is indeed the case. Because some of the control zones used in this approach received treatment in the form of ECs, we expect that the resulting estimates of

[28] All zone tracts were required to have poverty rates above twenty percent. Moreover, ninety percent of zone tracts were required to have poverty rates of at least twenty-five percent and fifty percent were required to have poverty rates of at least thirty-five percent. Tract unemployment rates were required to exceed 6.3%. The maximum population allowed within a zone was 200,000 or the greater of 50,000 or ten percent of the population of the most populous city within the nominated area.

[29] For example one urban EZ had to be located in an area where the most populous city contained 500,000 or fewer people. Another EZ was required to be in an area that included two states and had a combined population of 50,000 or less.

[30] See Greer (1995). Secretary Cisneros informed the inspector general's office that "he used the [HUD] staff's general input, as well as his personal knowledge and perspectives on individual community needs, commitment and leadership, in making the final designations and award decisions."

[31] Use of rejected applicants as controls as a means of mitigating selection biases has a long history in the literature on econometric evaluation of employment and training programs. See the monograph by Bell et al. (1995) for a review.

the impact of EZ designation will be biased towards zero, making our estimates relatively conservative.[32]

Since the majority of rejected and future zones are located in different cities than treated zones, we are able to assess the sensitivity of our estimates to geographic spillover effects. This is an important advantage of our work over the Abt study (and many of the studies of state level enterprise zones) which relied entirely upon within city comparisons. Two sorts of local spillovers are plausible. First, some of the "leveraged" outside funds flowing to EZs may have been diverted from other impoverished neighborhoods in the same cities or metropolitan areas. Such reallocations would serve to exaggerate the impact of EZ designation found by a within-city estimator since the control tracts would actually be receiving a negative treatment. Second, any true impact of EZ designation on labor or housing market conditions in EZ neighborhoods may spillover into adjacent neighborhoods. This could bias a within city estimator in the opposite direction, though the expected sign depends upon the outcome in question and the underlying economic parameters governing the process.[33] Without prior information on the size of these two spillover effects, one cannot know which effect will dominate or the composite direction of bias.

Though the use of rejected tracts as controls has many advantages, one may still be concerned that the cities that won first round EZs are fundamentally different from losing cities. A cursory inspection of Table 2.1 indicates that the three largest US cities all won EZs, while the remaining winners are large manufacturing intensive cities. If large cities experienced fundamentally different conditions over the 1990s than small cities, the comparison of observationally equivalent census tracts in winning and losing zones will be biased. To further explore this possibility we construct a set of "placebo zones" in each city receiving an EZ. Each placebo zone contains the same number of census tracts as the actual EZ in that city and possesses similar demographic characteristics. We compare the experience of these placebo zones over the 1990s to that of the rejected and later round zones and find no appreciable differences, bolstering our confidence in the credibility of our findings.

[32] ECs did not receive wage tax benefits but were allocated $3 million in SSBG funds and made eligible for tax exempt bond financing. As mentioned earlier, the bond financing does not appear to have been heavily utilized.

[33] Though one would normally expect improvements in the amenity value of one neighborhood to yield housing price increases in both that neighborhood and adjacent neighborhoods, it is possible, if neighborhoods are gross substitutes, for the prices of adjacent neighborhoods to be negatively correlated. Similarly, it is possible for job growth inside of EZs to occur at the expense of neighborhoods outside of EZs if firms merely relocate between neighborhoods without expanding total employment.

B Econometric Model

Let outcomes in application tract i in city c in decade t be represented by Y_{ict}.[34] Suppose that these outcomes are generated by a model of the form:

$$Y_{ict} = \mu_t\left(D_{ict}, Y_{ict-1}, X_{ict-1}, Z_{ct-1}, \eta_{ct}, \varepsilon_{ict}\right) + \theta_i, \tag{2.1}$$

where $\mu_t\left(.\right)$ is some function indexed by time, D_{ict} is a treatment dummy, Y_{ict-1} is the tract outcome lagged, X_{ict-1} is a vector of predetermined tract characteristics, Z_{ct-1} is a vector of predetermined city wide characteristics, θ_i is a tract fixed effect, η_{ct} is a random city specific year shock, and ε_{ict} is a serially correlated tract specific error term which is assumed to be independent of all other right-hand-side variables.

The class of stochastic processes encompassed by (2.1) is capable of capturing many of the key features one would expect to see in a panel of census tracts. It allows for mean reverting tract and city specific shocks and for conditional correlation of outcomes across tracts within a city and within tracts across time. Moreover, substantial heterogeneity across tracts is permitted, both in their mean outcomes and in their potential responses to EZ designation.

It will be convenient to reexpress the dependence of the function $\mu_t\left(D_{ict}, .\right)$ on EZ designation by writing $\mu_t\left(D_{ict}, .\right) = D_{ict}\mu_t^1\left(.\right) + \left(1 - D_{ict}\right)\mu_t^0\left(.\right)$. The (contemporaneous) effect of EZ designation on outcomes in a given tract may now be defined as $\beta_i = \mu_t^1\left(.\right) - \mu_t^0\left(.\right)$. Note that this effect is a potentially nonlinear function of the predetermined covariates Y_{ict-1}, X_{ict-1}, and Z_{ct-1}. This reflects the notion that neighborhoods with different degrees of pre-existing economic distress are likely to exhibit different responses to EZ designation.

In order to eliminate the tract fixed effect θ_i, let us rewrite (2.1) in first differences using the potential outcomes notation of Neyman (1923) and Rubin (1974):

$$\begin{aligned} \Delta Y_{ict}^1 &= \beta_i + h_t\left(\Omega_{it}, U_{ict}\right), \\ \Delta Y_{ict}^0 &= h_t\left(\Omega_{it}, U_{ict}\right), \end{aligned} \tag{2.2}$$

where $h_t\left(.\right) = \mu_t^0\left(.\right) - \mu_{t-1}^0\left(.\right)$, $\Omega_{it} = \left(Y_{ict-1}, X_{ict-1}, Z_{ct-1}, Y_{ict-2}, X_{ict-2}, Z_{ct-2}\right)$, and $U_{ict} = \left(\eta_{ct}, \varepsilon_{ict}, \eta_{ct-1}, \varepsilon_{ict-1}\right)$. Superscripts index potential outcomes under different treatment states. Because we have only one post-treatment decade in the data we only consider static treatment schemes (i.e. we do not consider potential outcomes associated with two decades of EZ designation or one decade of designation followed by a decade of non-designation). Thus, ΔY_{ict}^1 represents the change in Y_{ict} a tract would have experienced over the 1990s had it been awarded an EZ at the beginning of the decade, while ΔY_{ict}^0 represents the change

[34] From this point on we use the phrase "application tract" interchangeably with "proposed tract" to refer to application and future EZ tracts.

that would have occurred over the 1990s without an EZ. Because we only observe one of these potential outcomes per tract we may write $\Delta Y_{ict} = \Delta Y_{ict}^{1} D_{ict} + \Delta Y_{ict}^{0} (1 - D_{ict})$.

Suppose that application tracts were awarded Empowerment Zone status by HUD based upon the history of their Census covariates available in 1990 and other random factors. We model this selection mechanism as $D_{ict} = 1$ if $D_{ict}^{*} > 0$ and 0 otherwise where[35]

$$D_{ict}^{*} = \lambda \Omega_{it} + \upsilon_{ict}, \tag{2.3}$$

λ is a coefficient vector and υ_{ict} is a random error assumed to be independent of Ω_{it} and U_{ict}—an assumption we display here for future reference:

$$\upsilon_{ict} \perp (\Omega_{it}, U_{ict}). \tag{2.4}$$

In words, this means that conditional on covariates, EZ designation is independent of the experience a proposed census tract would have had over the 1990s in the absence of treatment. This assumption directly implies that the distribution of untreated potential tract outcomes $f\left(\Delta Y_{ict}^{0} | D_{ict}, \Omega_{it}\right)$ is independent of whether or not a tract actually received treatment so that $f\left(\Delta Y_{ict}^{0} | D_{ict}, \Omega_{it}\right) = f\left(\Delta Y_{ict}^{0} | \Omega_{it}\right)$. Rosenbaum and Rubin (1983) term this the Conditional Independence Assumption (CIA) and it forms the cornerstone of our difference-in differences identification strategy. The CIA has the following important implication:

$$E\left[\Delta Y_{ict}^{0} | \Omega_{it}, D_{ict} = 0\right] = E\left[\Delta Y_{ict}^{0} | \Omega_{it}, D_{ict} = 1\right], \tag{2.5}$$

which states that, conditional on covariates, EZ and non-EZ tracts would, on average, be expected to experience the same changes in outcomes during the 1990s in the absence of treatment.

Recall that the tract specific impact of EZ designation β_i is itself a function of the covariates. A standard parameter of interest in the program evaluation literature is the mean effect of treatment on the treated (Heckman and Robb, 1985), which may be defined as:

$$TT = E\left[\Delta Y_{ict}^{1} - \Delta Y_{ict}^{0} | D_{ict} = 1\right] = E\left[\beta_i | D_{ict} = 1\right].$$

As the name suggests, this concept measures the average impact of the program on those who take it up, or in this case, those tracts awarded EZ designation. Since EZ tracts have roughly similar numbers of people, weighting the effect on each tract equally approximates the national impact on EZ residents.

[35] This abstracts from the two step nature of the selection process inherent in EZ assignment. See Appendix II for a justification of the approach taken here.

Estimating TT requires identifying two moments. The first $E\left[\Delta Y_{ict}^{1}|D_{ict}=1\right]$ is trivially identified by the unweighted sample mean of treated observations on ΔY_{ict}. The second moment, $E\left[\Delta Y_{ict}^{0}|D_{ict}=1\right]$, is the counterfactual mean of the treated observations had they not been treated—a quantity with no directly observable sample analogue. We use two approaches to estimating $E\left[\Delta Y_{ict}^{0}|D_{ict}=1\right]$.

The first approach suggested by condition (2.5) is to approximate the function $E[\Delta Y_{ict}^{0}|\Omega_{it}, D_{ict}=0]$ using a parametric model and then to use that model to compute an estimate of $E\left[\Delta Y_{ict}^{0}|D_{ict}=1\right] = \int E\left[\Delta Y_{ict}^{0}|\Omega_{it}, D_{ict}=0\right] dF\left(\Omega_{it}|D_{ict}=1\right)$. We do this by fitting a flexible regression model to the untreated tracts and using the estimated regression coefficients to impute the counterfactual mean outcomes of each treated tract. The average difference between imputed counterfactual outcomes and actual values among treated tracts is then computed as an estimator of TT. This procedure, which can be thought of as a variant of the classic Blinder (1973) and Oaxaca (1973) approach to decomposing wage distributions, can be shown to consistently estimate TT given a sufficiently flexible model for $E\left[\Delta Y_{ict}^{0}|\Omega_{it}\right]$ (see Imbens, Newey, and Ridder, 2007). Thus for each tract we have an estimate of the tract specific treatment effect $\widehat{\beta}_i = \Delta Y_{ict}^{1} - \Delta\widehat{Y}_{ict}^{0}\left(\Omega_{it}\right)$ where $\Delta\widehat{Y}_{ict}^{0}\left(\Omega_{it}\right) = \widehat{E}\left[\Delta Y_{ict}^{0}|\Omega_{it}\right]$ is the prediction from a parametric linear regression function. We then estimate TT using:

$$\widehat{\text{B-O}} = \frac{1}{N_1}\sum_{i\in\{D=1\}}\widehat{\beta}_i.$$

The second approach is to estimate the counterfactual mean $E\left[\Delta Y_{ict}^{0}|D_{ict}=1\right]$ via propensity score reweighting.[36] The basic idea of the propensity score approach is to reweight the data in a manner that balances the distribution of covariates across treated and untreated tracts. This is accomplished by upweighting untreated tracts that "look like" treated tracts based upon their observed variables. Once the distribution of covariates is balanced across treatment and control groups a simple comparison of weighted means will, under the assumptions made thus far, identify TT. Moreover, the performance of the reweighting estimator in balancing the distribution of observed variables across groups can easily be assessed directly by comparing reweighted covariate moments.

A key assumption necessary for propensity score based approaches to identify TT is,

$$P\left(D_{ict}=1|\Omega_{it}\right) < 1. \tag{2.6}$$

This assumption, which is often referred to as the "common support" condition, states that no value of the covariates can deterministically predict receipt of treatment. The failure of

[36] Propensity score reweighting was proposed in the survey statistics literature by Horvitz and Thompson (1952) and adapted to causal inference by Rosenbaum (1987). In the economics literature such estimators have been used in a cross-sectional context by DiNardo et al. (1996) and extended to the panel setting by Abadie (2005). Recent work by Hirano, Imbens, and Ridder (2003) demonstrates that properly implemented reweighting estimators are asymptotically efficient in the class of semiparametric estimators.

this condition would present the possibility that some tracts with particular configurations of covariates would only be capable of being observed in the treated state, thereby preventing the construction of valid controls. As suggested by Heckman et al. (1998b) and Crump et al. (2006) we present results where observations with very high estimated propensity scores are dropped from the sample. This approach safeguards against violations of the overlap condition in finite samples and can substantially reduce the sampling variance of the estimator.[37]

Conditions (2.4) and (2.6) in conjunction with the results of Rosenbaum (1987) imply that[38]

$$E\left[\Delta Y^0_{ict}|D_{ict}=1\right]=E\left[\omega\left(\Omega_{it}\right)\Delta Y^0_{ict}|D_{ict}=0\right], \tag{2.7}$$

where $\omega\left(\Omega_{it}\right)=\frac{p(\Omega_{it})}{1-p(\Omega_{it})}\frac{1-\pi}{\pi}$, $p\left(\Omega_{it}\right)=P\left(D_{ict}=1|\Omega_{it}\right)$, and $\pi=P\left(D_{ict}=1\right)$. Thus the covariate distribution of untreated tracts can be made to mimic that of treated tracts by weighting observations by their conditional odds of treatment $\frac{p(\Omega_{it})}{1-p(\Omega_{it})}$ times the inverse of their unconditional odds $\frac{1-\pi}{\pi}$. Equation (2.7) simplifies estimation considerably since rather than estimating a very high dimensional conditional expectation, for which different tuning parameters might be required for different outcomes, one need only estimate a single propensity score $p\left(\Omega_{it}\right)=P\left(D_{ict}=1|\Omega_{it}\right)$ (Rosenbaum and Rubin, 1983).[39] In practice we estimate $p\left(\Omega_{it}\right)$ via a logit and π by $\frac{N_1}{N_1+N_0}$ the fraction of treated tracts in the estimation sample.

A useful corollary of (2.7) is that:

$$E\left[\omega\left(\Omega_{it}\right)|D_{ict}=0\right]=1. \tag{2.8}$$

Which merely states that the mean weight among the controls should equal one. We impose the sample analogue of this adding up condition when calculating our estimates in order to reflect the theoretical condition in (2.8).[40]

Given estimates $\widehat{p}\left(\Omega_{it}\right)$ and $\widehat{\pi}$ we estimate $E\left[\omega\left(\Omega\right)\Delta Y^0_{ict}|D_{ict}=0\right]$ with its sample

[37] Trimming slightly modifies the estimand to $E\left[\Delta Y^1_{ict}-\Delta Y^0_{ict}|\Delta D_{ict}=1,P\left(\Delta D_{ict}=1|\Omega_{it}\right)<k\right]$ where k is a scalar constant. As suggested by Crump et al. (2006) we choose $k=0.9$ throughout the paper. In most specifications this results in the trimming of a very small fraction (approximately 1%) of the sample.

[38] Proofs of conditions (2.7) and (2.8) are provided in Appendix III.

[39] As pointed out by Heckman et al. (1998a), propensity score approaches do not escape the curse of dimensionality since the function $p\left(\Omega_{it}\right)$ is unknown. The effects on asymptotic bias and variance of adjusting for the propensity score instead of the underlying covariates of which it is a function are ambiguous (see section 7 of that paper).

[40] Equation (2.8) actually provides us with an overidentifying restriction that can be used as a specification test on our model. Very large deviations from 1 of the mean estimated weight among untreated tracts are a sign of misspecification. In Appendix Table 2.A4 we conduct formal tests of this restriction.

analogue

$$\frac{1}{N_0}\sum\frac{\widehat{p}\left(\Omega_{it}\right)}{1-\widehat{p}\left(\Omega_{it}\right)}\frac{1-\widehat{\pi}}{\widehat{\pi}}\Delta Y_{ict}^{0}.$$

We then estimate TT by computing the weighted difference-in-difference (WDD):

$$\widehat{WDD}=\frac{1}{N_1}\sum_{i\in\{D=1\}}\Delta Y_{ict}^{1}-\frac{1}{N_0}\sum_{i\in\{D=0\}}\frac{\widehat{p}\left(\Omega_{it}\right)}{1-\widehat{p}\left(\Omega_{it}\right)}\frac{1-\widehat{\pi}}{\widehat{\pi}}\Delta Y_{ict}^{0}$$

Consistency follows subject to the usual regularity conditions by an appropriate law of large numbers.

Throughout the paper we show results from both the Blinder-Oaxaca (B-O) and reweighting approaches.[41] We prefer the reweighting based estimates on the grounds that they allow us to directly assess the suitability of our specification of the propensity score via visual inspection of covariate balance and simple diagnostics for the logit which are not outcome specific. It is also easier to check whether the overlap condition is satisfied with the reweighting approach than the B-O approach. On the other hand, a strength of the parametric B-O approach is that it can reliably estimate treatment effects even in the absence of overlap if the parametric model upon which it relies is approximately correct.[42]

C Inference Procedures

Confidence intervals and p-values for all estimators are obtained via a pairwise block bootstrapping algorithm described in Appendix IV. This procedure, which is analogous to cluster robust inference, resamples cities rather than tracts in order to preserve the within city dependence in the data. Because we are interested in evaluating the effect of EZ designation on a variety of outcomes, we use a sequential multiple testing procedure suggested by Benjamini and Hochberg (1995) to control the False Discovery Rate (FDR) of our inferences. The False Discovery Rate is defined as the expected fraction of rejections that are false and is closely related to the probability of a type I error. Details of the multiple testing procedure, which is a function of the single hypothesis p-values, are given in Appendix IV. For convenience we also report single hypothesis confidence intervals and p-values. From this point on, we shall refer to outcomes as “significant” at a given level of confidence if the estimated p-value ensures control of the FDR to the specified level. In general, the multiple testing procedure requires substantially lower p-values for a given level of significance than an equivalent single equation test. Failure to reject a single hypothesis in this multiple testing framework is equivalent to a failure to reject the joint

[41] See DiNardo (2002) for a discussion of the reweighting interpretation of Blinder-Oaxaca and Imbens, Newey, and Ridder (2007) for a demonstration of the first order equivalence of the two approaches.

[42] Another advantage implied by the results of Chen, Hong, and Tarozzi (2004) is that the B-O approach, which is a variant of their CEP-GMM estimator, reaches the semiparametric efficiency bound under weaker regularity conditions than propensity score reweighting.

null hypothesis that all of the treatment effects are zero.

V Results

A Characteristics of EZs and Controls

Table 2.3 shows average characteristics of winning and losing proposed zones before and after reweighting.[43] For our baseline specification we restrict the sample to zones in cities with population greater than 100,000. While the residents of rejected and future zones are poor and have high rates of unemployment we see from columns one and four of Table 2.3 that they are not quite as poor or detached from the labor force as residents of EZ areas. After reweighting, however, the mean characteristics of the two groups become substantially more comparable.

Figure 2.2 shows the time series behavior of the EZ and control tracts with and without reweighting. When reweighting methods are applied to the pooled set of controls their history over the past two decades mirrors that of actual Empowerment Zones remarkably well. There is no dip in outcomes prior to EZ designation of the sort found by Ashenfelter (1978) in studying training programs and for some outcomes the time series behavior of the treatment and control groups over the three decades prior to treatment is almost indistinguishable. One can actually see most of our results from these graphs themselves. The key labor market variables (employment, unemployment, and poverty) all seem to have improved in EZ neighborhoods relative to reweighted controls over the 1990s. A few demographic variables such as the fraction of the population with college degrees also appear to have been impacted by the program.

Columns two and three of Table 2.3 indicate that control tracts in treated cities have somewhat different characteristics from those in untreated cities. Moreover, our earlier discussion of spillover effects suggested that the use of controls in treated cities has the potential to confound a differences in differences estimator. Table 2.4 investigates whether pooling control tracts in treated cities with those in rejected cities is likely to introduce important biases into our analysis. This is accomplished by applying our difference in differences estimators to the sample of controls, coding tracts in future EZs in treated cities as the treated group and all other control tracts as untreated. The first column gives the results of a "naive" difference-in-differences analysis without covariate adjustments, the second column presents the results of our preferred reweighted difference-in-differences

[43] The variables included in the reweighting logits are reported in Appendix V. Our baseline specification minimizes the Akaike Information Criteria (see Appendix Table 2.A2). City population could not be included in the conditioning set because it came too close to perfectly predicting EZ receipt. That we cannot mimic the city population distribution of EZs via reweighting should be apparent from the list of winning cities in Table 2.1. To examine whether imbalance in city-wide population affects our DD results we try adding a third order polynomial in 1990 city population to our Blinder-Oaxaca estimator and experiment with a variety of different sample restrictions, each with a different distribution of city size.

estimator, the third column shows the results of the regression based Blinder-Oaxaca estimator, and the fourth column adds a third order polynomial in city population to the Blinder-Oaxaca model.

From the first column of Table 2.4 we see that over the 1990s, control tracts in treated cities experienced smaller increases in the share of residents with college degrees, slightly lower increases in rents, and a greater increase in the fraction of vacant houses than other controls. After conditioning on pre-treatment characteristics all of these relationships disappear. In fact, the magnitude of the differential experience of the two sets of controls over the 1990s tends to be very close to zero, though the reweighting estimator finds a rather large difference in the behavior of mean earnings. This aberrant earnings result disappears in the Blinder-Oaxaca based estimates. We take this as evidence that the two sets of control tracts are roughly exchangeable conditional on predetermined characteristics. In our subsequent analysis we pool together the two sets of controls in order to gain power and to improve the degree of covariate overlap with the EZ tracts.[44]

B Baseline Results

Table 2.5 presents numerical estimates of the impact of EZ designation on EZ neighborhoods. The naive DD estimator finds a large (29.7%) increase in the value of owner occupied housing, a 4 percentage point increase in the fraction of the neighborhood that is employed, a 4.1 percentage point decrease in the fraction of the neighborhood that is unemployed, and a 4.9 percentage point decrease in poverty. Reweighting the DD estimator for covariate imbalance changes the magnitude (though not the sign) of many of the point estimates. The estimated impact on housing values falls to 22.4 percent, while the impact on rents rises dramatically to 7.7% and becomes statistically significant. The reweighting estimator also finds a significant 2.3 percentage point increase in the fraction of residents with a college degree and a 2.6 percentage point decrease in the fraction of residents that are black. The estimated impacts on the labor market variables (employment, unemployment, earnings, and poverty) remain essentially unchanged.

For comparison we also report regression based Blinder-Oaxaca estimates in Column 3. The Blinder-Oaxaca method yields point estimates similar to those found by the reweighting estimator though the statistical precision of the estimates sometimes differs. It finds smaller (though still significant) effects of EZ designation on housing values, rents, poverty, unemployment, and employment. However, the estimated effects on the demographic composition of EZ neighborhoods are small and indistinguishable from zero.

[44] See Appendix Table 2.A5 for baseline results using the rejected tracts only. Dropping control tracts in treated cities reduces the power of the analysis but does not substantially affect the point estimates.

Taken together the WDD and B-O estimates suggest that EZs were effective in increasing the demand for the services of local residents. Employment rates rose, while unemployment and poverty rates fell. Housing markets also seem to have adjusted. Housing values increased as did, to a lesser extent, rents. Though the population of EZ neighborhoods does not appear to have changed substantially, the fraction college educated may have increased by as much as a third over 1990 levels, indicating that some changes in neighborhood composition took place. The magnitude and sign of the estimated impact on percent black is also consistent with this interpretation.

The general similarity between the reweighted and naive DD estimates reinforces our presumption that rejected and future EZ tracts are suitable controls for EZ tracts. To the extent that unadjusted comparisons are inaccurate, they seem to yield biases in the estimated impact on housing market and demographic outcomes. The difference between the reweighted and naive estimates suggest that Empowerment Zones were awarded to areas that would have experienced increases in percent black and decreases in rents and the fraction college educated relative to rejected tracts in the absence of treatment. It is also estimated that EZ housing values would have risen relative to rejected tracts without EZ designation, perhaps because of regional differences in the timing of the housing market boom of the late 1990s.

Column four assesses the importance of leaving city size out of the propensity score (see footnote 43) by adding a third order polynomial in city size to the regression model for the Blinder-Oaxaca specification. This parametrically corrects the estimator for any smooth relationship between changes in the outcomes and city population but substantially reduces the power of the analysis due to collinearity between city population and the other city level covariates.[45] We see from Column 4 that this estimator yields essentially the same results as the original WDD estimator that ignores city size but the estimates are less precise. Appendix Table 2.A5 presents further robustness checks, exploring the sensitivity of the estimates to changes in the sample of cities included in the treatment and control groups, and again finds that the conclusions reached by our preferred WDD estimator are essentially unchanged.

C Tests of the Conditional Independence Assumption

Despite the robustness of the results to modifications of the estimation sample and estimation technique, one may still question the conditional independence assumption (2.4) underlying our identification strategy. If unmeasured factors correlated with the future performance of neighborhoods influenced the process by which zones were awarded the

[45] This collinearity is especially pernicious in our setup as we have only 74 control cities. Our baseline B-O specification includes two lags of four city level covariates. Adding a third order polynomial in 1990 city population yields 11 city level parameters to be estimated from 74 aggregate observations.

treatment our estimates will be biased. To address such concerns, we now perform tests of the assumptions underlying our research design, starting with a series of "false experiments" involving the application of our estimator to samples in which none of the "treated" units received treatment. These experiments may be thought of as tests of the overidentifying restrictions provided by our statistical model.

The first such experiment involves applying our reweighting estimator to outcomes in 1990 before the EZs were assigned. Finding a non-zero "effect" in this time period would be an indication that either our conditioning set is insufficiently rich to characterize the dynamics of sample census tracts in the absence of treatment, or, that there is selection on the 1990 error components η_{c90} and ε_{ic90}.[46] The latter alternative is consistent with the notion that EZs were assigned based upon 1990 census characteristics (which include the innovations η_{c90} and ε_{ic90}) but would require that the 1990 innovation variance be a large fraction of the total cross sectional variance of outcomes over that period, an alternative we consider implausible given the frequency of our data. Thus, we interpret this false experiment as primarily a test of the specification of our conditioning set. Omitting important variables will make treated and untreated units uncomparable in the absence of treatment, yielding spurious estimated "treatment effects" over the 1980's. Table 2.6, however, shows that none of the estimators find any statistically significant effects in 1990 and that most of the point estimates are quite small. The preferred WDD estimator in column three fails to reject any of the hypotheses at even the 10% FDR level. Thus, it seems that the experience of the treated and untreated tracts with similar covariates was nearly identical over the 1980's, lending credence to the notion that they are comparable over the 1990's.

One may, however, feel uncomfortable with the supposition that the 1990s were simply more of the same. Indeed, Glaeser and Shapiro (2003) provide evidence that national trends in the performance of cities over the 1990s differed from those in the previous decade. Returning to our basic model which can be rewritten compactly as,

$$\Delta Y_{ict} = \beta_i D_{ict} + h_t\left(\Omega_{it}, U_{ict}\right), \tag{2.9}$$

one may suspect that city specific trends $\Delta\eta_{ct}$ were correlated with treatment status over the 1990s but not the 1980s, perhaps because HUD officials were able to perceive such trends as they emerged near the inception of the program. Hence, the latent index determining EZ assignment might be better represented by an equation of the form:

[46] As described in Appendix IV, the variables used in the reweighting procedure are from 1970 and 1980, so there is no mechanical reason to expect that the 1990 outcomes would be identical across treatment and control groups.

$$D^*_{ict} = \lambda \Omega_{it} + \rho \Delta \eta_{ct} + v_{ict}. \tag{2.10}$$

In the case where $\rho \neq 0$, the CIA condition is violated and the WDD estimator will not, in general, be consistent.

To test for such a problem we create a series of placebo zones in each treated city and compare their performance over the 1990s to that of future and rejected tracts using the WDD estimator. A finding of nonzero "treatment effects" would indicate a problem with the CIA assumption underlying our analysis. In order to construct the placebo zones we estimated a pooled propensity score model for all tracts in treated cities (see Appendix V for details) and then performed nearest neighbor propensity score matching without replacement in each city, choosing exactly one control tract for each treated EZ tract. This yields a set of placebo zones of the same size and with approximately the same census characteristics as each real EZ.

Figure 2.3 shows the EZ and placebo EZ tracts in Chicago. Tracts shaded black are the actual EZs designated by HUD, while those shaded grey are placebo zones. The placebo tracts tend to be geographically clustered in much the same way as actual EZs, reflecting the underlying spatial correlation of many of the covariates used in the analysis. One potentially troublesome feature of the placebo zones is that they tend to be located near actual EZ tracts. As discussed in Section IV, if EZ designation did in fact have an impact, the effects may have spilled over into adjacent communities. For this reason we also create two additional sets of placebo zones with the restriction that they be outside or inside of a one square mile radius of an EZ tract.

Table 2.7 shows the results of applying the WDD and B-O estimators to each set of placebo tracts.[47] The first column presents results for the pooled set of placebo tracts. None of the outcomes register statistically significant differences across placebo and control zones. Even if one were to ignore the multiple testing procedure, the only outcome close to registering a statistically significant effect is housing rents which despite the large point estimate possesses a single equation 95% confidence interval that includes zero. The second column shows the results of repeating the exercise with placebo tracts less than a mile from an EZ tract. Again, none of the differences are statistically significant. Finally, the third column examines the "impact" of the program on tracts a mile or more away from EZ tracts, yielding nearly identical results. The Blinder-Oaxaca estimates in columns four through six yield the same conclusions.

[47] In order to avoid complications we discard later round zones in the same city as first round EZs from the set of control zones. This results in a modest reduction in the total number of observations used in this part of the analysis.

The general agreement in Table 2.7 between the estimated impacts on closeby and faraway placebo tracts reassures us that any spillover effects that might have accompanied EZ designation are either offsetting or imperceptibly small. Moreover, the general failure to find any significant differences between the treatment and control groups across all three specifications bolsters our confidence in the assumptions underlying our research design.

As a final check on our research design we try converting the outcome variables to scaled within city ranks.[48] If our results are merely picking up city specific shocks then the rank of an average EZ tract in its city wide distribution of poverty rates, for example, should not change over the 1990s relative to the rank of a similar rejected tract in its city-wide distribution. We scale our ranks by the number of tracts in each city so that the transformed outcomes can be thought of as percentiles which are comparable across cities of different absolute size.[49]

Table 2.8 shows the results of applying the WDD and B-O estimators to the transformed outcomes. The point estimates represent the average impact of EZ designation on the percentile rank of EZ neighborhoods. For example, Column 1 indicates that EZ designation led EZ neighborhoods to fall 5.5 percentiles in the within city distribution of tract poverty rates. The results are in close agreement with the findings of Table 2.5, the only substantive difference being that the estimated effect on housing values falls to the point of statistical insignificance. Since housing values also exhibited large (though insignificant) point estimates in the false experiment in Table 2.6, we take this as evidence that the estimated impacts on housing values may not be robust. Column 2 of Table 2.8 shows that the Blinder-Oaxaca estimator with population controls yields point estimates similar to the reweighting estimator though the precision of the estimates is reduced. The remaining columns show that application of the reweighting and Blinder-Oaxaca estimators to the percentile outcomes over the 1980s and in the set of placebo tracts yields very small and statistically insignificant point estimates.

In conclusion, we interpret the results of the exercises considered in this section as demonstrating that the estimates provided in Table 2.5 are unlikely to have been generated by spurious correlation with city wide trends or by misspecification of the multivariate stochastic process generating tract level outcomes.

[48] In a previous version of this paper we experimented with a difference-in-differences-in-differences (DDD) estimator that sought to find within city controls for both actual and rejected EZ tracts. This estimator performed quite poorly severely failing our false experiment tests. This poor performance was caused by difficulties in finding suitable control tracts in rejected cities which are usually quite small. We believe the following percentile rank approach to be a much more transparent and robust approach to making within city comparisons.

[49] In other words, for any outcome Y_{ict} we form a new outcome $P_{ict} = rank_{cy}\left(Y_{ict}\right)/N_c$ where $rank_{cy}$ is the rank of Y_{ict} in the city wide distribution of the variable in that year and N_c is the number of tracts in the relevant city.

D Composition Constant Effects

An obvious concern with our difference in difference results is that some of the estimated labor market effects may be due to compositional changes in the residential population of EZs. Inspection of Table 2.3 indicates that residential mobility is quite high in EZ neighborhoods with only 56% of 1990 residents in the same house as in 1985. Although we have no statistics regarding mobility into and out of the Empowerment Zones, we think it likely that substantial neighborhood churning occurs between decades even if the demographic characteristics of EZ neighborhoods tend to remain relatively stable. For this reason we consider it impossible to determine with available data whether prior residents or new arrivals gained most from the EZ program. What can be done, however, is to assess whether the demographic groups that tended to live in EZs prior to EZ designation benefitted from the program. In this section we use tract level tabulations of labor market outcomes within detailed demographic cells to evaluate whether changes in demographic composition are driving our results. This is done by estimating within cell impacts and then averaging them using 1990 cell frequencies (see Appendix VI for details).

Table 2.9 displays racial composition constant effects on employment, unemployment, and poverty calculated from race specific employment rates. Estimates are calculated by using as the outcome variable the change in each tract's race specific labor market rate weighted by the 1990 racial shares. This adjustment does little to change our earlier conclusions from Table 2.5. Although the point estimates are slightly smaller, we still find substantial and statistically significant effects on employment, unemployment, and poverty. We also find that the fraction of residents with a college degree increased holding racial composition constant, suggesting that much of the estimated influx of the college educated to EZ neighborhoods occurred among blacks.

In order to determine whether the estimated labor market effects are due to changes in the age or educational composition of residents we also examine the impact of EZ designation on the racial composition constant employment rates of 16-19 year old high school graduates and dropouts. Surprisingly, we find very large and statistically significant employment effects on high school dropouts, most of whom, by virtue of our fixed weighting scheme, are black. Similar sized effects are present for high school graduates. We find no effect on students currently enrolled in high school which is unremarkable given that baseline employment rates of such youth are very low. In sum, EZs seem to have resulted in improvements in employment among young people who have either just graduated high school or dropped out – the two groups most likely to be actively seeking work. These youth, especially the dropouts, are unlikely to represent gentrifying families of the sort that one would think could confound interpretation of the previous results.

Our reading of this evidence is that changes in the demographic composition of the neighborhood are unlikely to have generated the large effects on labor market outcomes documented in Tables 5 and 8. This conclusion is broadly consistent with the anecdotal accounts of EZ stakeholders summarized in GAO (2006). The GAO assembled focus groups composed of EZ administrators, state and local officials, and EZ subgrantees and solicited testimonials regarding the impact of EZ designation in each city. The typical response was that EZs positively impacted labor and housing market outcomes, but that some of the observed improvements were the result of neighborhood turnover.

VI Impact Analysis

Our comparison of EZ neighborhoods to rejected and future EZ tracts in other cities strongly suggests that EZ designation substantially affected local labor and housing market conditions. EZs led to increases in local rates of employment on the order of four percentage points and roughly similar sized decreases in unemployment and poverty rates. The price of renting in EZs increased by around seven percent, while the value of owner occupied housing appears to have increased by nearly triple this amount (though the results of our robustness checks cast some doubt upon the validity of the latter estimates).

When compared with baseline employment, unemployment, and poverty rates of thirty six, fourteen, and forty six percent respectively, the estimated labor market impacts of EZ designation are quite substantial. Table 2.10 provides calculations converting the estimated treatment effects from Table 2.5 into effects on totals. The calculations yield an estimated increase in EZ employment of roughly 30,000 individuals, a decrease in unemployment of approximately 13,000 individuals, and a decrease in the poverty headcount of around 50,000 people. It is worth reiterating here that these estimates may well understate the true effect of EZ designation on residential neighborhoods since many of the control zones in our study received some smaller consolation treatment.

Combining the tax credits with the block grants and outside funds, we estimate that the amount of money actually spent in EZ neighborhoods over the course of our sample period is between one and three billion dollars. If we assume that the workers employed because of the EZ program earn the mean annual earnings of EZ residents, and that a third of the employment relationships created will be terminated each year with no effect on future employment probabilities, using a social discount factor of .9, we get a discounted present value of roughly $1.1 billion in extra output.[50]

A different approach is to use the housing market to value the impact of the program. EZ designation is estimated to have increased total annual rents paid by $78 million while the total value of owner-occupied housing is estimated to have increased by $470 million.

[50] The formula used here is $PV = \frac{E}{1-\beta(1-\delta)}$ where E is earnings, $\beta = .95$, and $\delta = 1/3$. The metric used for E is 1990 dollars.

If we use a 10% discount rate to convert the rent flow into an asset value and add it to the increase in total housing value we get a total increase in wealth of $1.2 billion. Even if we discard the estimated impact on housing values, which we have reason to suspect, we still get an estimated increase in wealth of $780 million which is fairly close to our estimates based upon the labor market. While these calculations are clearly flawed measures of the value of EZ designation, we believe they provide a reasonable illustration of the scale of the benefits generated by the program.

A key question raised by the estimates in this paper is why the EZs were able to have such a large impact on the EZ labor market. It is difficult with existing data to disentangle the relative contribution of grants and tax incentives in improving EZ neighborhoods. A lower bound estimate of the number of EZ employees for which firms claimed EZ wage credits can be obtained by dividing the total expenditure on credits in 2000 by the maximum credit of $3,000. This yields 15,000 employees. IRS analysis of 1996 tax return data suggests that this bound is quite loose as over a quarter of corporations claimed total credits less than the maximum for a single employee. If we instead divide the total expenditure by $2,000 we get roughly 23,000 employees claimed by firms. While this latter number is close to the estimated increase in employment, it seems likely that most of the credits were claimed on inframarginal hires or pre-existing workers. In fact, only 45% of firms surveyed by HUD who reported using the wage credits responded that the credits were "important" or "very important" for hiring decisions.[51] Thus we find the notion that the tax incentives are wholly responsible for the observed employment increases to be implausible.

The possibility that block grants and outside funding played an important role in redeveloping EZ neighborhoods is important for understanding the likely effects of the later round EZs and various disaster oriented zones, both of which rely almost entirely upon tax subsidies. The experience of the Round I EZs suggests that government entities may be able to play an important role in coordinating expectations among a wide group of non-profit, public, and private entities interested in investing in disadvantaged neighborhoods. The role of public seed money in leveraging outside investments in local economic development has been understudied.[52] Relatively small grants, in conjunction with sustained political support at the federal level, seem to have been successful in leveraging substantial outside investments in Round I EZ neighborhoods. These investments may have been responsible for stimulating the demand for EZ labor, perhaps through a series of local multiplier effects of the sort contemplated by regional planners (e.g. Treyz et al, 1992) or a form of local increasing returns as considered by Rauch (1993).

[51] Hebert et al. (2001) Exhibit 3-18.

[52] Andreoni (1998) has modeled the role of seed money in determining charitable contributions. To our knowledge the role of seed money in spurring economic development has not been explored in the academic literature.

While it is difficult to directly assess the impact of the non-tax expenditures on the physical and economic environment of EZ neighborhoods, there is some evidence that zone amenities improved over the 1990s. The 1997 wave of the HUD survey found that 45% of zone businesses perceived the neighborhood as an "improved" or "somewhat improved" place to do business since 1994/1995, while the 2000 wave of the survey found that 53% of businesses perceived such improvements since 1997, a statistically significant difference. The most common cited impediment to doing business in zones was crime and public safety in both surveys though concerns over crime seem to have been somewhat less prevalent in 2000. Without equivalent survey data in rejected areas we cannot disentangle these reporting patterns from general trends in the US economy over the 1990s, however, we think it reasonable to suspect that the billions of dollars spent in these neighborhoods might have resulted in substantial improvements to their public safety, physical appearance, and local infrastructure.

VII Conclusion

Our comparison of EZ neighborhoods to rejected and future EZ tracts in other cities strongly suggests that EZ designation substantially improved local labor and housing market conditions in EZ neighborhoods. The implications of these findings for the study of local economic development policies are manifold. First, it appears that the combination of tax credits and grants can be effective at stimulating local labor demand in areas with very low labor force participation rates. That this can occur without large changes in average earnings suggests either that labor force participation in such neighborhoods is very responsive to wages or that job proximity itself affects participation perhaps due to reductions in the cost of learning about vacancies or the cost of commuting to work.[53] Second, in the case of the EZs, the impact of these demand subsidies does not seem to have been captured by the relatively well off; economic development and poverty reduction seem to have accompanied one another in the manner originally hoped for by proponents of the program. Indeed, our use of disaggregate Census tabulations suggests that even young high school dropouts experienced improved labor market prospects as a result of the program. Third, while the treated communities appear to have avoided large scale gentrification over the period examined in this study, policymakers should consider carefully the potential impact of demand side interventions on the local cost of living. Given that the vast majority of EZ residents rent their homes, small changes in the cost of zone living can be expected to impose large burdens on the roughly two thirds of the EZ population who do not work. Tradeoffs of this sort should be taken into account when attempting to determine the incidence of the EZ subsidies. If authorities wish to use EZs as anti-poverty programs they may wish to consider combining housing assistance or incentives for the

[53] This latter alternative is often associated with Kain's (1968) "Spatial Mismatch Hypothesis".

development of mixed income housing as complements to demand side subsidies.

Though our results appear to corroborate the findings of the Abt study, we cannot, with our data, ascertain whether the employment gains of local residents are the result of job growth or the substitution of local workers for outside workers. A detailed analysis of matched employer-employee data might yield insights into whether the scale or substitution effects are responsible for generating the local employment gains observed. More research is also needed to determine whether any job creation that is occurring is due to existing firms expanding, new firms being born, or outside firms relocating.

Finally, this evaluation has only examined the first six years of the EZ program. Very little is known about the dynamics of neighborhood interventions. The decisions of residents, developers, and landlords that lead to neighborhood gentrification and turnover may respond to changes in housing values and rents with a lag. Moreover, as the program comes to a close, firms may move out of zones or close up altogether, reversing any employment gains in the process. Understanding these issues is key to determining the long run winners and losers of EZ designation.

Table 2.1: 1990 Characteristics of First Round Empowerment Zones

City	Total Population	Population Rank	Population in EZ	Poverty Rate in EZ	Unemp. Rate in EZ	EZ Area (sq. miles)
Atlanta	395,337	37	43,792	58	20	8.1
Baltimore	736,014	13	72,725	42	16	7.1
Chicago	2,783,484	3	200,182	49	28	14.3
Cleveland	505,556	23	52,985	47	27	6.3
Detroit	1,027,974	7	106,273	47	28	19.5
Los Angeles	3,512,777	2	234,829	40	19	26.1
New York	7,320,621	1	204,625	42	18	6.3
Philadelphia	1,594,339	5	52,440	50	23	4.3

Source: NCDB and HUD

Table 2.2: Total Spending, by category

	SSBG	Outside Money	Total
Total	$386,105,051	$2,847,510,204	$3,233,615,255
Expenditure by category			
Access to Capital	$82,614,577	$1,483,436,971	$1,566,051,548
Business Assistance	$56,263,375	$481,612,338	$537,875,713
Workforce Development	$48,040,383	$49,081,906	$97,122,289
Social Improvement	$76,367,835	$163,449,118	$239,816,953
Public Safety	$17,625,210	$254,618,150	$272,243,360
Physical Development	$14,266,234	$82,484,595	$96,750,829
Housing	$71,064,126	$325,951,575	$397,015,701
Capacity Improvement	$19,863,311	$6,875,551	$26,738,862
Average annual expenditure			
Access to Capital per firm			$20,881
Business Assistance per firm			$7,172
Workforce Development per unemployed person			$261
Social Improvement per housing unit			$138
Public Safety per person			$56
Physical Development per poor person			$44
Housing per housing unit			$229
Capacity Improvement per EZ			$891,295

Source: HUD PERMS data, Brashares (2000), and Decennial Census

Table 2.3: Sample Characteristics (1990)

	EZ's	Rejected/ Future Zones (outside EZ cities)	Rejected/ Future Zones (inside EZ cities)	Rejected/ Future Zones	Rejected/ Future Zones Reweighted	Unproposed tracts in treated cities
	[1]	[2]	[3]	[4]	[5]	[6]
Mean (census tracts)						
% Black	0.686	0.540	0.717	0.570	0.677	0.298
Employment Rate	0.379	0.466	0.438	0.461	0.380	0.559
Log(pop)	7.747	7.931	8.068	7.954	7.863	7.954
Log(Rent)	5.857	5.838	5.988	5.863	5.907	6.272
Log(House Value)	10.701	10.829	10.654	10.800	10.593	11.760
Log(Mean Earnings)	9.637	9.591	9.684	9.606	9.627	10.013
Poverty Rate	0.460	0.395	0.388	0.393	0.446	0.188
% Vacant Houses	0.147	0.141	0.121	0.138	0.135	0.069
Unemployment Rate	0.231	0.160	0.206	0.167	0.232	0.100
% In same house	0.560	0.494	0.570	0.506	0.555	0.579
% Travel less 20 min	0.473	0.668	0.447	0.632	0.466	0.429
Prop. age 65+	0.312	0.299	0.325	0.304	0.316	0.232
Prop. female-headed HH	0.623	0.555	0.593	0.562	0.631	0.326
Prop. Latino population	0.220	0.176	0.177	0.176	0.246	0.199
Prop. age <18	0.118	0.121	0.103	0.118	0.103	0.127
% College	0.056	0.090	0.060	0.085	0.053	0.196
% High school dropouts	0.311	0.260	0.291	0.265	0.319	0.191
Prop. of HHs with public assistance	0.353	0.241	0.293	0.250	0.362	0.135
Mean (city)						
Avg. across tracts % black	0.438	0.307	0.480	0.335	0.447	0.333
Total crime / population* 100	0.081	0.105	0.093	0.103	0.083	0.081
% College degree	0.175	0.143	0.168	0.147	0.173	0.148
% of workers in city government	0.049	0.047	0.043	0.046	0.047	0.079
Observations (number of census tracts)	257	1364	271	1635	1635	4495

Table 2.4: Balance of Control Samples

Difference-in-Differences Estimates

Model		**Naïve**	**Reweighted**	**Blinder-Oaxaca**	**B.O. City Pop.**
		[1]	**[2]**	**[3]**	**[4]**
Log(pop)	Coeff.	-0.022	-0.027	0.014	-0.057
	CI	[-0.071 , 0.018]	[-0.141 , 0.099]	[-0.083 , 0.113]	[-0.293 , 0.373]
	p-val	*0.283*	*0.559*	*0.753*	*0.574*
% In same house	Coeff.	0.005	-0.005	-0.007	-0.021
	CI	[-0.013 , 0.030]	[-0.073 , 0.046]	[-0.046 , 0.038]	[-0.140 , 0.062]
	p-val	*0.692*	*0.747*	*0.898*	*0.541*
% Black	Coeff.	0.002	-0.016	-0.003	-0.020
	CI	[-0.032 , 0.037]	[-0.065 , 0.017]	[-0.030 , 0.028]	[-0.113 , 0.112]
	p-val	*0.797*	*0.270*	*0.990*	*0.583*
% College	Coeff.	-0.010***	0.008*	0.000	0.009
	CI	[-0.016 , -0.005]	[0.004 , 0.043]	[-0.020 , 0.017]	[-0.092 , 0.067]
	p-val	*0.000*	*0.014*	*0.998*	*0.758*
Employment Rate	Coeff.	-0.010	0.017	0.017	0.038
	CI	[-0.043 , 0.030]	[-0.008 , 0.075]	[-0.007 , 0.052]	[-0.065 , 0.142]
	p-val	*0.695*	*0.118*	*0.152*	*0.227*
Unemployment Rate	Coeff.	-0.011	-0.002	-0.005	-0.004
	CI	[-0.034 , 0.019]	[-0.047 , 0.033]	[-0.038 , 0.020]	[-0.117 , 0.056]
	p-val	*0.427*	*0.886*	*0.635*	*0.829*
Log(Mean Earnings)	Coeff.	0.003	0.111*	0.028	0.064
	CI	[-0.055 , 0.049]	[0.072 , 0.297]	[-0.078 , 0.097]	[-0.204 , 0.243]
	p-val	*0.972*	*0.006*	*0.624*	*0.464*
Poverty Rate	Coeff.	0.016	-0.010	-0.013	-0.064
	CI	[-0.014 , 0.042]	[-0.060 , 0.045]	[-0.056 , 0.026]	[-0.166 , 0.159]
	p-val	*0.331*	*0.533*	*0.406*	*0.325*
Log(House Value)	Coeff.	0.118	0.095	-0.009	0.149
	CI	[-0.023 , 0.309]	[-0.027 , 0.470]	[-0.203 , 0.223]	[-0.727 , 0.522]
	p-val	*0.092*	*0.092*	*0.958*	*0.561*
Log(Rent)	Coeff.	-0.063*	0.029	0.030	0.121
	CI	[-0.110 , -0.009]	[-0.061 , 0.099]	[-0.021 , 0.101]	[-0.100 , 0.380]
	p-val	*0.022*	*0.414*	*0.196*	*0.180*
% Vacant Houses	Coeff.	0.029*	0.015	-0.005	0.009
	CI	[0.006 , 0.056]	[-0.019 , 0.043]	[-0.045 , 0.028]	[-0.098 , 0.090]
	p-val	*0.014*	*0.288*	*0.662*	*0.817*
Number of Tracts		1635	1502	1625	1625
Number of Cities		79	79	79	79

Estimators: All columns show difference-in-difference estimates in which the change in outcomes over the period 1990-2000 among control tracts in cities winning an EZ is compared with the change in outcomes among control tracts in rejected and future zones in other cities. **[1]** *Naïve* refers to difference in difference estimates without covariate adjustments. **[2]** *Reweighted* refers to propensity score reweighted estimates in which the propensity score was calculated using 1990 and 1980 tract and city level characteristics. **[3]** *Blinder-Oaxaca* computes counterfactual means of control tracts in treated cities via regression methods. **[4]** *B.O. City Pop.* is the Blinder-Oaxaca estimator augmented to include a 3rd order polynomial in 1990 city population. (See Sections IV-B, V-A and Appendix V for details).

Inference: 95% C*onfidence intervals (CI)* and *p-values* were obtained via a pairwise block bootstrap that resampled zones in order to preserve the within zone dependence of the data. See Appendix IV for details. *Significance levels*. A multiple testing procedure described in the Appendix was used to control the False Discovery Rate (FDR) to prespecified levels. The procedure yields lower threshold p-values for fixed level tests than in the single equation case. Stars indicate that a hypothesis can be rejected while controlling the FDR to specified levels: * rejected at 10% FDR, ** rejected at 5% FDR and *** rejected at 1% FDR.

Table 2.5: Impact of EZ Designation

Difference-in-Differences Estimates

Model		**Naïve**	**Reweighted**	**Blinder-Oaxaca**	**B.O. City Pop.**
		[1]	**[2]**	**[3]**	**[4]**
Log(pop)	Coeff.	-0.035	0.005	0.024	0.049
	CI	[-0.109 , 0.056]	[-0.085 , 0.105]	[-0.040 , 0.096]	[-0.024 , 0.187]
	p-val	*0.427*	*0.839*	*0.472*	*0.167*
% In same house	Coeff.	-0.009	-0.014	0.001	0.003
	CI	[-0.036 , 0.013]	[-0.052 , 0.006]	[-0.022 , 0.021]	[-0.023 , 0.032]
	p-val	*0.476*	*0.092*	*0.972*	*0.879*
% Black	Coeff.	-0.037	-0.026**	-0.012	-0.020
	CI	[-0.074 , 0.011]	[-0.069 , -0.005]	[-0.036 , 0.014]	[-0.048 , 0.011]
	p-val	*0.131*	*0.026*	*0.346*	*0.164*
% College	Coeff.	0.010	0.023**	0.012	0.014
	CI	[-0.005 , 0.025]	[0.010 , 0.047]	[-0.004 , 0.024]	[-0.006 , 0.027]
	p-val	*0.180*	*0.011*	*0.157*	*0.163*
Employment Rate	Coeff.	0.040**	0.038***	0.020*	0.023
	CI	[0.010 , 0.074]	[0.025 , 0.084]	[0.003 , 0.041]	[0.002 , 0.050]
	p-val	*0.009*	*0.000*	*0.022*	*0.039*
Unemployment Rate	Coeff.	-0.041**	-0.040**	-0.031**	-0.034*
	CI	[-0.072 , -0.013]	[-0.079 , -0.019]	[-0.053 , -0.013]	[-0.057 , -0.012]
	p-val	*0.005*	*0.012*	*0.003*	*0.007*
Log(Mean Earnings)	Coeff.	0.012	0.017	0.028	0.029
	CI	[-0.068 , 0.100]	[-0.050 , 0.114]	[-0.044 , 0.093]	[-0.049 , 0.102]
	p-val	*0.759*	*0.543*	*0.425*	*0.348*
Poverty Rate	Coeff.	-0.049***	-0.050***	-0.038**	-0.044
	CI	[-0.091 , -0.016]	[-0.103 , -0.028]	[-0.058 , -0.013]	[-0.067 , -0.007]
	p-val	*0.000*	*0.000*	*0.008*	*0.030*
Log(House Value)	Coeff.	0.297***	0.224**	0.158	0.183
	CI	[0.093 , 0.538]	[0.078 , 0.506]	[0.006 , 0.332]	[0.005 , 0.367]
	p-val	*0.001*	*0.020*	*0.040*	*0.044*
Log(Rent)	Coeff.	0.005	0.077***	0.044	0.054
	CI	[-0.061 , 0.070]	[0.053 , 0.155]	[-0.001 , 0.099]	[-0.003 , 0.130]
	p-val	*0.896*	*0.001*	*0.056*	*0.061*
% Vacant Houses	Coeff.	0.023	-0.001	0.014	0.006
	CI	[-0.006 , 0.048]	[-0.036 , 0.025]	[-0.007 , 0.037]	[-0.031 , 0.027]
	p-val	*0.128*	*0.681*	*0.177*	*0.790*
Number of Tracts		1892	1869	1892	1892
Number of Cities		82	82	82	82

Estimators: All columns show difference-in-difference estimates in which the change in outcomes over the period 1990-2000 among EZ tracts is compared with the change in outcomes among tracts in rejected and future zones. **[1]** *Naïve* refers to difference in difference estimates without covariate adjustment. **[2]** Reweighted refers to propensity score reweighted estimates in which the propensity score was calculated using 1990 and 1980 tract and city level characteristics. **[3]** Blinder-Oaxaca computes counterfactual means of EZ tracts via regression methods. **[4]** B.O. City Pop. is the Blinder-Oaxaca estimator augmented to include a 3rd order polynomial in 1990 city population. (See Section IV-B and Appendix V for details).

Inference: 95% C*onfidence intervals (CI)* and *p-values* were obtained via a pairwise block bootstrap that resampled zones in order to preserve the within zone dependence of the data. See Appendix IV for details. *Significance levels*. A multiple testing procedure described in the Appendix was used to control the False Discovery Rate (FDR) to prespecified levels. The procedure yields lower threshold p-values for fixed level tests than in the single equation case. Stars indicate that a hypothesis can be rejected while controlling the FDR to specified levels: * rejected at 10% FDR, ** rejected at 5% FDR and *** rejected at 1% FDR.

Table 2.6: False Experiment I (Lagged Model)

Difference-in-Differences Estimates

Model		Naïve [1]	Reweighted [2]	Blinder-Oaxaca [3]	B.O. City Pop. [4]
Log(pop)	Coeff.	-0.055	0.018	0.016	0.020
	CI	[-0.235 , 0.079]	[-0.081 , 0.122]	[-0.069 , 0.126]	[-0.069 , 0.114]
	p-val	*0.589*	*0.621*	*0.597*	*0.533*
% In same house	Coeff.	0.010	0.001	0.012	0.012
	CI	[-0.012 , 0.041]	[-0.033 , 0.029]	[-0.027 , 0.045]	[-0.030 , 0.046]
	p-val	*0.404*	*0.794*	*0.567*	*0.609*
% Black	Coeff.	-0.050	-0.024	-0.018	-0.017
	CI	[-0.109 , 0.036]	[-0.068 , 0.011]	[-0.066 , 0.022]	[-0.061 , 0.010]
	p-val	*0.228*	*0.170*	*0.385*	*0.185*
% College	Coeff.	0.003	0.001	0.004	0.009
	CI	[-0.007 , 0.011]	[-0.012 , 0.009]	[-0.010 , 0.013]	[-0.009 , 0.018]
	p-val	*0.460*	*0.681*	*0.643*	*0.258*
Employment Rate	Coeff.	0.015	-0.016	-0.017	-0.012
	CI	[-0.015 , 0.039]	[-0.058 , -0.001]	[-0.042 , 0.015]	[-0.041 , 0.021]
	p-val	*0.338*	*0.046*	*0.248*	*0.410*
Unemployment Rate	Coeff.	0.010	0.012	0.004	-0.006
	CI	[-0.009 , 0.035]	[0.001 , 0.054]	[-0.020 , 0.035]	[-0.031 , 0.026]
	p-val	*0.314*	*0.045*	*0.633*	*0.758*
Log(Mean Earnings)	Coeff.	0.007	-0.013	0.019	0.037
	CI	[-0.076 , 0.064]	[-0.109 , 0.049]	[-0.068 , 0.085]	[-0.060 , 0.115]
	p-val	*0.836*	*0.578*	*0.687*	*0.364*
Poverty Rate	Coeff.	-0.022	0.034*	0.020	0.010
	CI	[-0.050 , 0.011]	[0.021 , 0.087]	[-0.019 , 0.059]	[-0.028 , 0.050]
	p-val	*0.206*	*0.005*	*0.238*	*0.448*
Log(House Value)	Coeff.	0.091	-0.050	-0.087	-0.100
	CI	[-0.124 , 0.284]	[-0.282 , 0.134]	[-0.292 , 0.170]	[-0.300 , 0.148]
	p-val	*0.400*	*0.515*	*0.442*	*0.391*
Log(Rent)	Coeff.	0.036	-0.041	-0.006	-0.011
	CI	[-0.078 , 0.129]	[-0.139 , 0.005]	[-0.085 , 0.102]	[-0.084 , 0.092]
	p-val	*0.502*	*0.063*	*0.931*	*0.817*
% Vacant Houses	Coeff.	-0.002	0.015	-0.002	-0.004
	CI	[-0.029 , 0.027]	[-0.011 , 0.041]	[-0.032 , 0.017]	[-0.031 , 0.018]
	p-val	*0.992*	*0.247*	*0.710*	*0.593*
Number of Tracts		1891	1882	1891	1891
Number of Cities		82	82	82	82

Estimators: All columns show difference-in-difference estimates in which the change in outcomes over the period 1980-1990 among EZ tracts is compared with the change in outcomes among tracts in rejected and future zones. **[1]** *Naïve* refers to difference in difference estimates without covariate adjustment. **[2]** Reweighted refers to propensity score reweighted estimates in which the propensity score was calculated using 1980 and 1970 tract and city level characteristics. **[3]** Blinder-Oaxaca computes counterfactual means of EZ tracts via regression methods. **[4]** B.O. City Pop. is the Blinder-Oaxaca estimator augmented to include a 3rd order polynomial in 1980 city population. (See Sections IV-B, V-C and Appendix V for details).

Inference: 95% C*onfidence intervals (CI)* and *p-values* were obtained via a pairwise block bootstrap that resampled zones in order to preserve the within zone dependence of the data. See Appendix IV for details. *Significance levels*. A multiple testing procedure described in the Appendix was used to control the False Discovery Rate (FDR) to prespecified levels. The procedure yields lower threshold p-values for fixed level tests than in the single equation case. Stars indicate that a hypothesis can be rejected while controlling the FDR to specified levels: * rejected at 10% FDR, ** rejected at 5% FDR and *** rejected at 1% FDR.

Table 2.7: False Experiment II (Placebo Zones)

Difference-in-Differences Estimates

Model		**Reweighted**			**Blinder-Oaxaca City Pop.**		
Sample		**All**	**Close Tracts**	**Faraway Tracts**	**All**	**Close Tracts**	**Faraway Tracts**
		[1]	[2]	[3]	[4]	[5]	[6]
Log(pop)	Coeff.	0.018	0.046	0.049	0.052	0.071	0.024
	CI	[-0.130 , 0.459]	[-0.087 , 0.481]	[-0.093 , 0.509]	[-0.024 , 0.166]	[-0.004 , 0.203]	[-0.049 , 0.143]
	p-val	*0.894*	*0.523*	*0.565*	*0.149*	*0.062*	*0.415*
% In same house	Coeff.	-0.005	-0.007	-0.002	0.010	0.013	0.006
	CI	[-0.098 , 0.034]	[-0.099 , 0.036]	[-0.093 , 0.037]	[-0.013 , 0.040]	[-0.006 , 0.048]	[-0.022 , 0.039]
	p-val	*0.677*	*0.586*	*0.830*	*0.333*	*0.158*	*0.616*
% Black	Coeff.	0.015	0.016	0.030	-0.010	-0.006	-0.015
	CI	[-0.163 , 0.077]	[-0.159 , 0.083]	[-0.150 , 0.090]	[-0.037 , 0.020]	[-0.035 , 0.028]	[-0.048 , 0.020]
	p-val	*0.639*	*0.633*	*0.441*	*0.367*	*0.612*	*0.343*
% College	Coeff.	0.014	0.011	0.028	0.011	0.008	0.024
	CI	[-0.012 , 0.086]	[-0.010 , 0.079]	[-0.009 , 0.109]	[-0.018 , 0.035]	[-0.018 , 0.033]	[-0.018 , 0.061]
	p-val	*0.255*	*0.245*	*0.142*	*0.472*	*0.561*	*0.275*
Employment Rate	Coeff.	0.002	-0.004	-0.003	0.001	-0.005	0.016
	CI	[-0.023 , 0.061]	[-0.031 , 0.052]	[-0.029 , 0.053]	[-0.026 , 0.033]	[-0.029 , 0.021]	[-0.015 , 0.051]
	p-val	*0.544*	*0.828*	*0.786*	*0.904*	*0.645*	*0.294*
Unemployment Rate	Coeff.	-0.018	-0.005	-0.023	-0.009	-0.001	-0.018
	CI	[-0.082 , 0.118]	[-0.068 , 0.128]	[-0.080 , 0.120]	[-0.028 , 0.012]	[-0.016 , 0.018]	[-0.039 , 0.006]
	p-val	*0.349*	*0.545*	*0.311*	*0.384*	*0.942*	*0.152*
Log(Mean Earnings)	Coeff.	0.005	-0.006	0.010	0.040	0.045	0.019
	CI	[-0.128 , 0.130]	[-0.148 , 0.116]	[-0.127 , 0.153]	[-0.020 , 0.111]	[-0.032 , 0.129]	[-0.051 , 0.106]
	p-val	*0.952*	*0.873*	*0.897*	*0.144*	*0.194*	*0.557*
Poverty Rate	Coeff.	-0.013	0.002	0.002	-0.022	-0.021	-0.022
	CI	[-0.092 , 0.041]	[-0.072 , 0.052]	[-0.071 , 0.050]	[-0.052 , 0.015]	[-0.055 , 0.023]	[-0.054 , 0.016]
	p-val	*0.657*	*0.928*	*0.980*	*0.202*	*0.243*	*0.235*
Log(House Value)	Coeff.	0.058	0.058	0.005	0.152	0.143	0.135
	CI	[-0.160 , 1.556]	[-0.181 , 1.584]	[-0.217 , 1.484]	[-0.008 , 0.309]	[-0.060 , 0.341]	[-0.028 , 0.287]
	p-val	*0.741*	*0.736*	*0.955*	*0.057*	*0.103*	*0.081*
Log(Rent)	Coeff.	0.055	0.047	0.073	0.042	0.032	0.051
	CI	[-0.024 , 0.143]	[-0.024 , 0.125]	[-0.006 , 0.166]	[-0.009 , 0.117]	[-0.020 , 0.112]	[-0.008 , 0.138]
	p-val	*0.114*	*0.148*	*0.066*	*0.107*	*0.193*	*0.092*
% Vacant Houses	Coeff.	0.006	0.009	-0.006	-0.003	-0.002	-0.007
	CI	[-0.109 , 0.038]	[-0.107 , 0.042]	[-0.123 , 0.026]	[-0.043 , 0.017]	[-0.044 , 0.019]	[-0.047 , 0.012]
	p-val	*0.840*	*0.744*	*0.676*	*0.613*	*0.652*	*0.405*
Number of Tracts		1892	1867	1892	1892	1867	1892
Number of Cities		82	82	82	82	82	82

Estimators: All columns show difference-in-difference estimates in which the change in outcomes over the period 1990-2000 among EZ tracts is compared with the change in outcomes among tracts in rejected and future zones. **[1]** *Naïve* refers to difference in difference estimates without covariate adjustment. **[2]** Reweighted refers to propensity score reweighted estimates in which the propensity score was calculated using 1990 and 1980 tract and city level characteristics. **[3]** Blinder-Oaxaca computes counterfactual means of EZ tracts via regression methods. **[4]** B.O. City Pop. is the Blinder-Oaxaca estimator augmented to include a 3rd order polynomial in 1990 city population. (See Sections IV-B, V-C and Appendix V for details) .

Definition of placebo zones. All: tracts inside treated EZ cities but outside the EZ that are a nearest neighbor match for an EZ tract based upon the estimated pscore. **Close/Faraway:** Tracts inside treated EZ cities but outside the EZ and less/more than a mile away from it, that are a nearest neighbor match for an EZ tract based upon the estimated pscore.

Inference: 95% Confidence intervals (CI) and p-values were obtained via a pairwise block bootstrap that resampled zones in order to preserve the within zone dependence of the data. See Appendix IV for details. *Significance levels* . A multiple testing procedure described in the Appendix was used to control the False Discovery Rate (FDR) to prespecified levels. The procedure yields lower threshold pvalues for fixed level tests than in the single equation case. Stars indicate that a hypothesis can be rejected while controlling the FDR to specified levels: * rejected at 10% FDR, ** rejected at 5% FDR and *** rejected at 1% FDR.

Table 2.8: Impact of EZ Designation on Percentile Rank Outcomes

Difference-in-Differences Estimates

		Real Experiment		*False Experiment*		*Placebo Experiment*	
Model		**Reweighted**	**B.O. City Pop.**	**Reweighted**	**B.O. City Pop.**	**Reweighted**	**B.O. City Pop.**
		[1]	[2]	[3]	[4]	[5]	[6]
Log(pop)	Coeff.	-0.001	0.000	0.010	0.014	-0.009	0.010
	CI	[-0.036 , 0.037]	[-0.036 , 0.074]	[-0.032 , 0.051]	[-0.025 , 0.048]	[-0.073 , 0.248]	[-0.021 , 0.074]
	p-val	*0.984*	*0.979*	*0.545*	*0.340*	*0.601*	*0.553*
% In same house	Coeff.	-0.018	0.034	0.012	0.026	-0.012	0.037
	CI	[-0.125 , 0.021]	[-0.021 , 0.085]	[-0.073 , 0.086]	[-0.070 , 0.107]	[-0.293 , 0.074]	[-0.027 , 0.103]
	p-val	*0.144*	*0.182*	*0.984*	*0.593*	*0.594*	*0.193*
% Black	Coeff.	-0.031	-0.031	-0.003	0.020	0.006	-0.016
	CI	[-0.072 , 0.005]	[-0.059 , 0.009]	[-0.024 , 0.026]	[-0.007 , 0.048]	[-0.090 , 0.068]	[-0.045 , 0.018]
	p-val	*0.076*	*0.106*	*0.929*	*0.110*	*0.669*	*0.287*
% College	Coeff.	0.055*	0.034	-0.009	0.031	-0.002	0.012
	CI	[0.003 , 0.115]	[-0.024 , 0.079]	[-0.063 , 0.018]	[-0.014 , 0.062]	[-0.091 , 0.176]	[-0.065 , 0.075]
	p-val	*0.036*	*0.262*	*0.266*	*0.157*	*0.987*	*0.761*
Employment Rate	Coeff.	0.049**	0.028	-0.019	-0.015	-0.013	-0.010
	CI	[0.016 , 0.107]	[-0.016 , 0.095]	[-0.074 , 0.003]	[-0.058 , 0.049]	[-0.059 , 0.117]	[-0.065 , 0.080]
	p-val	*0.005*	*0.216*	*0.066*	*0.607*	*0.707*	*0.655*
Unemployment Rate	Coeff.	-0.060	-0.064	0.006	-0.034	-0.011	-0.023
	CI	[-0.125 , 0.007]	[-0.119 , -0.003]	[-0.019 , 0.076]	[-0.113 , 0.004]	[-0.095 , 0.166]	[-0.080 , 0.033]
	p-val	*0.072*	*0.041*	*0.207*	*0.067*	*0.777*	*0.451*
Log(Mean Earnings)	Coeff.	0.053	0.059	0.019	0.027	0.025	0.033
	CI	[-0.004 , 0.143]	[-0.003 , 0.122]	[-0.029 , 0.105]	[-0.044 , 0.083]	[-0.180 , 0.126]	[-0.016 , 0.103]
	p-val	*0.067*	*0.057*	*0.256*	*0.320*	*0.521*	*0.167*
Poverty Rate	Coeff.	-0.055**	-0.053	0.010	0.000	-0.004	-0.019
	CI	[-0.106 , -0.015]	[-0.079 , -0.019]	[-0.014 , 0.047]	[-0.030 , 0.020]	[-0.189 , 0.072]	[-0.066 , 0.024]
	p-val	*0.010*	*0.014*	*0.271*	*0.799*	*0.969*	*0.418*
Log(House Value)	Coeff.	0.082	0.057	0.002	-0.005	-0.005	0.010
	CI	[-0.009 , 0.184]	[-0.044 , 0.148]	[-0.051 , 0.056]	[-0.091 , 0.077]	[-0.115 , 0.449]	[-0.072 , 0.097]
	p-val	*0.070*	*0.239*	*0.804*	*0.943*	*0.851*	*0.798*
Log(Rent)	Coeff.	0.069**	0.050	-0.011	0.017	0.050	0.040
	CI	[0.038 , 0.142]	[-0.003 , 0.178]	[-0.067 , 0.037]	[-0.024 , 0.074]	[-0.027 , 0.164]	[-0.036 , 0.185]
	p-val	*0.004*	*0.062*	*0.394*	*0.402*	*0.219*	*0.330*
% Vacant Houses	Coeff.	0.004	0.015	-0.040	-0.064	0.027	-0.011
	CI	[-0.140 , 0.092]	[-0.079 , 0.103]	[-0.135 , 0.039]	[-0.182 , 0.041]	[-0.601 , 0.152]	[-0.089 , 0.055]
	p-val	*0.813*	*0.782*	*0.275*	*0.208*	*0.749*	*0.657*
Number of Tracts		1869	1882	1882	1882	1892	1892
Number of Cities		82	82	82	82	82	82

Note. For details regarding estimation and experiments see Sections IV-B, V-C and Appendix V as well as notes to Tables 4-7.**Inference:** 95% Confidence intervals (CI) and p-values were obtained via a pairwise block bootstrap. Stars indicate that a hypothesis can be rejected while controlling the FDR to specified levels: * rejected at 10% FDR, ** rejected at 5% FDR and *** rejected at 1% FDR.

Table 2.9: Composition-Constant Impact of EZ Designation

Difference-in-Differences Estimates

		Real Experiment		*Placebo Experiment*	
Model		**Reweighted**	**B.O. City Pop.**	**Reweighted**	**B.O. City Pop.**
		[1]	[2]	[3]	[4]
Employment Rate	Coeff.	0.033***	0.024	-0.002	0.004
	CI	[0.019 , 0.079]	[0.000 , 0.054]	[-0.031 , 0.061]	[-0.025 , 0.040]
	p-val	0.001	0.053	0.867	0.760
Unemployment Rate	Coeff.	-0.029**	-0.023	-0.014	-0.003
	CI	[-0.066 , -0.003]	[-0.044 , 0.000]	[-0.078 , 0.141]	[-0.020 , 0.016]
	p-val	0.035	0.047	0.442	0.797
Poverty Rate	Coeff.	-0.045***	-0.036	-0.012	-0.022
	CI	[-0.097 , -0.025]	[-0.061 , 0.001]	[-0.109 , 0.043]	[-0.057 , 0.016]
	p-val	0.000	0.055	0.653	0.198
Employment Rate 16-19 in HS	Coeff.	-0.003	-0.008	0.022	0.009
	CI	[-0.059 , 0.061]	[-0.105 , 0.029]	[-0.042 , 0.164]	[-0.069 , 0.055]
	p-val	0.804	0.507	0.277	0.855
Employment Rate 16-19 HS drop.	Coeff.	0.103**	0.101	0.040	0.062
	CI	[0.039 , 0.193]	[-0.072 , 0.188]	[-0.117 , 0.165]	[-0.104 , 0.163]
	p-val	0.008	0.143	0.570	0.337
Employment Rate 16-19 HS grad.	Coeff.	0.129**	0.096	0.032	0.062
	CI	[0.046 , 0.369]	[-0.040 , 0.300]	[-0.132 , 0.360]	[-0.064 , 0.199]
	p-val	0.012	0.141	0.529	0.318
% College	Coeff.	0.012	-0.018	0.010	0.029
	CI	[-0.022 , 0.040]	[-0.083 , 0.084]	[-0.030 , 0.054]	[-0.042 , 0.092]
	p-val	0.443	0.472	*0.621*	*0.306*
Number of Tracts		1869	1869	1892	1892
Number of Cities		82	82	82	82

Note: *Racial-composition-constant outcomes* fix the racial composition of a census tract to that observed in 1990. For details see Sections IV-B, V-D and Appendix V as well as notes to Tables 4-7. *Inference:* 95% Confidence intervals (CI) and p-values were obtained via a pairwise block bootstrap. Stars indicate that a hypothesis can be rejected while controlling the FDR to specified levels: * rejected at 10% FDR, ** rejected at 5% FDR and *** rejected at 1% FDR.

Table 2.10: Impact Calculations

Panel (A)	**Decennial census data** *(inside Empowerment Zones, 1990)*		**Notes**
Population	Total Population	967,851	
	Total Population 16+	718,202	
	Employed	259,271	
	Persons in labor force	330,373	
Earnings	Average annual wage (in 1990 dollars)	16,182	
Panel (B)	**Estimated effects of the Empowerment Zones program**		
Labor Market	Decrease in the unemployed between 1995 and 2000	13,215	0.041 x Persons in labor force
	Decrease in poverty headcount between 1995 and 2000	48,393	0.049 x Total population
	Employment Increase between 1995 and 2000	27,292	0.039 x Population 16+
	Wage value of the number of jobs created	441,634,034	(0.039 x Population 16+) x Ave. Annual wage
	Present discounted value of jobs created	1,104,085,085	Wage value/[1-β(1-d)], with d=1/3 (separation rate) and β=0.9 (social disc. factor)
Housing Market	Average change in annual tractwide rent	319	0.077 * Median annual rent inside EZs
	Total increase in annual rents inside EZ	78,011,523	Average increase in monthly rent x 12 x # of rented houses
	Present Value	780,115,228	Total Annual Rents / 0.1
	Average change in tract-wide owner occupied housing value	9,707	0.224 * Median housing value inside EZs
	Total increase in owner-occupied housing value inside EZ	474,497,043	Average change in tract-wide value x # of owner occupied houses
	Total increase of value of EZ housing units	1,254,612,272	Increase in rent asset value + Increase in value benefiting EZ residents

Note: The coefficients 0.040 (for unemployment rate), 0.050 (for poverty rate), and 0.038 (for employment rate), 0.077 (for rent) and 0.224 (for housing values) were obtained from Column 2 of Table 5.

Table 2.A1: Treatment by city

City	EZ (Round I) (1994)	Application (1994)	Round 1 (1994)	Round 2 (2000)	Round 3 (2002)
Akron		x	EC-1		
Albany		x	EC-1		
Albuquerque		x	EC-1		
Anniston		x			
Atlanta	x	x			
Austin		x			
Baltimore	x	x			
Benton Harbor		x			
Birmingham		x	EC-1		
Boston		x	EEC-1	EZ-2	
Bridgeport		x	EC-1		
Buffalo					RC
Charleston		x			RC
Charlotte		x	EC-1		
Chattanooga					RC
Chester		x			
Chicago	x	x			RC
Cincinnati				EZ-2	
Cleveland	x	x			
Columbia				EZ-2	
Columbus				EZ-2	
Corpus Christi					RC
Cumberland				EZ-2	
Dallas		x	EC-1		
Denver		x	EC-1		
Des Moines		x	EC-1		
Detroit	x	x			RC
El Paso		x	EC-1	EZ-2	
Fairbanks		x			
Flint		x			RC
Fort Lauderdale		x			
Fort Worth		x			
Fresno		x			EZ-3
Gary		x		EZ-2	
Greeley		x			
Hamilton					RC
Harrisburg		x	EC-1		
Hartford		x			
Houston		x	EEC-1		
Huntington				EZ-2	
Indianapolis		x	EC-1		
Jackson		x	EC-1		
Jacksonville		x			EZ-3
Kansas		x	EEC-1		
Knoxville		x		EZ-2	
Lake Charles		x			
Las Vegas		x	EC-1		
Lawrence					RC
Little Rock		x	EC-1		EZ-3
Los Angeles	x	x			RC
Louisville		x	EC-1		
Lowell					RC
Manchester		x	EC-1		
Memphis		x			RC
Miami		x	EC-1	EZ-2	
Milwaukee		x			RC
Minneapolis		x	EC-1	EZ-2	
Mobile		x			RC
Monroe		x			RC
Muskegon		x	EC-1		
Nashville-Davidsc		x	EC-1		
New Haven		x	EC-1	EZ-2	
New Orleans		x			RC
New York	x	x			
Newburgh		x	EC-1		
Niagara Falls					RC
Norfolk		x	EC-1	EZ-2	
Oakland		x	EEC-1		
Ogden		x	EC-1		
Oklahoma		x	EC-1		EZ-3
Omaha		x	EC-1		
Orange		x			
Peoria		x			
Philadelphia	x	x			RC
Phoenix		x	EC-1		
Pine Bluff		x			
Pittsburgh		x	EC-1		
Port Arthur		x			
Portland		x	EC-1		
Providence		x	EC-1		
Richmond		x			
Rochester		x			RC
Sacramento		x			
San Antonio		x	EC-1		EZ-3
San Diego		x			RC
San Francisco					RC
Santa Ana				EZ-2	
Savannah		x			
Schenectady					RC
Seattle		x	EC-1		
Shreveport		x			
Sioux		x			
Springfield		x	EC-1		
St. Louis		x	EC-1	EZ-2	
St. Paul		x	EC-1		
Steubenville		x			
Syracuse					EZ-3
Tacoma		x			RC
Tampa		x	EC-1		
Tucson		x			EZ-3
Waco		x	EC-1		
Washington		x	EC-1		UEnZ
Wilmington		x	EC-1		
Yakima					RC
Yonkers					EZ-3
Youngstown					

Note: EC-1 refers to Enterprise Community awarded in Round I (1994), EEC-1 refers to Enhanced Enterprise Community awarded in Round I (1994), EZ-2 refers to Empowrment Zone awarded in Round II (2000), RC refers to Renewal Community awarded in Round III (2002), EZ-3 refers to Empowrment Zone awarded in Round III (2002) and UEnZ Urban Enterprise Zone awarded in Round III (2002)

Table 2.A2: Logit Model Selection

	BIC		AIC	
Model:	*1 Lag*	*2 Lags*	*1 Lag*	*2 Lags*
Linear models				
(1) Basic	1413	1347	1391	1308
(2) Basic + Other outcomes	1405	1342	1366	1270
(3) Basic + Other outcomes + Other tract-level covariates	1207	1236	1096	1025
(4) Basic + Other outcomes + Other tract-level covariates + City covariates	1182	1135	1049	880
*Linear + Squares models**				
(5) Basic	1413	1347	1391	1308
(6) Basic + Other outcomes	1405	1342	1366	1270
(7) Basic + Other outcomes + Other tract-level covariates	1207	1236	1096	1025

Covariates included in each model:

Basic: Poverty Rate, Log(pop), Unemployment Rate.

Other outcomes: % Black, % Travel less 20 min, Employment Rate, Log(Rent), Log(House Value), Log(Mean Earnings), % Vacant Houses, % in same house, % College degree.

Other tract-level covariates: Prop. female-headed HH, Log family income*, Prop. Latino population, Prop. age <18, Prop. age 65+, % High school dropouts, Log rent, Log house value, Prop. of HHs with public assistance.

City level covariates: % black, Total crime / population* 100, % 65+ years old, % College degree, % of workers in manufacturing, % of workers in city government.

Note: * These models include linear and squared terms of tract level variables.

Table 2.A3: Logits

		Real Experiment							False Experiment				
		Baseline	Big City	All	No NY-LA	No CLE-LA	Balance		Baseline	Big City	All	No NY-LA	No CLE-LA
% Black	[1980]	1.482	1.889	0.275	-0.846	-2.834	-7.95	[1970]	1.184	1.179	0.692	1.311	2.044
		[0.141]	[0.056]*	[0.717]	[0.244]	[0.454]	[0.009]***		[0.086]*	[0.090]*	[0.400]	[0.065]*	[0.019]**
	[1990]	-1.576	-1.856	-1.321	-0.647	1.404	3.832	[1980]	-0.746	-0.716	-1.063	-1.984	-1.345
		[0.210]	[0.116]	[0.231]	[0.529]	[0.711]	[0.184]		[0.413]	[0.438]	[0.216]	[0.019]**	[0.337]
% College	[1980]	-7.112	-7.886	-5.638	-1.957	2.519	23.706	[1970]	-10.398	-10.976	-12.682	-17.509	-15.034
		[0.060]*	[0.025]**	[0.072]*	[0.614]	[0.599]	[0.000]***		[0.006]***	[0.005]***	[0.024]**	[0.001]***	[0.041]**
	[1990]	6.778	7.74	4.09	0.826	11.394	2.523	[1980]	-0.422	-1.103	0.581	1.487	6.916
		[0.078]*	[0.035]**	[0.139]	[0.631]	[0.006]***	[0.551]		[0.871]	[0.654]	[0.840]	[0.595]	[0.074]*
% High school dropouts	[1980]	-1.802	-2.17	1.477	1.25	7.615	9.024	[1970]	2.479	2.009	1.961	-1.265	0.56
		[0.472]	[0.419]	[0.633]	[0.701]	[0.002]***	[0.020]**		[0.612]	[0.680]	[0.723]	[0.831]	[0.866]
	[1990]	5.056	5.105	6.461	6.242	6.84	-3.261	[1980]	-0.119	-0.27	1.847	0.966	7.168
		[0.035]**	[0.022]**	[0.021]**	[0.056]*	[0.006]***	[0.302]		[0.975]	[0.945]	[0.623]	[0.801]	[0.015]**
% Travel less 20 min	[1980]	0.206	-0.154	-0.054	0.846	-2.376	-8.078	[1980]	-2.238	-2.339	-1.542	-1.62	-5.95
		[0.774]	[0.823]	[0.959]	[0.376]	[0.025]**	[0.000]***		[0.024]**	[0.009]***	[0.131]	[0.201]	[0.000]***
	[1990]	-3.171	-3.725	-1.575	-3.411	-9.303	-19.747						
		[0.064]*	[0.023]**	[0.398]	[0.053]*	[0.000]***	[0.000]***						
% Vacant Houses	[1980]	-2.783	-3.286	-1.791	-2.364	-2.722	6.71	[1970]	0.307	0.024	3.722	0.459	1.174
		[0.260]	[0.169]	[0.554]	[0.408]	[0.138]	[0.001]***		[0.861]	[0.990]	[0.093]*	[0.856]	[0.608]
	[1990]	2.577	2.314	4.714	4.456	0.986	-6.374	[1980]	-3.98	-4.365	-3.865	-1.663	-6.137
		[0.098]*	[0.157]	[0.000]***	[0.000]***	[0.671]	[0.004]***		[0.044]**	[0.045]**	[0.093]*	[0.466]	[0.023]**
% in same house	[1980]	-0.883	-0.987	-1.629	-2.122	4.448	4.872	[1970]	0.235	-0.285	0.094	-0.387	-0.119
		[0.430]	[0.335]	[0.134]	[0.027]**	[0.000]***	[0.003]***		[0.863]	[0.823]	[0.957]	[0.857]	[0.936]
	[1990]	2.667	2.246	3.029	2.211	4.59	7.93	[1980]	-0.263	-0.418	-0.241	-0.587	2.679
		[0.069]*	[0.148]	[0.031]**	[0.137]	[0.003]***	[0.007]***		[0.804]	[0.718]	[0.855]	[0.705]	[0.008]***
Employment Rate	[1980]	-3.512	-4.222	-0.832	-4.855	-8.2	2.199	[1970]	-1.044	-0.354	2.38	2.274	3.015
		[0.225]	[0.136]	[0.814]	[0.181]	[0.040]**	[0.438]		[0.695]	[0.897]	[0.467]	[0.541]	[0.427]
	[1990]	2.429	2.624	9.254	9.919	7.347	8.501	[1980]	-2.772	-3.105	2.236	0.789	-5.462
		[0.502]	[0.490]	[0.000]***	[0.000]***	[0.042]**	[0.004]***		[0.513]	[0.464]	[0.460]	[0.798]	[0.225]
Log(Area)	[1980]	-0.247	-0.181	-0.754	-0.678	-0.352	-0.615	[1970]	-0.393	-0.36	-0.825	-0.99	-0.485
		[0.469]	[0.588]	[0.002]***	[0.002]***	[0.187]	[0.069]*		[0.192]	[0.238]	[0.000]***	[0.000]***	[0.104]
Log(House Value)	[1980]	-1.094	-1.036	-1.94	-2.047	-1.386	-1.732	[1980]	-0.677	-0.711	-1.249	-0.732	-1.281
		[0.005]***	[0.020]**	[0.000]***	[0.000]***	[0.058]*	[0.035]**		[0.123]	[0.118]	[0.044]**	[0.148]	[0.069]*
	[1990]	-0.396	-0.309	-0.866	-0.131	-2.022	-2.815						
		[0.387]	[0.459]	[0.111]	[0.713]	[0.000]***	[0.000]***						
Log(Mean Earnings)	[1980]	-0.119	0.028	0.03	-0.596	-0.992	-4.703	[1970]	3.901	3.94	3.264	3.387	7.368
		[0.886]	[0.972]	[0.978]	[0.552]	[0.243]	[0.001]***		[0.003]***	[0.002]***	[0.007]***	[0.007]***	[0.001]***
	[1990]	0.87	0.766	0.898	1.472	1.344	0.582	[1980]	1.125	1.181	1.066	0.627	-0.229
		[0.038]**	[0.081]*	[0.082]*	[0.000]***	[0.046]**	[0.624]		[0.134]	[0.140]	[0.219]	[0.516]	[0.787]
Log(Rent)	[1980]	-0.388	0.017	-0.328	-0.365	-0.385	-5.461	[1980]	0.725	0.979	-0.197	-0.342	1.24
		[0.743]	[0.988]	[0.829]	[0.816]	[0.661]	[0.000]***		[0.451]	[0.333]	[0.833]	[0.693]	[0.241]
	[1990]	-0.525	-0.637	-2.309	-1.447	1.514	3.22	[1970]	-0.737	-0.713	-0.526	-0.402	-1.322
		[0.499]	[0.422]	[0.017]**	[0.124]	[0.192]	[0.008]***		[0.114]	[0.140]	[0.307]	[0.519]	[0.021]**
Log(pop)	[1980]	-1.257	-1.018	-1.205	-1.266	-2.34	-0.014	[1980]	0.366	0.38	0.038	0.006	1.052
		[0.009]***	[0.052]*	[0.002]***	[0.010]***	[0.031]**	[0.991]		[0.403]	[0.395]	[0.944]	[0.992]	[0.006]***
	[1990]	0.987	0.761	0.929	1.07	2.615	1.326						
		[0.062]*	[0.181]	[0.006]***	[0.005]***	[0.033]**	[0.314]						

* significant at 10%; ** significant at 5%; *** significant at 1%. Clustered-robust pvalues in brackets

Table 2.A3: Logits *(Cont.)*

		Real Experiment							False Experiment				
		Baseline	Big City	All	No NY-LA	No CLE-LA	Balance		Baseline	Big City	All	No NY-LA	No CLE-LA
Poverty Rate	[1980]	1.335	1.663	1.878	0.695	4.279	-1.167	[1970]	4.164	3.954	3.145	-0.76	5.606
		[0.409]	[0.323]	[0.311]	[0.703]	[0.215]	[0.769]		[0.009]***	[0.018]**	[0.151]	[0.664]	[0.010]**
	[1990]	4.018	3.96	5.422	6.032	5.849	3.823	[1980]	5.302	5.48	7.44	6.742	5.587
		[0.011]**	[0.014]**	[0.000]***	[0.000]***	[0.044]**	[0.149]		[0.001]***	[0.002]***	[0.000]***	[0.000]***	[0.009]***
Prop. Latino population	[1980]	-3.816	-3.757	-0.45	-0.619	-7.791	-8.159	[1970]	-0.558	-0.82	0.47	0.57	-3.201
		[0.016]**	[0.021]**	[0.867]	[0.843]	[0.017]**	[0.021]**		[0.646]	[0.487]	[0.776]	[0.727]	[0.061]*
	[1990]	6.158	6.535	2.247	3.18	11.393	9.342	[1980]	3.06	3.461	1.966	2.722	6.349
		[0.016]**	[0.013]**	[0.383]	[0.347]	[0.015]**	[0.000]***		[0.046]**	[0.020]**	[0.337]	[0.278]	[0.002]***
Prop. age 65+	[1980]	-3.022	-4.613	-5.906	-3.97	-5.533	-14.262	[1970]	0.296	0.365	0.827	-1.112	-0.07
		[0.350]	[0.111]	[0.037]**	[0.027]**	[0.209]	[0.027]**		[0.903]	[0.878]	[0.732]	[0.652]	[0.980]
	[1990]	-8.383	-7.448	-12.481	-11.73	-12.038	-5.988	[1980]	-10.339	-10.893	-11.898	-10.211	-13.855
		[0.003]***	[0.006]***	[0.001]***	[0.001]***	[0.002]***	[0.064]*		[0.000]***	[0.000]***	[0.000]***	[0.000]***	[0.000]***
Prop. age <18	[1980]	-2.42	-3.184	-1.863	-2.455	-5.487	-7.649	[1970]	-0.984	0.462	2.01	-0.261	6.307
		[0.276]	[0.102]	[0.466]	[0.210]	[0.221]	[0.188]		[0.797]	[0.910]	[0.672]	[0.962]	[0.255]
	[1990]	0.657	1.34	1.874	1.7	3.34	3.574	[1980]	-2.859	-3.027	-1.432	-2.757	-8.337
		[0.736]	[0.477]	[0.408]	[0.391]	[0.337]	[0.497]		[0.263]	[0.239]	[0.605]	[0.289]	[0.022]**
Prop. female-headed HH	[1980]	-2.613	-2.855	-1.56	-1.159	-5.7	-4.285	[1970]	1.662	1.918	4.211	2.978	2.692
		[0.015]**	[0.003]***	[0.140]	[0.314]	[0.000]***	[0.256]		[0.348]	[0.283]	[0.055]*	[0.211]	[0.133]
	[1990]	-0.01	-0.062	0.382	1.049	0.214	0.456	[1980]	-3.121	-3.08	-3.122	-1.46	-5.018
		[0.992]	[0.946]	[0.753]	[0.355]	[0.869]	[0.808]		[0.000]***	[0.000]***	[0.000]***	[0.099]*	[0.000]***
Prop. of HHs with public assistan	[1980]	7.117	6.861	5.561	7.039	7.591	5.829	[1970]	-0.416	0.248	-3.375	1.062	-2.115
		[0.000]***	[0.000]***	[0.004]***	[0.000]***	[0.023]**	[0.353]		[0.894]	[0.935]	[0.438]	[0.778]	[0.543]
	[1990]	1.484	1.825	3.889	2.051	2.736	-0.017	[1980]	5.866	5.693	7.093	7.268	7.617
		[0.199]	[0.123]	[0.027]**	[0.151]	[0.166]	[0.996]		[0.002]***	[0.003]***	[0.000]***	[0.000]***	[0.007]***
Unemployment Rate	[1980]	-3.537	-3.681	-4.502	-1.322	-0.315	15.025	[1970]	-0.129	1.076	-2.872	-6.211	0.289
		[0.231]	[0.220]	[0.315]	[0.689]	[0.908]	[0.000]***		[0.978]	[0.820]	[0.610]	[0.321]	[0.955]
	[1990]	-1.431	-0.888	-1.634	-2.676	3.132	3.962	[1980]	-2.352	-3.026	-0.767	-0.292	2.072
		[0.291]	[0.533]	[0.282]	[0.082]*	[0.195]	[0.372]		[0.388]	[0.303]	[0.785]	[0.913]	[0.490]
% of workers in city government	[1980]	-55.044	-44.041	80.061	-63.682	-62.182	-190.819	[1980]	-20.672	-16.928	-9.072	-2.729	-26
		[0.087]*	[0.068]*	[0.002]***	[0.051]*	[0.039]**	[0.034]**		[0.237]	[0.293]	[0.575]	[0.850]	[0.195]
	[1990]	42.793	34.769	-109.78	34.27	28.817	76.057						
		[0.047]**	[0.011]**	[0.000]***	[0.133]	[0.172]	[0.389]						
% of workers in manufacturing	[1980]	-34.106	-24.077	-19.647	22.921	-47.946	-92.395	[1980]	3.029	2.368	1.699	-9.367	-0.078
		[0.010]**	[0.021]**	[0.667]	[0.524]	[0.000]***	[0.000]***		[0.533]	[0.617]	[0.775]	[0.085]*	[0.990]
	[1990]	63.893	47.931	35.374	-48.203	80.004	142.721						
		[0.001]***	[0.002]***	[0.640]	[0.410]	[0.000]***	[0.000]***						
Avg. across tracts % black	[1980]	20.675	17.082	27.777	4.664	21.784	43.019	[1970]	6.186	4.891	-4.675	-21.697	-0.096
		[0.052]*	[0.084]*	[0.025]**	[0.674]	[0.108]	[0.021]**		[0.581]	[0.645]	[0.672]	[0.014]**	[0.995]
	[1990]	-10.717	-8.047	-4.863	12.45	-9.452	-20.394	[1980]	0.075	1.208	11.978	26.357	4.881
		[0.254]	[0.367]	[0.639]	[0.167]	[0.429]	[0.214]		[0.993]	[0.877]	[0.171]	[0.000]***	[0.645]
Total crime / population* 100	[1980]	21.31	20.382	56.392	67.035	51.116	53.809	[1980]	4.845	-2.868	-11.601	-29.952	-1.479
		[0.193]	[0.080]*	[0.114]	[0.006]***	[0.012]**	[0.033]**		[0.658]	[0.595]	[0.464]	[0.044]**	[0.925]
	[1990]	-25.329	-23.89	-65.141	-72.652	-53.908	-45.003						
		[0.009]***	[0.006]***	[0.001]***	[0.000]***	[0.000]***	[0.027]**						
Constant	[1980]	8.925	6.951	21.014	15.7	15.045	81.345	[1970]	-39.069	-40.327	-28.545	-25.012	-53.686
		[0.380]	[0.484]	[0.119]	[0.178]	[0.141]	[0.000]***		[0.026]**	[0.030]**	[0.084]*	[0.083]*	[0.014]**
Observations		1892	2067	1785	1769	1621	1635		1891	2054	1784	1768	1621

* significant at 10%; ** significant at 5%; *** significant at 1%. Clustered-robust pvalues in brackets

Table 2.A4: Specification Checks

	Baseline	All	No NY or LA	No CLE or LA	Rejected	Balance
	[1]	[2]	[3]	[4]	[5]	[6]
Log (Mean weight) for the untreated	0.014	0.056	-0.020	-0.154	2.291	1.441
CI	[-1.421 , 1.038]	[-3.330 , 1.005]	[-0.604 , 1.449]	[-0.383 , 0.976]	[-4.490 , 7.017]	[-4.686 , 5.131]
p-value	0.670	0.767	0.388	0.764	0.332	0.331
Pseudo R^2	0.476	0.474	0.584	0.569	0.721	0.799
Ratio of treated/untreated tracts s.t. pscore in [0.25-0.50]	0.076	0.076	0.065	0.068	0.055	0.042
Ratio of treated/untreated tracts s.t. pscore in [min(D=0)-0.25]	0.619	0.565	0.566	0.682	0.378	0.607
Ratio of treated/untreated tracts s.t. pscore in [0.50-0.75]	1.717	1.814	1.708	1.519	2.211	1.500
Ratio of treated/untreated tracts s.t. pscore in [0.75-max(D=0)]	3.917	3.385	7.444	5.500	16.300	33.667
% of treated tracts s.t. pscore > max{pscore|D=0}	7.393	7.393	6.218	16.854	4.297	2.214

Columns [1]-[5] show specification tests for the treatment equation estimation estimated via a logit; Column [6] show specification tests for a logit model in which the dependent variable is 1 if the obs. is a control tract in an EZ city and 0 if it is in a rejected/future tract in a non-EZ city.[1] *Baseline* sample refers to a sample of treated and rejected/future zones in EZ cities; [2] *All* refers the complete sample of accepted and rejected/future zones (i.e. no constraint on population or number of tracts); [3] *No NY or LA* presents results on a sample that excludes New York City and Los Angeles. [4] *No Clevaland or LA* presents results on a sample that excludes the SEZs Cleveland and Los Angeles. [5] *Rejected* uses as controls only census tracts outside treated cities. [6] Balance refers to the model estimated on a sample of control tracts.

Note: For an explanation of the mean weight test see Appendix IV. P-values and confidence intervals were obtained by block bootstrap.

Table 2.A5: Impact of EZ Designation (Robustness Checks)

Difference-in-Differences Estimates

		No NY or LA	No Cleveland or LA	Rejected	All
		[1]	[2]	[3]	[4]
Log(pop)	Coeff.	-0.056**	-0.026	-0.071	-0.009
	CI	[-0.207 , -0.024]	[-0.159 , 0.012]	[-0.217 , 0.084]	[-0.096 , 0.073]
	p-val	*0.022*	*0.083*	*0.231*	*0.765*
% In same house	Coeff.	0.002	-0.007	0.002	-0.013*
	CI	[-0.021 , 0.072]	[-0.041 , 0.055]	[-0.060 , 0.061]	[-0.050 , 0.001]
	p-val	*0.541*	*0.773*	*0.960*	*0.055*
% Black	Coeff.	-0.005	-0.024*	-0.037	-0.025*
	CI	[-0.038 , 0.031]	[-0.075 , -0.004]	[-0.124 , 0.012]	[-0.069 , 0.000]
	p-val	*0.994*	*0.030*	*0.103*	*0.046*
% College	Coeff.	0.027**	0.025***	0.024	0.024**
	CI	[0.011 , 0.062]	[0.017 , 0.060]	[-0.018 , 0.067]	[0.009 , 0.047]
	p-val	*0.012*	*0.000*	*0.204*	*0.012*
Employment Rate	Coeff.	0.046***	0.042***	0.036	0.043***
	CI	[0.028 , 0.103]	[0.038 , 0.100]	[-0.003 , 0.092]	[0.030 , 0.090]
	p-val	*0.000*	*0.000*	*0.061*	*0.000*
Unemployment Rate	Coeff.	-0.056***	-0.042***	-0.028	-0.042**
	CI	[-0.106 , -0.042]	[-0.078 , -0.035]	[-0.075 , 0.055]	[-0.078 , -0.019]
	p-val	*0.001*	*0.000*	*0.364*	*0.006*
Log(Mean Earnings)	Coeff.	0.002	0.027	0.017	0.018
	CI	[-0.088 , 0.113]	[-0.058 , 0.182]	[-0.134 , 0.140]	[-0.044 , 0.113]
	p-val	*0.871*	*0.400*	*0.792*	*0.514*
Poverty Rate	Coeff.	-0.055***	-0.064***	-0.054	-0.052***
	CI	[-0.125 , -0.036]	[-0.151 , -0.051]	[-0.121 , -0.016]	[-0.105 , -0.031]
	p-val	*0.000*	*0.000*	*0.014*	*0.000*
Log(House Value)	Coeff.	0.115	0.234	0.117	0.211**
	CI	[-0.288 , 0.287]	[-0.049 , 0.502]	[-0.015 , 0.686]	[0.051 , 0.471]
	p-val	*0.922*	*0.094*	*0.070*	*0.017*
Log(Rent)	Coeff.	0.064**	0.081***	0.069	0.064**
	CI	[0.027 , 0.154]	[0.064 , 0.195]	[-0.035 , 0.168]	[0.029 , 0.129]
	p-val	*0.006*	*0.001*	*0.134*	*0.003*
% Vacant Houses	Coeff.	0.015*	0.006	0.034	-0.002
	CI	[-0.001 , 0.065]	[-0.012 , 0.050]	[-0.020 , 0.085]	[-0.035 , 0.021]
	p-val	*0.059*	*0.248*	*0.163*	*0.579*
Number of Tracts		1742	1736	1480	2042
Number of Cities		80	80	82	104

Estimators: All columns show reweighted difference-in-difference estimates in which the change in outcomes over the period 1990-2000 among tracts in EZs is compared with the change in outcomes among tracts in rejected and future zones. **[1]** *No NY or LA* presents results on a sample that excludes New York City and Los Angeles. [2] *No Cleveland or LA* presents results on a sample that excludes the SEZs Cleveland and Los Angeles. [3] Rejected uses as controls only census tracts nominated for round I EZs that were rejected by HUD. [4] All presents results on the complete sample of accepted and rejected/future zones (i.e. no constraint on population or number of tracts). (See Section IV-B, V-B and Appendix V for details).

Inference: 95% C*onfidence intervals (CI)* and *p-values* were obtained via a pairwise block bootstrap that resampled zones in order to preserve the within zone dependence of the data. See Appendix IV for details. *Significance levels*. A multiple testing procedure described in the Appendix was used to control the False Discovery Rate (FDR) to prespecified levels. The procedure yields lower threshold p-values for fixed level tests than in the single equation case. Stars indicate that a hypothesis can be rejected while controlling the FDR to specified levels: * rejected at 10% FDR, ** rejected at 5% FDR and *** rejected at 1% FDR.

Figure 2.1: Chicago Empowerment Zone

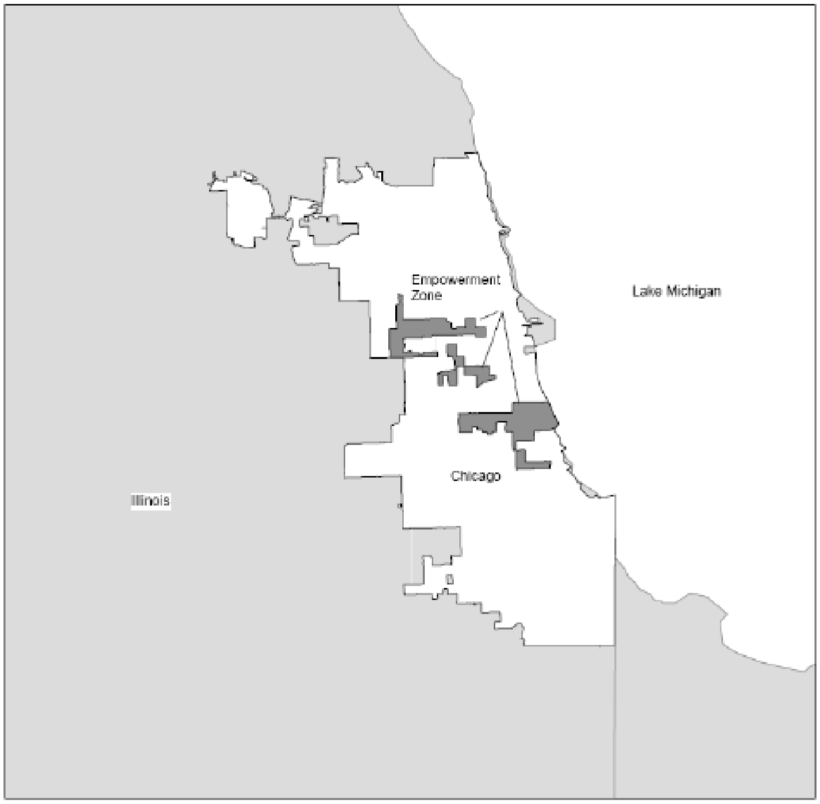
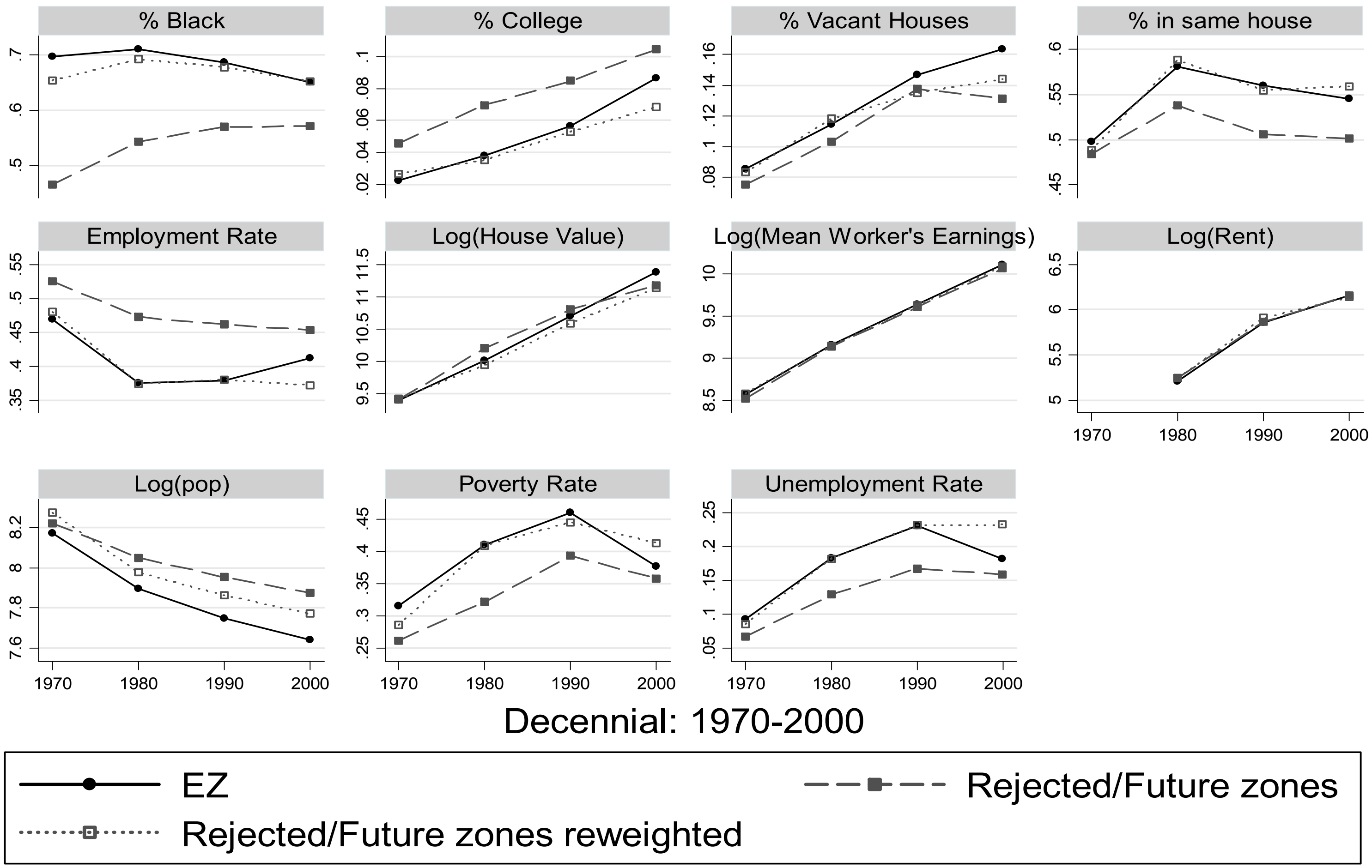
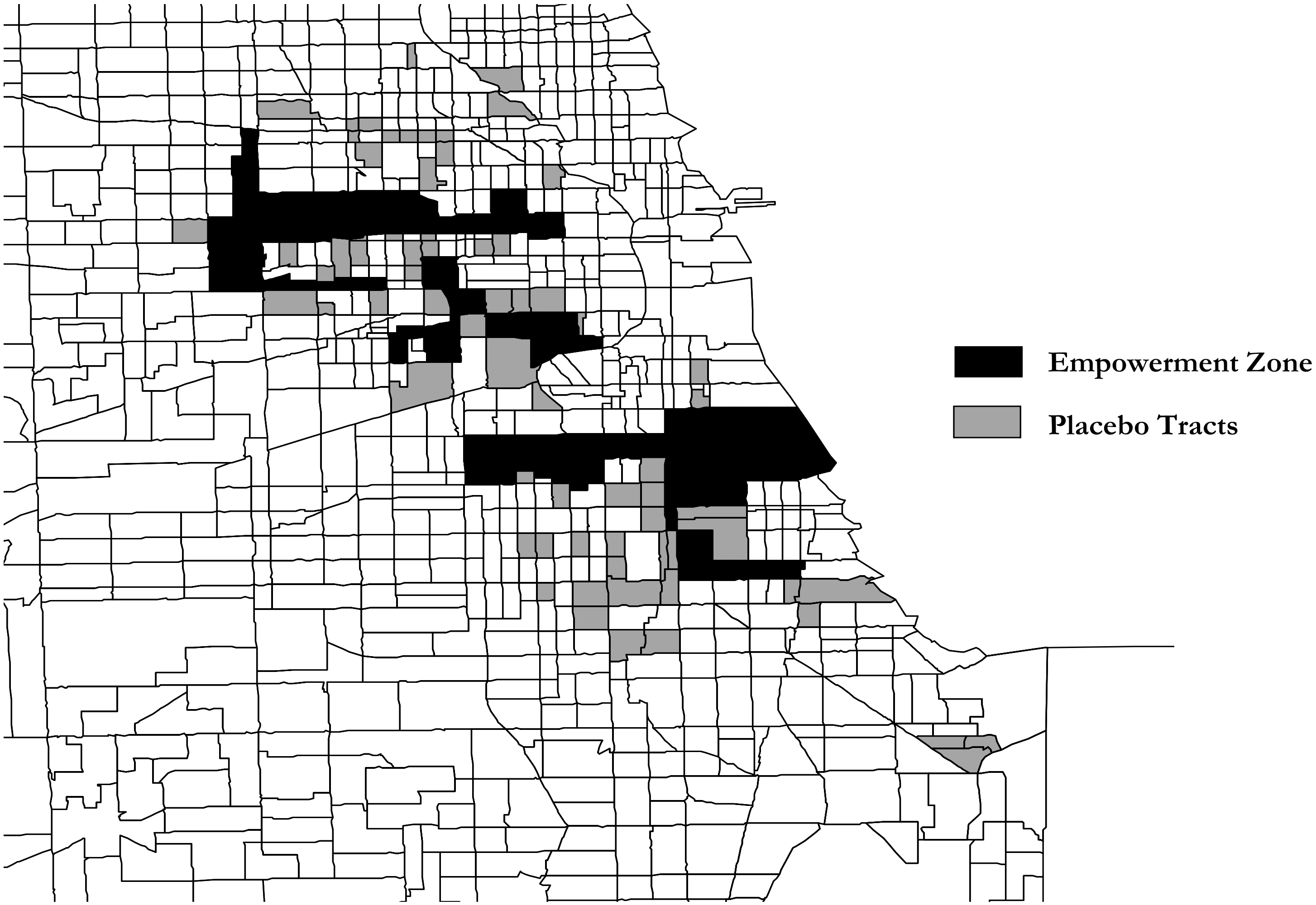


Figure 2.2: Time Series Characteristics



Graphs by outcome

Figure 2.3
Chicago: Empowerment Zone and Placebo Tracts



Appendix I: Data Description and Details

NCDB. The NCDB remaps data from 1970, 1980, and 1990 tracts to 2000 tract boundaries. Coverage in 1970 and 1980 is limited as the US was not entirely divided into tracts at that time, although most areas that were not covered were rural. By 1990 the US was fully divided into census tracts. The remapping process involves mapping tracts in each decade using a GIS program and determining when tract boundaries changed. In the event of a change weights were assigned to tracts from earlier periods based upon population overlap in order to ensure accurate computation of count totals, means, and fractions. Details of the process are given in Appendix J of NCDB Users Guide Provided by Geolytics.

County/City Databook. We extract from the County/City Databook (CCD) variables that are not part of the Decennial Census of population (and therefore are not in the NCDB) such as crime rate, percentage of workers in the manufacturing sector and percentage of workers working in the government. When possible, city level variables were constructed by aggregating the NCDB tract information by city using Geocorr correspondences between tracts and cities. Cross referencing the constructed variables to their analogues in the CCD yielded virtually identical figures.

HUD. We have information on 73 of the 78 applications sent to HUD. We have repeatedly requested the 5 missing applications to no avail. Our dataset also includes all census tracts that belong to any urban EZ, EC, EEC, RC, or UEZ of all the first three rounds. (See Table 2.A1 for more details).

Geocorr. The MABLE/Geocorr engine generates files showing the correspondence between a wide variety of Census and cartographic geographies in the United States. We use Geocorr 2000 to match each census tract to one or more places (cities, townships, villages, etc.). Each census tract that crosses city boundaries was allocated to the city where the majority of the tract's population is located.

Missing Data Some variables used in the estimation procedure exhibited mild missing data problems. Approximately 8.6% of the tracts in our estimation sample had missing mean 1990 housing values and 1.4% had missing mean 1990 rents. Overall we lost approximately 13% of the sample in our baseline specification because of missing values of the control variables. All tables in the paper restrict the estimation sample to the set of tracts (both treated and untreated) with complete covariate information. In results not shown we tried imputing the missing values via sequential regression methods and performed a full case analysis. This procedure yielded very similar results for all outcomes except for housing values which exhibited moderately smaller point estimates.

Appendix II: Alternative Derivation of Propensity Score Model

The assignment model in (2.3) ignores the two step nature of EZ treatment assignment. Here we demonstrate that the hierarchical nature of the assignment process does not present any additional complications to our analysis. Let P_{ic} be an indicator for whether a tract is proposed, W_c an indicator for whether a city wins an EZ, and D_{ic} an indicator for whether a tract gets EZ designation. For a tract to receive EZ designation it must be proposed and its city-wide proposal must be accepted by HUD, so that:

$$D_{ic} = P_{ic} \times W_c.$$

Suppose that tract proposal is a function of covariates Ω_{ic}, unobserved trends ε_{ic}, and a random error ξ_{ic} so that

$$\begin{aligned} P_{ic}^* &= \lambda\Omega_{ic} + \rho\varepsilon_{ic} + \xi_{ic}, \\ P_{ic} &= I\left[P_{ic}^* > 0\right]. \end{aligned}$$

Note that when $\rho \neq 0$ there is selection on unobserved variables in the proposal process. In contrast assume that HUD's decision to award zones is based solely upon the distribution of covariates in a city and random factors independent of the future performance of the proposed neighborhoods, so that

$$\begin{aligned} W_c^* &= T\left(F_c\left(\Omega\right)\right) + \zeta_c, \\ W_c &= I\left[W_c^* > 0\right], \end{aligned}$$

where $F_c\left(\Omega\right)$ is the Empirical Distribution Function (EDF) of covariates in city c, $T\left(.\right)$ is some functional of the EDF, and ζ_c is a random error in the assignment process.

The above equations in conjunction with (2.2) imply that

$$\text{(A.1.)} \qquad \Delta Y_{ict}^1, \Delta Y_{ict}^0 \perp D_{ic} | \Omega_{ic}, P_{ic} = 1.$$

In words, proposed tracts are comparable conditional on their individual level covariates. This follows because $U_{ict} \perp \zeta_c, T\left(F_c\left(\Omega\right)\right)$ for any functional $T\left(.\right)$ – i.e. because conditional on a tract's own covariate levels, its outcomes don't depend on the citywide distribution of covariates or the random assignment error. These are the key assumptions implicit in (2.2). In results not shown, we have tested the assumption that tract outcomes do not depend on the citywide distribution of covariates by including the characteristics of neighboring tracts in regressions and in our reweighting logits. We find virtually identical results. We take this as evidence that cross-tract dependence in the evolution of outcomes is minimal.

By the Rosenbaum & Rubin (1983) theorem $(A.1.)$ implies

$$\Delta Y_{ict}^{1}, \Delta Y_{ict}^{0} \perp D_{ic} | P\left(D_{ic}=1|\Omega_{ic}, P_{ic}=1\right).$$

Now note that

$$\text{(A.2.)} \qquad \begin{aligned} P\left(D_{ict}=1|\Omega_{ic}, P_{ic}=1\right) &= P\left(W_{c}=1|\Omega_{ic}, P_{ic}=1\right) \\ &= E\left[I\left[T\left(F_{c}\left(\Omega\right)\right) < -\zeta_{c}\right]|\Omega_{ic}, P_{ic}=1\right] \\ &= h\left(\Omega_{ic}\right), \end{aligned}$$

where $h\left(.\right)$ is some function. Thus $P\left(D_{ic}=1|\Omega_{ic}, P_{ic}=1\right)$ varies across tracts within a given city. This may seem puzzling given that conditional on being proposed entire cities must either win or lose EZ designation. However, we are not considering $P\left(W_{c}=1|T\left(F_{c}\left(\Omega\right)\right), P_{ic}=1\right)$ but rather $P\left(W_{c}=1|\Omega_{ic}, P_{ic}=1\right)$. The former quantity only varies across cities and is what we are thinking about when we say "the probability of winning." The latter quantity is the probability of a tract being in a winning city given its characteristics and is what the Rosenbaum and Rubin theorem requires we condition on when making inferences. This quantity is consistently estimated via a flexible logit of tract assignment on tract level covariates.

Note also that $(A.2.)$ can be rewritten in a latent variable framework as

$$\begin{aligned} D_{ic}^{*} &= h\left(\Omega_{ic}\right) + \vartheta_{ic}, \\ D_{ic} &= I\left[D_{ic}^{*} > 0\right], \end{aligned}$$

where $\vartheta_{ic} \perp \Omega_{ic}, U_{ic} | P_{ic}=1$ which is equivalent to the expression in (2.3).

Appendix III: Proofs

Proof of (2.7) A similar proof was first derived by Dehejia and Wahba (1997).

$$\begin{aligned}
E\left[\Delta Y_{ict}^{0}|D_{ict}=1\right] &= E\left[E\left[\Delta Y_{ict}^{0}|D_{ict}=1,\Omega_{it}\right]|D_{ict}=1\right] \\
&= E\left[E\left[\Delta Y_{ict}^{0}|D_{ict}=0,\Omega_{it}\right]|D_{ict}=1\right] \\
&= \int E\left[\Delta Y_{ict}^{0}|D_{ict}=0,\Omega_{it}\right] dF\left(\Omega_{it}|D_{ict}=1\right) \\
&= \int E\left[\Delta Y_{ict}^{0}|D_{ict}=0,\Omega_{it}\right] \frac{dF\left(\Omega_{it}|D_{ict}=1\right)}{dF\left(\Omega_{it}|D_{ict}=0\right)} dF\left(\Omega_{it}|D_{ict}=0\right) \\
&= \int E\left[\Delta Y_{ict}^{0}|D_{ict}=0,\Omega_{it}\right] \omega\left(\Omega_{it}\right) dF\left(\Omega_{it}|D_{ict}=0\right) \\
&= E\left[\omega\left(\Omega_{it}\right) E\left[\Delta Y_{ict}^{0}|D_{ict}=0,\Omega_{it}\right]|D_{ict}=0\right] \\
&= E\left[\omega\left(\Omega_{it}\right) \Delta Y_{ict}^{0}|D_{ict}=0\right],
\end{aligned}$$

by Bayes rule,

$$\omega\left(\Omega\right)=\frac{dF\left(\Omega|D_{ict}=1\right)}{dF\left(\Omega|D_{ict}=0\right)}=\frac{P\left(D_{ict}=1|\Omega\right)}{1-P\left(D_{ict}=1|\Omega\right)}\frac{1-P\left(D_{ict}=1\right)}{P\left(D_{ict}=1\right)}=\frac{p\left(\Omega\right)}{1-p\left(\Omega\right)}\frac{1-\pi}{\pi}.$$

Proof of (2.8)

$$E\left[\omega\left(\Omega_{it}\right)|D_{ict}=0\right]=\int \omega\left(\Omega_{it}\right) dF\left(\Omega_{it}|D_{ict}=0\right)=\int dF\left(\Omega_{it}|D_{ict}=1\right)=1.$$

Appendix IV: Inference Procedures

Bootstrap Procedures

We use a nonparametric block bootstrap procedure to assess the sampling variability of the WDD estimator. The steps used are as follows:

1. Sample 8 treated cities and 74 untreated cities with replacement from the original sample.

2. Estimate the propensity score.

3. Compute the statistic of interest T_b^k.

4. Go to step 1 if number of reps is less than 9999, otherwise stop.

We used the empirical bootstrap distribution of T_b^k to calculate single equation p-values and confidence intervals. Asymmetric bootstrap confidence intervals and p-values were constructed using the method described by Davidson and Mackinnon (2004, pp.187-188). P-values and confidence intervals for Naive and OLS models used a studentized bootstrap procedure in order to obtain an asymptotic refinement. None of the tests involving reweighted estimators were studentized.

Benjamini and Hochberg Multiple Testing Procedure

It is well known that conducting multiple tests with a fixed rejection probability does not control the probability of making at least one Type I error across all tests. Standard solutions to the multiple testing problem such as the use of Bonferonni bounds are overly conservative when the tests are correlated or when some of the nulls are false. Benjamini and Hochberg (1995) propose a procedure that controls what they term the False Discovery Rate. Define F as the number of falsely rejected nulls, C as the number of correctly rejected nulls and $R = F + C$ as the total number of rejected hypotheses. The fraction of rejections that are false is a random variable $Q = F/R$. If we define $Q = 0$ in the case where $R = 0$, then the false discovery rate can be written $FDR = E[Q]$. Note that $E[Q] = P(R > 1) E[Q|R > 1]$ and so the FDR can be thought of as the probability of rejecting a null times the expected fraction of rejections that are false given that at least one rejection has occurred. In the case where all nulls are true, the false discovery rate equals the probability of a Type I error (also known as the Family Wide Error Rate) since when all rejections are false $FDR = P(R > 1) = P(F > 1)$. When some nulls are false however, the FDR differs from the probability of making a Type I error. It can be shown that in general $FDR \leq P(F > 1)$. As the fraction of nulls that are false increases, the two concepts diverge and the greater will be the gain in power from controlling the FDR instead of $P(F > 1)$.

From a practical perspective, control of the FDR may better approximate the nature of confidence that researchers desire when attempting to make multiple inferences since the seriousness of a false rejection presumably declines in proportion to the total number of rejections made. Control of the FDR provides an average level of confidence in the decisions made rather than a level of confidence in the entire joint decision. However, control of the FDR also provides a proper test of the joint null that all hypotheses are true, for under such a null, the FDR is equivalent to the Family Wide Error Rate and rejection of a single hypothesis constitutes a rejection of the joint null at the specified level. Failure to reject a single hypothesis in the FDR multiple testing framework constitutes a failure to reject the joint null. Because the FDR approach does not rely upon normality, we have a rather robust replacement for conventional χ^2 tests of joint nulls which are known to have poor finite sample performance.

The Benjamini and Hochberg procedure is conducted by listing the p-values $p_1, p_2, ..., p_m$ of the individual tests in increasing order. The level α test procedure rejects all null hypotheses with $p_i < p_k$ where k is the largest i for which $p_i < \frac{i}{m}\alpha$. For convenience we conduct the procedure at three different levels of α. Benjamini and Hochberg's procedure is robust to arbitrary correlation between the tests and maintains control of the FDR regardless of the fraction of nulls which are false.

Appendix V: Specification of Reweighting and Blinder-Oaxaca Estimators

The covariates used in the study are given in Table 2.A3. For the WDD estimator applied to outcomes over the period 1990-2000 we used a linear logit specification with two lags (i.e. 1990 and 1980 values) of all time varying tract and city level variables. For the WDD false experiment which was computed on outcomes over the period 1980-1990, we used two lags (i.e. 1980 and 1970 values) of all time varying tract and city level variables except the following for which we only included one lag due to the presence of frequent missing values in 1970: log rents, log housing values, % travel less than 20 minutes, citywide % employment in manufacturing, citywide % employment in government, and citywide crime rate.The Blinder-Oaxaca models use the same set of covariates as the reweighting logits but also include squares of all tract level variables and interaction terms between tract level poverty, unemployment, population, and housing values.

To construct placebo zones we performed nearest neighbor matching without replacement on a propensity score estimated on all tracts in the eight cities receiving EZs. The propensity score was estimated on the sample of all tracts in the eight treated cities, using a logit of assignment status on two lags of all time varying tract and city level variables, a set of city dummies, and the interaction of the lags of tract level poverty, unemployment, and population with the city dummies. In calculating, the treatment effect on the placebo zones we replaced the treated tracts by the placebo tracts and proceed to compute $\widehat{\mathrm{OB}}$ and $\widehat{WDD}$ (using the previously estimated weights).

We showed in Appendix III that $E[\omega(\Omega_{it})|D_{ict} = 0] = 1$. This provides us with an overidentifying restriction that can be used as a specification test. Large deviations of the mean estimated weight among untreated observations from one are a sign of model misspecification. Appendix Table 2.A4 presents the log of the mean weight, its confidence interval and a pvalue of the null that the population mean weight equals one. Confidence intervals and pvalues were calculated via the block bootstrap. In all models we fail to reject the null that the sum of weights among the untreated is one at conventional levels of significance. We also present two standard measures of goodness of fit: (1) The pseudo r-squared which is defined as $1 - \frac{\log L_0}{\log L_u}$ where L_0 is the likelihood function restricted to all coefficients being zero and L_u is the unrestricted maximized likelihood and (2) The relative frequency of correct positive predictions of treatment. Finally, to get a sense of how much overlap exists in the propensity score distribution across treatment and controls we show the number of treated tracts per untreated tract by strata of the propensity score.

Appendix VI: Construction of Composition Constant Changes

Let p^u_{rt} be the fraction of individuals in some universe u (e.g. 16-19 year old dropouts) belonging to race r at time t and h^u_{rt} the hazard of individuals in such categories experiencing one of the outcomes in Table 2.9 – e.g. employment, unemployment, poverty, or college education. The mean hazard rates at time t can be written

$$R^u_t = \sum_r p^u_{rt} h^u_{rt}.$$

The "composition constant" rate in 2000 assigns 1990 weights to the 2000 hazards

$$\widetilde{R}^u_{00} = \sum_r p^u_{r90} h^u_{r00}.$$

So that the composition constant change in rates is

$$\begin{aligned} D^u_{00} &= \widetilde{R}_{00} - R_{90} \\ &= \sum_r p^u_{r90} \left(h^u_{r00} - h^u_{r90} \right). \end{aligned}$$

The construction of the composition constant changes was hampered somewhat by the fact that some tracts had no members of a particular racial group in some years preventing estimation of the hazards. This did not present a problem in the case where one of the p^u_{r90} 's was missing for in such cases regardless of what is imputed for h^u_{r90} the entire term will be zero. But when p^u_{r00} was missing and p^u_{r90} was not we faced a nontrivial censoring problem. We solved this problem by imputing missing values of h^u_{r00} using a linear regression of the observed hazards on all of the covariates used in our reweighting logits plus a dummy for being in an EZ. Imputations were constructed as the sum of the predicted values from the imputation regression plus a draw from a normal distribution with standard deviation equal to the residual mean squared error of the imputation regression. With these imputed hazards we proceeded to compute values of D_{00} for all tracts capable of inclusion in the universe (e.g. all tracts having 16-19 year old dropouts).

Bibliography

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 2005, *72* (1),1-19.

Andreoni, James, "Towards a Theory of Charitable Fund Raising," *Journal of Political Economy*, 1998, *106* (6), 1186-1213.

Ashenfelter, Orley, "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 1978, *60* (1), 47-57.

Bartik, Timothy J., *Who Benefits From State and Local Economic Development Policies?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. 1991.

Bartik, Timothy J., "Evaluating the Impacts of Local Economic Development Policies on Local Economic Outcomes: What Has Been Done and What is Doable?," Upjohn Institute Staff Working Paper #03-89. 2002.

Bell, Stephen, Larry Orr, John Blomquist, and Glenn Cain, *Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test.* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. 1995.

Benjamini, Yoav and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society Series B (Methodological)*, 1995, *57* (1), 289-300.

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

Boarnet, Marlon G. and William T. Bogart, "Enterprise Zones and Employment: Evidence from New Jersey," *Journal of Urban Economics*, 1996, *40* (2), 198-215.

Bondonio, Daniele and John Engberg, "Enterprise Zones and Local Employment: Evidence from the States' Programs," *Regional Science & Urban Economics*, 2000, *30* (5), 519-549.

Bondonio, Daniele, "Do Tax Incentives Affect Local Economic Growth? What Mean Impacts Miss in the Analysis of Enterprise Zone Policies," Center for Economic Studies Working Paper 03-17. 2003.

Brashares, Edith,"Empowerment Zone Tax Incentive Use: What the 1996 Data Indicate," *Statistics of Income Bulletin*, 2000, Vol. 3, 236-252.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi, "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data, and Treatment Effects," Mimeo. 2004.

Chouteau, Dale L., "HUD's Oversight of the Empowerment Zone Program, Office of Community Planning and Development, Multi-Location Review," Department of Housing and Urban Development, Office of Inspector General. Audit Case # 99-CH-156-0001. 1999.

Crump, Richard, Joseph Hotz, Guido Imbens, and Oscar Mitnik, "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand," Unpublished manuscript. 2006.

Dehejia, Rajeev H. and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," in *"Econometric Methods for Program Evaluation,"* Cambridge: Rajeev H. Dehejia, Ph.D. Dissertation, Harvard University, 1997, chapter 1.

Davidson, Russell and James Mackinnon.*Econometric Theory and Methods.* Oxford: Oxford University Press. 2004.

Department of Housing and Urban Development, "Introduction to the RC/EZ Initiative," 2003. Accessed online at: http://www.hud.gov/offices/cpd /economicdevelopment/programs/rc/about/ezecinit.cfm

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

Dinardo, John, "Propensity Score Reweighting and Changes in Wage Distributions," Mimeo. 2002.

Elvery, Joel, "The Impact of Enterprise Zones on Residents' Employment: An Evaluation of the Enterprise Zone Programs of California and Florida," Unpublished manuscript. 2003.

Engberg, John and Robert Greenbaum, "State Enterprise Zones and Local Housing Markets," *Journal of Housing Research*, 1999, *10* (2), 163-187.

General Accounting Office, "Businesses' Use of Empowerment Zone Tax Incentives," Report # RCED-99-253. 1999.

General Accounting Office, “Community Development: Federal Revitalization Programs Are Being Implemented, but Data on the Use of Tax Programs Are Limited,” Report # 04-306. 2004.

General Accounting Office, “Empowerment Zone and Enterprise Community Program: Improvements Occurred in Communities, But The Effect of The Program Is Unclear,” Report # 06-727. 2006.

Glaeser, Edward and Jesse Shapiro, “Urban Growth in the 1990s: Is City Living Back?,” *Journal of Regional Science*, 2003, emph43 (1), 139-165.

Greer, Chris, “Audit of Empowerment Zone, Enterprise Community and Economic Development Initiative Grant Selection Processes,” Office of Inspector General, Audit Case No. 95-HQ-154-0002. 1995.

Hebert, S., A. Vidal, G. Mills, F. James, and D. Gruenstein, “Interim Assessment of the Empowerment Zones and Enterprise Communities (EZ/EC) Program: A Progress Report,” Office of Policy Development and Research. 2001. Available online at: www.huduser.org/ Publications/pdf/ezec_report.pdf

Heckman, James J. and Richard Robb, “Alternative Methods for Evaluating the Impact of Interventions,” In *Longitudinal Analysis of Labor Market Data,* ed. James Heckman and Burton Singer, 156-246, Cambridge: Cambridge University Press. 1986.

Heckman, James J., Hidehiko Ichimura, and Petra Todd, “Matching as an Econometric Evaluation Estimator,” *Review of Economic Studies*, 1998a, *65* (2), 261-294.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra Todd, “Characterizing Selection Bias Using Experimental Data,” *Econometrica*, 1998b, *66* (5), 1017-1098.

Heckman, James J. and Jeffrey A. Smith, “The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies,” *Economic Journal*, 1999, *109* (457), 313-348.

Heeringa, Steven G. and John S. Haeussler, “The Small Business Benefits Survey: A Survey Design to Study Small Business Employee Benefits in Local Labor Markets”, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor. Unpublished manuscript. 1993.

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

Horvitz, D.G. and D.J. Thompson, "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 1952, *47* (260), 663-685.

Imbens, Guido, Whitney Newey, and Geert Ridder, Mean Squared Error Calculations for Average Treatment Effects," Mimeo. 2007.

Internal Revenue Service, "Tax Incentives for Distressed Communities," Publication 954 Cat. No. 20086A. 2004.

Kain, John, "Housing Segregation, Negro Employment, and Metropolitan Decentralization," *Quarterly Journal of Economics*, 1968, *82* (2), 175-197.

Neyman, Jerzy, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," As reprinted in *Statistical Science*, 1923, *5* (4), 465-480.

Neyman, Jerzy and Elizabeth Scott, "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 1948, *16* (1), 1-32.

Nichols, Albert and Richard Zeckhauser, "Targeting Transfers Through Restrictions on Recipients," *American Economic Review*, 1982, *72* (2), 372-377.

Nolan, Alistair, and Ging Wong, *Evaluating Local Economic and Employment Development: How to Assess What Works Among Programmes and Policies.* Paris: Organization for Economic Cooperation and Development. 2004.

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

Papke, Leslie, "What Do We Know About Enterprise Zones?," NBER Working Paper #4251. 1993.

Papke, Leslie, "Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program," *Journal of Public Economics*, 1994, *54* (1), 37-49.

Peters, Alan H. and Peter S. Fisher, *State Enterprise Zone Programs: Have They Worked?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. 2002.

Rauch, James. E., "Does History Matter Only When It Matters Little? The Case of City Industry Location," *Quarterly Journal of Economics*, 1993, *108* (3), 843-867.

Rosenbaum, Paul, "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 1987, *82* (398), 387-394.

Rosenbaum, Paul and Donald Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 1983, *70* (1), 41-45.

Rubin, Donald, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 1974, *66* (5), 688-701.

Treyz, George, Dan Rickman, and Gang Shao, "The REMI Economic-Demographic Forecasting and Simulation Model," *International Regional Science Review*, 1992, *14* (3), 221-253.

Wallace, Marc, "An Analysis of Presidential Preferences in the Distribution of Empowerment Zones and Enterprise Communities," *Public Administration Review*, 2003, *63* (5), 562-572.

Wolfe, Heath, "HUD's Oversight of Empowerment Zone Program: Office of Community Planning and Development Multi-Location Review," Department of Housing and Urban Development, Office of Inspector General. Audit Case # 2003-CH-0001. 2003.

Chapter 3

Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects [1]

This paper explores the finite sample properties of semiparametric estimators of average treatment effects. Such estimators are standard in the program evaluation literature and have become increasingly popular in the applied microeconometric literature. These estimators rely on two assumptions. The first assumption is that assignment to treatment is randomized conditional on a set of observed covariates. The second assumption is more technical and asserts that no value of the observed covariates assures treatment assignment.[2] Intuitively, these assumptions allow for treatment to covary with observed characteristics, but require that there be some unexplained variation in treatment assignment left over after conditioning and that the unexplained aspect of treatment resembles an experiment.[3]

Estimation of program impacts under these assumptions could proceed using traditional parametric estimation methods such as maximum likelihood. However, an early result of Rosenbaum and Rubin (1983) is that if treatment is randomized conditionally on the observed covariates, then it is randomized conditional on the (scalar) propensity score, the conditional probability of treatment given the observed covariates. Influenced by this result, the subsequent econometric and statistical literatures have focused on semiparametric estimators that eschew parametric assumptions on the relationship between the outcome and observed covariates. Empirical literatures, particularly in economics, but also in medicine, sociology and other disciplines, feature an extraordinary number of program impact estimates based on such semiparametric estimators.

[1] This paper was written with John DiNardo and Justin McCrary.

For comments that improved the paper, we thank Alberto Abadie, Matias Cattaneo, Marie Davidian, Keisuke Hirano, Guido Imbens, Jack Porter and Jeff Smith. We would also like to thank Markus Frölich for providing us copies of the code used to generate the results in his paper.

[2] Selection on observed variables is defined in Section I, as is the second assumption, which is typically referred to as an overlap assumption. In Section II, we emphasize the correct interpretation of selection on observed variables using a specific parametric model.

[3] In other words, there exists an instrument which is unobserved by the researcher.

Perhaps surprisingly in light of their ubiquity in empirical work, formal large sample results for these estimators have only recently been derived in the literature. Heckman, Ichimura and Todd (1997) report large sample properties of estimators based on kernel and local linear matching on the true and an estimated propensity score. Hirano, Imbens and Ridder (2003) report large sample properties of a reweighting estimator that uses a nonparametric estimate of the propensity score. This is essentially the same reweighting estimator that was introduced to the economics literature by DiNardo, Fortin and Lemieux (1996) and Dehejia and Wahba (1997), and it is related to an estimator due to Horvitz and Thompson (1952). Importantly, Hirano et al. (2003) establish that their version of a reweighting estimator achieves the semiparametric efficiency bound (SEB) established by Hahn (1998) for this problem. Robins, Rotnitzky and Zhao (1994) and Robins and Rotnitzky (1995) establish large sample properties and the efficiency of a regression-adjusted reweighting estimator that uses the estimated propensity score. Finally, Abadie and Imbens (2006) establish the large sample properties and near-efficiency of kth nearest-neighbor matching using the true propensity score.[4]

To date, no formal finite sample properties have been established for any of the estimators discussed, and there is limited simulation evidence on their performance. It is generally desirable to learn about the finite sample properties of estimators used in empirical research, since not all data sets are big enough for asymptotic theory to be a useful guide to estimator properties. It is particularly desirable to learn about the finite sample properties of semiparametric estimators of average treatment effects, given the literature's substantive focus on treatment effect heterogeneity.[5] In the face of heterogeneity, treatment effects must effectively be estimated for various subsamples.[6] For many economic data sets, these subsamples are modest in size, perhaps numbering in the hundreds or even dozens, where asymptotic theory may be a particularly poor guide to finite sample performance.

[4] It deserves mention that Chen, Hong and Tarozzi (2008) study the large sample properties and efficiency of sieve estimators in this setting. We do not study the finite sample properties of these estimators due to space constraints.

[5] Understanding the sources of treatment effect heterogeneity is critical if the analyst hopes to extrapolate from the findings of a given study to broader forecasts of the likely impacts of policies not yet implemented. These issues are a key focus of the program evaluation literature (see, for example, Heckman and Vytlacil 2005 and Heckman, Urzua and Vytlacil 2006).

[6] Importantly, the intrinsic dimensionality of treatment effect heterogeneity cannot be massaged by appealing to the dimension reduction of the propensity score. The Rosenbaum and Rubin (1983) result that a conditionally randomized treatment is randomized conditional on the scalar propensity score has been interpreted as justification for matching on the propensity score rather than on the full set of covariates. However, the Rosenbaum and Rubin result does not imply that units with the same value of the propensity score have the same treatment effect. Examples of empirical investigation of treatment effect heterogeneity along dimensions different from the propensity score include Card (1996), Katz, Kling and Liebman (2001), Haviland and Nagin (2005) and Kent and Hayward (2008), for example.

In this paper, we examine the relative performance of several leading semiparametric estimators of average treatment effects in samples of size 100 and 500.[7] We focus on the performance of propensity score reweighting and matching estimators for estimating the average treatment effect (ATE) and the average effect of treatment on the treated (TOT). We consider a range of matching strategies, including nearest neighbor, kernel, local linear, and ridge matching, and blocking. We also consider several varieties of reweighting estimators, the so-called double robust estimator (Robins et al. 1994), and a specific version of Hahn's (1998) general estimator, which we term a control function estimator. We consider settings with good overlap in the distribution of propensity scores for treatment and control units, as well as settings of poor overlap. In settings of poor overlap, we investigate the performance of various trimming methods proposed and used in the literature. Finally, we consider the implications for performance of misspecification of the propensity score, both in terms of an incorrect parametric model for treatment as well as conditioning on the wrong set of covariates.

A summary of our findings is as follows. First, reweighting is approximately unbiased and semiparametrically efficient, even for sample sizes of 100. Our assessment is that reweighting exhibits the best overall finite sample performance of any of the estimators we consider. Second, pair matching shares the good bias performance of reweighting, but has a variance that is roughly 30 percent greater than that of reweighting. Third, kth nearest-neighbor matching, with k chosen by leave-one-out cross-validation, does reduce the excessive variance of pair matching, but at the cost of substantially greater bias. Fourth, kernel, local linear, and ridge matching perform similarly to k-th nearest neighbor matching in exhibiting little variance but much bias when $n = 100$. Once $n = 500$, ridge and local linear matching are both competitive with reweighting on bias and variance grounds.[8] Fifth, both in terms of bias and variance, the popular blocking matching estimator performs neither as badly as kth nearest-neighbor and kernel matching, nor as well as local linear and ridge matching, and is generally dominated by reweighting. Sixth, the double robust estimator is competitive with reweighting, but appears to be slightly more variable and slightly more biased. Seventh, the control function estimator is approximately unbiased, even for samples of size 100, and is approximately semiparametrically efficient once $n = 500$. Eighth, when there is misspecification of the propensity score either due to parametric assumptions or the lack of availability of important covariates, the relative performance of the estimators is approximately as described above. However, in that context, if problems with bias are suspected and variance is less important, pair matching is the preferred estimator.

[7] This issue has been previously taken up by Lunceford and Davidian (2004), Frölich (2004), Zhao (2004), Zhao (2008), and Freedman and Berk (n.d.).

[8] All three of these kernel-based estimators use leave-one-out cross-validation to select a bandwidth, and this model selection issue may be an important aspect of performance for smaller sample sizes.

The above conclusions hold when the propensity score model is correctly specified and when there is good overlap in the distribution of propensity scores for treatment and control units. Our investigations highlight the problems with semiparametric estimators of average treatment effects when overlap is poor. Khan and Tamer (2007) emphasize this point from a theoretical perspective, noting that when overlap is poor the semiparametric efficient bound derived by Hahn (1998) for this problem can be infinite, leading to a failure of $\sqrt{n}$-consistency. Consistent with this conclusion, our results indicate that when overlap is poor, none of the estimators studied work well. In cases where overlap is poor, although technically sufficient to guarantee $\sqrt{n}$-consistency, we document poor performance for $n = 100$, but adequate performance for $n = 500$. This suggests that larger sample sizes may be needed for threshold cases.

A standard empirical approach to problems with overlap is to trim observations with extreme values of the propensity score. We investigate four of the trimming strategies used in the literature. Our simulations suggest that some of these procedures can be effective but only in situations in which the treatment effect is similar for all the observations in the sample. Finally, we provide evidence that as problems with overlap arise, the limiting distribution of semiparametric estimators becomes nonstandard.

Our conclusions run contrary to those of the existing literature on the finite sample performance of reweighting and matching. Our simulations indicate that reweighting is a generally robust estimator whose performance in small samples is as effective as in large samples, where it has been shown to be optimal in a certain sense. The matching methods we consider work poorly for samples of size 100, although some of the methods become effective for samples of size 500. In contrast to these findings, the existing finite sample literature is generally negative regarding reweighting and tends to conclude that matching estimators are best. We review this literature. We show that nearly all of the results from the existing finite sample literature are based on data generating processes (DGPs) for which $\sqrt{n}$-consistent semiparametric estimators do not exist, or DGPs where $\sqrt{n}$-consistency is close to failing. Our own investigations are unusual, in that we focus on DGPs where semiparametric estimators are expected to perform well. We show that this difference in DGPs accounts for our different conclusions.[9]

The remainder of the paper is organized as follows. Section I sets notation, defines estimators, discusses estimands and efficiency bounds, and emphasizes the connections among the many estimators we consider by casting them in the common framework of weighted regression. In particular, this section provides a 3-step interpretation of matching

[9] To be clear, we do not advocate the use of reweighting estimators—or any of the estimators studied here—in settings of failure and near failure of $\sqrt{n}$-consistency of semiparametric estimators of average treatment effects. At present, relatively little is known about appropriate estimation and testing procedures in these settings.

that clarifies the conceptual similarities and differences between the two approaches. In Section II, we describe our benchmark DGP. This DGP is chosen so that semiparametric estimates of average treatment effects are $\sqrt{n}$-consistent. Results for the benchmark DGP are presented in Section III. In Section IV, we take up the issue of DGPs for which $\sqrt{n}$-consistency may be compromised. Results for such DGPs are presented in IV. Section VI compares our results to those of the existing finite sample literature. Section VII concludes.

I Notation and Background

Let $Y_i(1)$ denote the outcome for unit i that would obtain under treatment and $Y_i(0)$ the outcome that would obtain under control. Treatment is denoted by the binary variable T_i. We observe $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$, but never the pair $(Y_i(0), Y_i(1))$. The data $(X_i, Y_i, T_i)_{i=1}^n$ are taken to be independent across i, but are potentially heteroscedastic. Let the propensity score, the conditional probability of treatment, be denoted $p(x) \equiv P(T_i = 1|X_i = x)$. Let the covariate-specific average treatment effect by denoted $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$.

We focus on the relative performance of semiparametric estimators of population averages of $\tau(x)$. Here, semiparametric means an estimator that models the relationship between the probability of receiving treatment and the covariates X_i, but remains agnostic regarding the relationship between the counterfactual outcomes $(Y_i(0), Y_i(1))$ and the covariates.

Semiparametric estimators of treatment effects are typically justified by an appeal to (1) selection on observed variables and (2) sufficient overlap. Selection on observed variables means that treatment is randomized given X_i, or that $(Y_i(0), Y_i(1), Z_i) \perp\!\!\!\perp T_i|X_i$, where Z_i is any characteristic of the individual that is not affected by treatment assignment (e.g. pre-program earnings).[10]This assumption has traditionally been referred to as selection on observed variables in the economics literature (e.g., Heckman and Robb 1985). In the statistics and more recent econometrics literature this assumption is instead referred to as ignorability or unconfoundedness (e.g., Rosenbaum and Rubin 1983, Imbens 2004).[11]

Selection on observed variables is not by itself sufficient to semiparametrically identify average treatment effects. The DGPs we focus on in Section III are consistent with both selection on observed variables and a *strict overlap* assumption: $\xi < p(x) < 1 - \xi$ for almost every x in the support of X_i, for some $\xi > 0$. This assumption is stronger than the *standard overlap* assumption that $0 < p(x) < 1$ for almost every x in the support of

[10] Notice that some pre-program covariates may be affected by anticipation of treatment.

[11] Lechner (2005) shows that some control variables may be influenced by the treatment. However, this endogeneity does not matter for consistency of the treatment effect estimator, as long as the usual formulation of the conditional independence assumption holds.

X_i (e.g., Rosenbaum and Rubin 1983, Heckman et al. 1997, Hahn 1998, Wooldridge 2002, Imbens 2004, Todd 2007), but is also common in the literature (e.g., Robins et al. 1994, Abadie and Imbens 2006, 2008, Crump, Hotz, Imbens and Mitnik 2007a,b). Both the standard overlap and the strict overlap assumptions are strong. Khan and Tamer (2007) emphasize that something akin to the strict overlap assumption is needed to deliver $\sqrt{n}$-consistency of semiparametric estimators in this context. We take up the issue of DGPs that violate strict overlap, but satisfy standard overlap, in Sections IV and V, below.

A Estimands and Estimators

As noted, we focus on the performance of estimators for target parameters that are averages of $\tau(x)$. The specific averages we consider are the average treatment effect $\alpha = E[\tau(X_i)]$ and the average treatment effect on the treated $\theta = E[\tau(X_i)|T = 1]$. We refer to these estimands as ATE and TOT, respectively. Although we focus on the performance of estimators for these estimands, we emphasize that ATE and TOT are not the only estimands of interest. However, the performance of these estimators for ATE and TOT is likely to be similar to the performance of these estimators when adapted to estimation of other averages of $\tau(x)$.

We consider fourteen estimators: nine matching estimators, three reweighting estimators (sometimes termed inverse propensity score weighting estimators, or IPW), one control function estimator, and the so-called double robust estimator.[12] Each of these estimators are two-step estimators relying on a first-step estimate of the propensity score. The nine matching estimators include pair matching, kth nearest neighbor matching, kernel matching, local linear matching, ridge regression matching, and blocking. Aside from pair matching, each of these matching strategies employs a cross-validation method for choosing a tuning parameter. Kernel, local linear, and ridge matching all further require the choice of a kernel. Following Frölich (2004), we consider both the Epanechnikov kernel and the Gaussian kernel.

The three reweighting estimators include a reweighting estimator in which the sum of the weights is allowed to be stochastic (IPW1), a reweighting estimator in which the sum of the weights is forced to be 1 (IPW2), and an asymptotically optimal combination of the former two estimators (IPW3) that is due to Lunceford and Davidian (2004).

We also consider the so-called double robust estimator due to Robins and Rotnitzky (1995), which has recently received a good deal of attention in the literature (e.g., Imbens 2004). This procedure can be thought of as a regression-adjusted version of reweighting.

[12] The breadth of coverage arises from an attempt to encompass many of the estimators used in the literature as well as to be consistent with previous finite sample evidence on the topic. Nonetheless, there are of course other potentially effective estimators whose performance is not covered by the analysis here.

The regression adjustments are more similar in spirit to an older approach to the problem of estimating treatment effects. We complete our analysis by studying the performance of a control function estimator. This estimator is essentially the same as the double robust estimator, but is unweighted. The version of the control function estimator we implement models the regression function of the outcome given the covariates and treatment status as a polynomial in the estimated propensity score, with additive and possibly interacted terms for treatment status. This procedure is described in Wooldridge (2002) for the case of ATE, and is in the spirit of Oaxaca (1973) and Blinder (1973) decompositions and Hahn's (1998) general estimator.

While it is true that at least some versions of reweighting and matching are believed to be semiparametrically efficient in large samples, and while both approaches are based on the same first-step propensity score estimate, it is far from clear that the two approaches would perform similarly in finite samples. First, most matching estimators rely on tuning parameters. It is possible that use of tuning parameters could improve finite sample performance relative to reweighting. Second, the approaches take advantage of very different properties of the propensity score. Matching requires of the estimated propensity score only that it be a balancing score (Rosenbaum and Rubin 1983). In contrast, reweighting requires that the propensity score be a conditional probability. For example, matching on the square root of the propensity score should work just as well as matching on propensity score; in contrast, reweighting with the square root of the propensity score should do badly.

B Weighted Least Squares as a Unifying Framework

Both matching and reweighting estimators of average treatment effects can be understood as the coefficient on the treatment indicator in a weighted regression, with weighting functions that differ by estimator. This common structure clarifies that the essential difference between the estimators is the weighting function implicitly used.

That reweighting estimators have this form is widely understood. A general notation for reweighting estimators for the TOT and ATE is

$$\widehat{\theta} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{j=1}^{n}(1-T_j)Y_j\overline{w}(j), \tag{3.1}$$

$$\widehat{\alpha} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i \overline{w}_1(i) - \frac{1}{n_0}\sum_{j=1}^{n}(1-T_j)Y_j\overline{w}_0(j). \tag{3.2}$$

The weights in equations (3.1) and (3.2) only add up to one for some versions of reweighting estimators. When the TOT weights add up to one in the sense of $\frac{1}{n_0}\sum_{j=1}^{n}(1-T_j)\overline{w}(j) = 1$, the TOT estimate can be obtained using standard statistical software from the coefficient on treatment in a regression of the outcome on a constant and a treatment indicator using weights $W = T + (1-T)\overline{w}(\cdot)$. When the weights do not add up to one, the TOT estimate can be calculated directly using equation (3.1). When the ATE weights add up to one in the sense that $\frac{1}{n_0}\sum_{j=1}^{n}(1-T_j)\overline{w}_0(j) = 1$ and $\frac{1}{n_1}\sum_{j=1}^{n}T_j\overline{w}_1(j) = 1$, the ATE estimate can be obtained from the same regression described, but with weights $W = T\overline{w}_1(\cdot) + (1-T)\overline{w}_0(\cdot)$. The reweighting estimators we consider are characterized below by enumerating the weighting functions used.

Weights Used for Reweighting Estimators

Effect	**Treatment,** t	**Estimator**	**Weighting Function** $\overline{w}_t(j)$
TOT	1	IPW1	$\frac{\hat{p}(X_j)}{1-\hat{p}(X_j)} \Big/ \frac{\hat{p}}{1-\hat{p}}$
TOT	1	IPW2	$\frac{\hat{p}(X_j)}{1-\hat{p}(X_j)} \Big/ \frac{1}{n_0}\sum_{k=1}^{n}\frac{(1-T_k)\hat{p}(X_k)}{1-\hat{p}(X_k)}$
TOT	1	IPW3	$\frac{\hat{p}(X_j)}{1-\hat{p}(X_j)}\left(1-C_j\right) \Big/ \frac{1}{n_0}\sum_{k=1}^{n}\frac{(1-T_k)\hat{p}(X_k)}{1-\hat{p}(X_k)}\left(1-C_k\right)$
ATE	0	IPW1	$\frac{1-\hat{p}}{1-\hat{p}(X_j)}$
ATE	1	IPW1	$\frac{\hat{p}}{\hat{p}(X_j)}$
ATE	0	IPW2	$\frac{1}{1-\hat{p}(X_j)} \Big/ \frac{1}{n_0}\sum_{k=1}^{n}\frac{1-T_k}{1-\hat{p}(X_k)}$
ATE	1	IPW2	$\frac{1}{\hat{p}(X_j)} \Big/ \frac{1}{n_1}\sum_{k=1}^{n}\frac{T_k}{\hat{p}(X_k)}$
ATE	0	IPW3	$\frac{1}{1-\hat{p}(X_j)}(1-C_j^0) \Big/ \frac{1}{n_0}\sum_{k=1}^{n}\frac{1-T_k}{1-\hat{p}(X_k)}(1-C_k^0)$
ATE	1	IPW3	$\frac{1}{\hat{p}(X_j)}(1-C_j^1) \Big/ \frac{1}{n_1}\sum_{k=1}^{n}\frac{T_k}{\hat{p}(X_k)}(1-C_k^1)$

Note: $\hat{p} \equiv \frac{n_1}{n}$, $A_i = \frac{1-T_i}{1-\hat{p}(X_i)}, B_i = \frac{T_i}{\hat{p}(X_i)}, C_i = \frac{\left(1-\frac{\hat{p}(X_i)}{\hat{p}}A_i\right)\frac{1}{n}\sum_{j=1}^{n}\left(1-\frac{\hat{p}(X_j)}{\hat{p}}A_j\right)}{\frac{1}{n}\sum_{j=1}^{n}\left(1-\frac{\hat{p}(X_j)}{\hat{p}}A_j\right)^2}$,
$C_i^0 = \frac{\frac{1}{1-\hat{p}(X_i)}\frac{1}{n}\sum_{j=1}^{n}(A_j\ \hat{p}(X_j)-T_i)}{\frac{1}{n}\sum_{j=1}^{n}(A_j\ \hat{p}(X_j)-T_i)^2}$ and $C_i^1 = \frac{\frac{1}{\hat{p}(X_i)}\frac{1}{n}\sum_{j=1}^{n}(B_j(1-\hat{p}(X_j))-(1-T_i))}{\frac{1}{n}\sum_{j=1}^{n}(B_j(1-\hat{p}(X_j))-(1-T_i))^2}$, which are IPW3 correction factors that are small when the propensity score model is well specified.

The functional form given by IPW1 can be found in many treatments in the literature (e.g., Dehejia and Wahba 1997, Wooldridge 2002, Hirano et al. 2003). IPW2 is advocated by Johnston and DiNardo (1996) and Imbens (2004). Since most applied work is based on regression software, which naturally rescales weights, most estimates in the empirical literature are probably IPW2. With a well-specified propensity score model, the weights used in IPW1 should nearly add up to one and IPW1 and IPW2 should not differ dramatically. This is because, ignoring estimation error in $\hat{p}(X_i)$ and $\hat{p}$, iterated expectations shows that $E[W_i] = 1$ for both TOT and ATE. However, in finite samples for some DGPs, the sum of the weights can depart substantially from 1. Unlike IPW2, IPW3 is not commonly implemented in the empirical literature. This estimator, derived by Lunceford and Davidian (2004) for the case of ATE, is the (large sample) variance-minimizing linear combination of IPW1 and IPW2.[13]

[13] The TOT version of IPW3 is new, but follows straightforwardly, if tediously, from the approach outlined

While it is widely understood that reweighting estimators can be implemented as a weighted regression, it is less widely understood that matching estimators share this property.[14] We demonstrate that matching estimators are weighted regressions for the case of TOT.[15] A general notation for a matching estimator of the TOT is (cf., Smith and Todd 2005, eq. 10)

$$\widehat{\theta} = \frac{1}{n_1} \sum_{i \in I_1} \left\{ Y_i - \sum_{j \in I_0} w(i,j) Y_j \right\}, \tag{3.3}$$

where $w(i,j)$ is the weight that the control observation j is assigned in the formation of an estimated counterfactual for the treated observation i, I_1 is the set of n_1 treated units and I_0 is the set of n_0 control units. The weights $w(i,j)$ are in general a function of the distance in the covariates. In the case of propensity score matching, that distance is measured by the difference in the estimated propensity scores. We now describe the matching estimators we consider by enumerating the TOT weighting functions $w(i,j)$.[16]

Weights Used for Matching Estimators for TOT

Estimator	**Weighting Function** $w(i,j)$
kth Nearest Neighbor	$\frac{1}{k}\mathbf{1}(\hat{p}(X_j) \in \mathcal{J}_k(i))$
Kernel	$K_{ij} \Big/ \sum_{j \in I_0} K_{ij}$
Local Linear	$\left(K_{ij} L_i^2 - K_{ij} \hat{\Delta}_{ij} L_i^1\right) \Big/ \sum_{j \in I_0} \left(K_{ij} L_i^2 - K_{ij} \hat{\Delta}_{ij} L_i^1 + r_L\right)$
Ridge	$K_{ij} \Big/ \sum_{j \in I_0} K_{ij} + \widetilde{\Delta}_{ij} \Big/ \sum_{j \in I_0} \left(K_{ij} \widetilde{\Delta}_{ij}^2 + r_R h \lvert \widetilde{\Delta}_{ij} \rvert\right)$
Blocking	$\sum_{m=1}^{M} \mathbf{1}(\hat{p}(X_i) \in B_m) \mathbf{1}(\hat{p}(X_j) \in B_m) \Big/ \sum_{m=1}^{M} \mathbf{1}(\hat{p}(X_j) \in B_m)$

All of the matching estimators enumerated can be understood as the coefficient on the

by those authors (see Appendix I for details).

[14] However, there are clear antecedents in the literature. For example equations (3) and (4) of Abadie and Imbens (2006) clarify this common structure.

[15] The case of ATE then follows since a matching estimator for the ATE is a convex combination of the average treatment effect for the treated and for the untreated, with convex parameter equal to the fraction treated.

[16] The notation is as follows: $\mathcal{J}_k(i)$ is the set of k estimated propensity scores among the control observations that are closest to $\hat{p}(X_i)$, $\hat{\Delta}_{ij} = \hat{p}(X_i) - \hat{p}(X_j)$, $K_{ij} = K(\hat{\Delta}_{ij}/h)$ for $K(\cdot)$ a kernel function and h a bandwidth, $L_i^p = \sum_{j \in I_0} K_{ij} \hat{\Delta}_{ij}^p$, for $p = 1, 2$, $\widetilde{\Delta}_{ij} = \hat{p}(X_j) - \overline{p}(X_i)$, $\overline{p}_i = \sum_{j \in I_0} p_j K_{ij} / \sum_{j \in I_0} K_{ij}$, r_L is an adjustment factor suggested by Fan (1993), r_R is an adjustment factor suggested by Seifert and Gasser (2000), B_m is an interval such as $[0, 0.2]$ that gives the mth block for the blocking estimator, and M is the number of blocks used. For a Gaussian kernel, $r_L = 0$ and for an Epanechnikov kernel, $r_L = 1/n^2$. For a Gaussian kernel, $r_R = 0.35$ and for an Epanechnikov kernel, $r_R = 0.31$.

treatment indicator in a weighted regression. To see this, rewrite

$$\begin{aligned}
\widehat{\theta} &= \frac{1}{n_1}\sum_{i=1}^{n} T_i \left\{ Y_i - \sum_{j=1}^{n} w(i,j)(1-T_j)Y_j \right\} \\
&= \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \sum_{j=1}^{n}(1-T_j)Y_j \frac{1}{n_1}\sum_{i=1}^{n} w(i,j)T_i \\
&\equiv \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{j=1}^{n}(1-T_j)Y_j \overline{w}(j), \qquad (3.4)
\end{aligned}$$

where $\overline{w}(j) = \frac{n_0}{n_1}\sum_{i=1}^{n} w(i,j)T_i$ is proportional to the average weight that a control observation is given, on average across all treatment observations.[17] Viewing matching estimators as weighted least squares is useful as a means of understanding the relationships among the various estimators used in the literature. For example, the weight used by kernel matching can be written as

$$\overline{w}(j) = \frac{n_0}{n_1}\sum_{i=1}^{n} T_i w(i,j) = \frac{\sum_{i=1}^{n} T_i K_{ij} \big/ \sum_{i=1}^{n} K_{ij}}{\sum_{j=1}^{n}(1-T_j)K_{ij} \big/ \sum_{i=1}^{n} K_{ij}} \bigg/ \frac{\hat{p}}{1-\hat{p}}.$$

Ignoring estimation error in the propensity score, $\sum_{i=1}^{n} T_i K_{ij} \big/ \sum_{i=1}^{n} K_{ij}$ is a kernel regression estimate of $P(T_i = 1|p(X_i) = p(X_j))$, which is equivalent to $p(X_j)$.[18] If the kernel in question is symmetric, then $\sum_{i=1}^{n}(1-T_i)K_{ij} \big/ \sum_{i=1}^{n} K_{ij}$ is similarly a kernel regression estimate of $P(T_i = 0|p(X_i) = p(X_j))$, which is equivalent to $1 - p(X_j)$. Thus, for kernel matching with a symmetric kernel, we have

$$\overline{w}(j) \approx \frac{p(X_j)}{1-p(X_j)} \bigg/ \frac{p}{1-p},$$

which is the same as the target parameter of the TOT weight used by reweighting.

This result provides a 3-step interpretation to symmetric kernel matching for the TOT:

1. Estimate the propensity score, $\hat{p}(X_i)$

[17] The matching estimators proposed in the literature require no normalization on the weights involved in the second sum in equation (3.4). This follows because the matching estimators that have been proposed define the weighting functions $w(i,j)$ in such a way that $\sum_{j \in I_0} w(i,j)Y_j$ is a legitimate average of the controls, in that sense that for every treated unit i, $\sum_{j \in I_0} w(i,j) = \sum_{j=1}^{n} w(i,j)(1-T_j) = 1$. This has the important implication that the weights in the second sum of equation (3.4) automatically add up to one:

$$\frac{1}{n_0}\sum_{j=1}^{n}(1-T_j)\overline{w}(j) = \frac{1}{n_0}\sum_{j=1}^{n}\left\{(1-T_j)\left[\frac{n_0}{n_1}\sum_{i=1}^{n} w(i,j)T_i\right]\right\} = \frac{1}{n_1}\sum_{i=1}^{n}\left\{T_i\left[\sum_{j \in I_0} w(i,j)\right]\right\} = \frac{1}{n_1}\sum_{i=1}^{n} T_i = 1.$$

[18] Generally, if X and Y are random variables such that $m(X) = E[Y|X]$ exists, then $E[Y|m(X)] = m(X)$ by iterated expectations.

2. For each observation j in the control group, compute $\overline{p}(X_j) = \sum_{i=1}^n T_i K_{ij} \big/ \sum_{i=1}^n K_{ij}$. In words, this is the fraction treated among those with propensity scores near $\hat{p}(X_j)$. Under smoothness assumptions on $p(X_i)$, this will be approximately $\hat{p}(X_j)$.
3. Form the weight $\overline{w}(j) = \left(\overline{p}(X_j)/(1-\overline{p}(X_j))\right) \Big/ \left(\hat{p}/\left(1-\hat{p}\right)\right)$ and run a weighted regression of Y_i on a constant and T_i with weight $W_i = T_i + (1-T_i)\overline{w}(i)$.

Reweighting differs from this procedure in that, in step 2, it directly sets $\overline{p}(X_j) = \hat{p}(X_j)$. The simulation suggests that this shortcut is effective at improving small sample performance.

C Mixed Methods

We also consider the performance of an estimator known as "double robust" that is neither reweighting nor matching but is a hybrid procedure combining reweighting with more traditional regression techniques. This procedure is discussed by Robins and Rotnitzky (1995) in the related context of imputation for missing data. Imbens (2004) provides a good introductory treatment.

To describe the intuition behind this estimator, we first return to a characterization of reweighting. The essential idea behind reweighting is that in large samples, reweighting ensures orthogonality between the treatment indicator and any possible function of the covariates. That is, for any bounded continuous function $g(\cdot)$,

$$\begin{aligned} E\left[g(X_i)|T_i=1\right] &= E\left[g(X_i)\frac{p(X_i)}{1-p(X_i)}\Big/\frac{p}{1-p}\Big|T_i=0\right], \\ E\left[g(X_i)\frac{p}{p(X_i)}\Big|T_i=1\right] &= E\left[g(X_i)\frac{1-p}{1-p(X_i)}\Big|T_i=0\right] = E\left[g(X_i)\right]. \end{aligned} \tag{3.5}$$

This implies that the joint distribution of X_i is equal in weighted subsamples defined by $T_i = 1$ and $T_i = 0$, using either TOT or ATE weights.[19] This in turn implies that in the reweighted sample, treatment is *unconditionally* randomized, and estimation can proceed by computing the (reweighted) difference in means, as described in subsection B, above. A standard procedure in estimating the effect of an unconditionally randomized treatment is to include covariates in a regression of the outcome on a constant and a treatment indicator. It is often argued that this procedure improves the precision of estimated treatment effects. By analogy with this procedure, a reweighting estimator may enjoy improved precision if the weighted regression of the outcome on a constant and a treatment indicator is augmented by covariates.

The estimator just described is the double robust estimator. Reweighting computes average treatment effects by running a weighted regression of the outcome on a constant

[19] Given the standard overlap assumption, this result follows from iterated expectations. A proof for the case of TOT is given in McCrary (2007, fn. 35).

and a treatment indicator. Double robust estimation computes average treatment effects by running a weighted regression of the outcome on a constant, a treatment indicator, and some function of the covariates such as the propensity score.

The gain in precision associated with moving from a reweighting estimator to a double robust estimator is likely modest with economic data.[20] However, a potentially important advantage is that the estimator is more likely to be consistent, in a particular sense. Suppose that the model for the treatment equation is misspecified, but that the model for the outcome equation is correctly specified. Then the double robust estimator would retain consistency, despite the misspecification of the treatment equation model.[21] We implement the double robust estimator by including the estimated propensity score linearly into the regression model, for both ATE and TOT.[22]

The double robust estimator is related to another popular estimator that we call a control function estimator. For the case of ATE, the control function estimator is the slope coefficient on a treatment indicator in a regression of the outcome on a constant, the treatment indicator, and functions of the covariates X_i. For the case of TOT, we obtain the control function estimator by running a regression of the outcome on a constant and a cubic in the propensity score, separately by treatment status.[23] For each model, we form predicted values, and compute the average difference in predictions, among the treated observations. This procedure is in the spirit of the older Oaxaca (1973) and Blinder (1973) procedure and is related to the general estimator proposed by Hahn (1998).

D Tuning Parameter Selection

The more complicated matching estimators require choosing tuning parameters. Kernel-based matching estimators require selection of a bandwidth, nearest-neighbor matching

[20] Suppose the goal is to obtain a percent reduction of q in the standard error on the estimated treatment effect. Approximate the standard error of the treatment effect by the spherical variance matrix least squares formula. Then reducing the standard error of the estimated treatment effect by q percent requires reducing the regression root mean squared error by q percent, since the "matrix part" of the standard error is affected only negligibly by the inclusion of covariates, due to the orthogonality noted in equation (3.5). This requires reducing the regression mean squared error (MSE) by roughly $2q$ percent when q is small. A $2q$ percent reduction in the regression MSE requires that the F-statistic on the exclusion of the added covariates be a very large $2qn/K$, where n is the overall sample size and K is the number of added covariates. Consider one of the strongest correlations observed in economic data, that between log-earnings and education. In a typical U.S. Census file with 100,000 observations, the t-ratio on the education coefficient in a log-earnings regression is about 100 (cf., Card 1999). The formula quoted suggests that including education as a covariate with an outcome of log earnings would improve the standard error on a hypothetical treatment indicator by only 5 percent.

[21] In the case described, the double robust estimator would be consistent, but inefficient relative to a regression-based estimator with no weights, by the Gauss-Markov Theorem.

[22] We include the $\hat{p}(X_i)$ rather than X_i because the outcome equation in our DGPs is a function of $p(X_i)$.

[23] In simulations not shown, we computed the MSE for the control function estimator in which the propensity score entered in a polynomial of order 1,...,5. The cubic polynomial had the lowest MSE on average across contexts.

requires choosing the number of neighbors, and blocking requires choosing the blocks.

In order to select both the bandwidth h to be used in the kernel-based matching estimators and the number of neighbors to be utilized in nearest neighbor matching, we implement a simple leave-one-out cross-validation procedure that chooses h as

$$h^* = \arg\min_{h \in H} \sum_{i \in I_0} \left[Y_i - \hat{m}_{-i}\left(p\left(X_i\right)\right)\right]^2 ,$$

where $\hat{m}_{-i}\left(p\left(X_i\right)\right)$ is the predicted outcome for observation i, computed with observation i removed from the sample, and $m\left(\cdot\right)$ is the non-parametric regression function implied by each matching procedure. For kernel, local linear and ridge matching the bandwidth search grid H is $0.01 \times |\kappa|\, 1.2^{g-1}$ for $g = 1, 2, ..., 29, \infty$. For nearest-neighbor matching the grid H is $\{1, 2, ..., 20, 21, 25, 29, ..., 53, \infty\}$ for a sample size smaller than 500 and $\{1, 2, 5, 8, .., 23, 28, 33, ..., 48, 60, 80, 100, \infty\}$ for 500 or more observations.[24]

For the blocking estimator, we first stratify the sample into M blocks defined by intervals of the estimated propensity score. We continue to refine the blocks until within each block we cannot reject the null that the expected propensity score among the treated is equal to the expected propensity score among the controls (Rosenbaum and Rubin 1983, Dehejia and Wahba 1999). In order to perform this test we used a simple t-test with a 99 percent confidence level. Once the sample is stratified, we can compute the average difference between the outcome of treated and control units that belong to each block, $\hat{\tau}_m$. Finally, the blocking estimator computes the weighted average of $\hat{\tau}_m$ across M blocks, where the weights are the proportion of observations in each block, either overall (ATE) or among the treated only (TOT).

E Efficiency Bounds

In analyzing the performance of the estimators we study, it useful to have an idea of a lower bound on the variance of the various estimators for a given model. Estimators which attain a variance lower bound are best, in a specific sense.

We consider two variants of efficiency bounds. The first of these is the Cramér-Rao lower bound, which can be calculated given a fully parametric model. The semiparametric models motivating the estimators under study in this paper do not provide sufficient detail on the putative DGP to allow calculation of the Cramér-Rao bound. Nonetheless, since we assign the DGP in this study, we can calculate the Cramér-Rao bound using this knowledge. This forms a useful benchmark. For example, we will see that in some models, the variance of a semiparametric estimator is only slightly greater than the Cramér-Rao bound. These

[24] For more details on this procedure see Stone (1974) and Black and Smith (2004) for an application.

are then models in which there is little cost to discarding a fully parametric model in favor of a semiparametric model.

The second efficiency bound we calculate is the semiparametric efficiency bound. These bounds can be viewed as the smallest variance that can be obtained without imposing parametric assumptions on the outcome equation. Alternatively, the SEB can be viewed as the least upper bound of the Cramér-Rao bounds, among the set of DGPs consistent with the parametric assumptions placed on the treatment equation. An introductory discussion of the SEB concept is given in Newey (1990). Hahn (1998, Theorems 1, 2) shows that under selection on observed variables and standard overlap, the SEB is given by

$$\begin{aligned}
(3.6)\qquad SEB_{k/u}^{ATE} &= E\left[\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1-p(X_i)} + (\tau(X_i)-\alpha)^2\right],\\
(3.7)\qquad SEB_{u}^{TOT} &= E\left[\frac{\sigma_1^2(X_i)p(X_i)}{p^2} + \frac{\sigma_0^2(X_i)p(X_i)^2}{p^2\left(1-p(X_i)\right)} + \frac{p(X_i)}{p^2}\left(\tau(X_i)-\theta\right)^2\right],\\
(3.8)\qquad SEB_{k}^{TOT} &= E\left[\frac{\sigma_1^2(X_i)p(X_i)}{p^2} + \frac{\sigma_0^2(X_i)p(X_i)^2}{p^2\left(1-p(X_i)\right)} + \frac{p(X_i)^2}{p^2}\left(\tau(X_i)-\theta\right)^2\right],
\end{aligned}$$

where the subindex $l = k, u$ indicates whether the propensity score is known or unknown and $\sigma_t^2(X_i)$ is the conditional variance of $Y_i(t)$ given X_i.

Reweighting using a nonparametric estimate of the propensity score achieves the bounds in equations (3.6) and (3.7), as shown by Hirano et al. (2003) for both the ATE and TOT case. Nearest neighbor matching on covariates using a Euclidean norm also achieves these bounds when the number of matches is large. Abadie and Imbens (2006, Theorem 5) demonstrate this for the case of ATE and the case of TOT follows from the machinery they develop.[25] However, nearest neighbor matching is inconsistent when there is more than one continuous covariate to be matched.

Efficiency results for other matching estimators are not yet available in the literature. In Appendix II we provide a derivation of the SEB for the models used in our simulations. In Appendix Table 3.A6 we show the SEB and the CRLB for all of the data generating processes used in this paper. We turn now to a description of these models.

[25] Using their equation (13) and the results of the unpublished proof of their Theorem 5, it is straightforward to derive the large sample variance of the kth nearest-neighbor matching estimator for the TOT as

$$(3.9)\qquad SEB_u^{TOT} + \frac{1}{2k}E\left[\frac{\sigma_0^2(X_i)}{p}\left(\frac{1}{1-p(X_i)} - (1-p(X_i))\right)\right] \xrightarrow[k\to\infty]{} SEB_u^{TOT}$$

II Data Generating Process

The DGPs we consider are all special cases of the latent index model

$$\begin{aligned} T_i^* &= \eta + \kappa X_i - u_i, && (3.10) \\ T_i &= 1\left(T_i^* > 0\right), && (3.11) \\ Y_i &= \beta T_i + \gamma m(p(X_i)) + \delta T_i m(p(X_i)) + \varepsilon_i, && (3.12) \end{aligned}$$

where u_i and ε_i are independent of X_i and of each other, $m(\cdot)$ is a curve to be discussed, and $p(X_i)$ is the propensity score implied by the model, or the probability of treatment given X_i. The covariate X_i is taken to be distributed standard normal. Our focus is on cross-sectional settings, so ε_i is independent across i, but potentially heteroscedastic. This is achieved by generating e_i as an independent and identically distributed standard normal sequence and then generating

$$\varepsilon_i = \psi\left(e_i p(X_i) + e_i T_i\right) + (1-\psi)e_i. \tag{3.13}$$

We consider several different distributional assumptions for the treatment assignment equation residual, u_i. As we discuss in more detail in Sections IV and V below, the choice of distribution for u_i can be relevant to both the finite and large sample performance of average treatment effect estimators. Let the distribution function for u_i be denoted generally by $F(\cdot)$. Then the propensity score is given by

$$p(X_i) \equiv P(T_i^* > 0) = F(\eta + \kappa X_i). \tag{3.14}$$

The model given in equations (3.10) through (3.12) nests a basic latent index regression model, in which treatment effects vary with X_i but are homogeneous, residuals are homoscedastic, and the conditional expectation of the outcome under control is white noise. The model is flexible, however, and can also accommodate heterogeneous treatment effects, heteroscedasticity, and nonlinear response functions.

Heterogeneity of treatment effects is controlled by the parameter δ in equation (3.12). When $\delta = 0$, the covariate-specific treatment effects are constant: $\tau(x) = \beta$ for all x. Thus under this restriction the average treatment effect (ATE) and the average effect of treatment on the treated (TOT) both equal β in the population and in the sample.[26] When $\delta \neq 0$, the covariate-specific treatment effect, given by $\tau(x) = \beta + \delta m(p(x))$, depends on the covariate and ATE and TOT may differ.[27] Heteroscedasticity is controlled by the parameter ψ in equation (3.13). When $\psi = 0$, we obtain homoscedasticity. When $\psi \neq 0$,

[26] For a discussion of the distinction between sample and population estimands, see Imbens (2004), for example.

[27] For a discussion of other estimands of interest, see Heckman and Vytlacil (2005), for example.

the residual variance depends on treatment as well as on the propensity score. The function $m(\cdot)$ and the parameter γ manipulate the non-linearity of the outcome equation that is common to both treated and non-treated observations.[28]

We assess the relative performance of the estimators described in Section I in a total of twenty-four different contexts. These different contexts are characterized by four different settings, three different designs, and two different regression functions. We now describe these contexts in greater detail.

The four settings we consider correspond to four different combinations of the parameters in equation (3.12): β, γ, δ, and ψ. In each of these four settings, we set $\beta = 1$ and $\gamma = 1$. However, we vary the values of the parameters δ and ψ, leading to four combinations of homogeneous and heterogeneous treatment effects and homoscedastic and heteroscedastic error terms. The specific configurations of parameters used in these four settings are summarized below:

Setting	β	γ	δ	ψ	**Description**
I	1	1	0	0	homogeneous treatment, homoscedastic
II	1	1	1	0	heterogeneous treatment, homoscedastic
III	1	1	0	2	homogeneous treatment, heteroscedastic
IV	1	1	1	2	heterogeneous treatment, heteroscedastic

The two regression functions we consider, $m(\cdot)$, correspond to the functional forms used by Frölich (2004). The first curve considered is a simple linear function. The second curve is nonlinear and rises from around 0.7 at $q = 0$ to 0.8 near $q = 0.4$, where the curve attains its peak, before declining to 0.2 at $q = 1$. The precise equations used for these two regression functions are summarized below:

Curve	**Formula**	**Description**
1	$m_1(q) = 0.15 + 0.7q$	Linear
2	$m_2(q) = 0.2 + \sqrt{1-q} - 0.6(0.9-q)^2$	Nonlinear

Finally, the three designs we consider correspond to different combinations of the parameters in equation (3.10): η and κ. These parameters control degrees of overlap between the densities of the propensity score of treated and control observations as well as different ratios of control to treated units. The specific configurations of parameter values for η and κ are different in Sections III and V and are enumerated in those sections.

[28] Note that we only consider DGPs in which $\gamma \neq 0$. When $\gamma = 0$, all estimators of the TOT for which the weighting function $\overline{w}(j)$ adds up to 1 can analytically be shown to be finite sample unbiased. When $\gamma \neq 0$, no easy analytical finite sample results are available, and simulation evidence is much more relevant.

III Results: Benchmark Case

We begin by focusing on the case of X_i distributed standard normal and u_i distributed standard Cauchy.[29] As we discuss in more detail below, an initial focus on this DGP allows us to sidestep some important technical issues that arise with poor overlap in propensity score distributions between treatment and control units. We defer discussion of these complications until Sections IV and V. The specific configurations of the parameters η and κ used in these three designs are summarized below:

Design	η	κ	**Treated-to-Control Ratio**
A	0	0.8	1:1
B	0.8	1	2:1
C	-0.8	1	1:2

An important feature of these DGPs is the behavior of the conditional density functions of the propensity score, $p(X_i)$, conditional on treatment. Figure 3.1A displays the conditional density of the propensity score given treatment status. This figure features prominently in our discussion, and we henceforth refer to such a figure as an *overlap plot*.

The figures point to several important features of our benchmark DGPs. First, for all three designs considered, the strict overlap assumption is satisfied. As noted by Khan and Tamer (2007), this is a sufficient assumption for $\sqrt{n}$-consistency of semiparametric treatment effects estimators. Second, the ratio of the treatment density height to that for control gives the treatment-to-control sample size ratio. From this we infer that it is more challenging to estimate the TOT in design C than in designs A or B. Third, design A is symmetric and estimation of the ATE is no more difficult than estimation of the TOT.

We turn next to an analysis of the results of the simulation. In Section III.A we assume that the propensity score model is correctly specified, and estimation proceeds using a maximum likelihood Cauchy binary choice model that includes X_i as the sole covariate. In Section III.B we study the impact of misspecification of the propensity score model on performance.

In both sections III.A and III.B, and throughout the paper, we report separate estimates of the bias and the variance of the estimators. In addition, for each estimator we test the hypothesis that the bias is equal to zero, and we test the hypothesis that the variance is equal to the SEB. These choices reflect our view that it is difficult to bound the bias a researcher would face, across the possible DGPs the researcher might confront, unless the

[29] In principle, we could have let u_i follow a normal distribution with parameters selected in a manner that they allow for good overlap. In such a case, because in the normal case the parameters that manipulate overlap also change the ratio of treated to control observations in the designs, we would not be able to explore designs as the ones we can study when u_i is distributed Cauchy.

estimator is unbiased or nearly so. Bounding the bias is desirable under an objective of minimizing the worst case scenario performance of the estimator, across possible DGPs.

A Correct Parametric Specification of Treatment Assignment

Table 3.1 examines the performance of our 14 estimators in the Normal-Cauchy model for $n = 100$ and $n = 500$. For ease of exposition, we do not show estimates of the bias and variance for all twenty-four contexts.[30] Instead, we summarize these estimates by presenting the simulated root mean square bias (RMSB) and average variance, both overall across the twenty-four contexts and separately for the settings described in Section II.[31] There are 14 columns, one for each estimator under consideration.

Estimates of the RMSB are presented in the first and second panels of Table 3.1 for $n = 100$ and $n = 500$, respectively. As an aid to summarizing the results, we additionally perform F-tests of the null hypothesis that the bias is zero jointly across the twenty-four contexts and jointly across the designs and curves in any given setting.[32] The value of the F-statistic for the joint test across twenty-four contexts is reported below the setting-specific RMSB estimates, and p-values for these F-tests are reported in brackets.[33] The values of the F-statistics for the setting-specific tests are suppressed in the interest of space. For these tests, we place an asterisk next to the RMSB when the hypothesis is rejected at the 1% significance level.

Average variances are presented in the third and fourth panels of Table 3.1 for $n = 100$ and $n = 500$, respectively. We provide a reference point for these variances using the SEB.[34] Below the average variances we report the percentage difference between the estimated variance and the SEB on average across all twenty-four contexts. We also perform a F-test of the equality of the variance estimates and the SEB, jointly across all twenty-four contexts

[30] As described above, a *context* here means a bundle of setting, design, and curve. We consider four settings, three designs and two curves.

[31] In the main text, we focus on TOT and report summary tables. A series of appendix tables present summary tables for ATE. Detailed tables for both TOT and ATE, as well as Stata data sets containing all of the replication results, are available at `http://www.econ.berkeley.edu/~jmccrary`.

[32] Practically, these tests are implemented as Wald tests using a feasible generalized least squares model for the 240,000 replications less their (context-specific) target parameters. To keep the power of these tests constant across sample sizes, we keep nR constant at one million, where R is the number of replications. This implies 10,000 replications for $n = 100$ and 2,000 replications for $n = 500$. This also spares significant computational expense.

[33] Logical equivalence of null hypotheses implies that these F-tests can be viewed as (i) testing that all twenty-four biases are zero, (ii) testing that all four setting-specific RMSB are zero, or (iii) testing that the overall RMSB is zero.

[34] Table 3.A6 presents the SEB for each of the twenty-four contexts in question and contrasts this semi-parametric bound with the parametric Cramér-Rao bound. Details of these computations are provided in Appendix II. Because the overlap is generally good in the Normal-Cauchy model, the SEB is only 8% higher than the Cramér-Rao bound on average across contexts and never more than 28% higher. Note that the variances reported in Table 3.1 for $n = 100$ are to be compared to $10 \times SEB$ ($10 \times SEB = 1000 \times (SEB/100)$) and for $n = 500$ are to be compared to $2 \times SEB$.

and separately for each setting.[35] The F-statistic for the joint test across all twenty-four contexts is presented below the average percent discrepancy between the variances and the SEBs. For the setting-specific test, we suppress the value of the statistic in the interest of space. For these tests, we place an asterisk next to the average variance when the hypothesis is rejected at the 1% significance level.

We turn now to a discussion of the results, beginning with the evidence on bias for $n = 100$. The results suggest several important conclusions. First, the pair matching, reweighting, double robust, and control function estimators are all approximately unbiased. Of these, IPW1 and IPW2 are probably the least biased, performing even better than pair matching. Double robust seems to acquire slightly greater bias in settings with treatment effect heterogeneity, whereas the other unbiased estimators acquire slightly less. The F-statistics reject the null of zero bias at the 5% level of significance for all estimators except IPW1, IPW2, and control function. Second, all matching estimators that rely upon tuning parameters are noticeably biased. We suspect that this is due to the difficulty of accurate estimation of nonparametric tuning parameters.[36] Of these estimators, ridge matching performs best, particularly when the Epanechnikov kernel is used.

For $n = 500$, pair matching, reweighting, double robust and control function remain approximately unbiased. In terms of bias, these estimators perform remarkably similarly for this sample size. For the more complicated matching estimators, we see reduced bias in all cases as expected, and local linear and ridge matching become competitive with reweighting with the larger sample size. Although we can still reject the null of no bias, blocking becomes much less biased. The bias of nearest-neighbor and kernel matching remains high in all settings.

When analyzing the performance within settings (see appendix tables) we observe similar patterns of relative performance. First, reweighting, double robust, and control function estimators are all unbiased regardless of the shape of the overlap plots and regardless of the ratio of treated to control observations. Second, treatment effect heterogeneity, homoscedasticity, and nonlinearity of the regression response function all affect relative performance negligibly.

We next discuss the variance results, presented in the bottom half of Table 3.1. These results reveal several important findings. First, pair matching presents the largest variance

[35] Practically, these tests are implemented as Wald tests using a generalized least squares model for the twenty-four estimated variances less their (context-specific) SEB. The variance of the variance can be approximated quite accurately under an auxiliary assumption that the estimates of the TOT are distributed normally. In that case, the variance of the variance is approximately $2\hat{V}^2/(R-1)$, where $\hat{V}$ is the sample variance itself and R is the number of replications. See Wishart (1928) and Muirhead (2005, Chapter 3).

[36] Loader (1999) reports that the rates of convergence of cross validation is $O_p(n^{-1/10})$ which could explain the bad performance of these estimators in small samples. See also, Galdo and Black (2007) for further discussion on alternative cross-validation methods.

of all the estimators under consideration in all four settings, for both $n = 100$ and $n = 500$. Second, for $n = 100$, IPW2, IPW3 and double robust have the lowest variance among unbiased estimators. Once $n = 500$, the SEB is essentially attained by all of the unbiased estimators except for pair matching. Compared to the SEB, IPW3 has on average a variance for $n = 100$ that is 3.5% in excess, IPW2 a variance that is 4% in excess, and double robust a variance that is 6.4% in excess. Once $n = 500$, these percentages decline to 1%, 1.2%, and 1.4%, respectively.[37] Third, among the biased estimators, those with highest bias (nearest-neighbor and kernel matching) are the ones that present the lowest variance. On average the variance of these estimators is below the SEB. This suggests that if these estimators are asymptotically efficient, then they have a variance which approaches the SEB from below. This conjecture is particularly plausible since local linear and ridge matching, the least biased among the matching estimators, exhibit variance similar to that of the reweighting estimators.

In sum, our analysis indicates that when good overlap is present and misspecification is not a concern, there is little reason to use an estimator other than IPW2 or perhaps IPW3. These estimators are trivial to program, typically requiring 3 lines of computer code, appear to be subject to minimal bias, and are minimal variance among approximately unbiased estimators.

B Incorrect Specification of Treatment Assignment

We investigate two different types of misspecification of the propensity score. First, we assume that $p\left(X_i\right) = F\left(\eta + \kappa X_i\right)$ when in fact the true DGP is $p\left(X_i\right) = F(\eta + \kappa X_{1i} + X_{2i} + X_{3i})$ where X_{ji} follows a standard normal distribution and $F\left(\cdot\right)$ is a Cauchy distribution. We call this a misspecification in terms of covariates, X_i. This kind of misspecification occurs when the researcher fails to include all confounding variables in the propensity score model. Second, we proceed with estimation as if $p\left(X_i\right) = \tilde{F}\left(\eta + \kappa X_i\right)$ when in fact the true DGP is $p\left(X_i\right) = F\left(\eta + \kappa X_i\right)$. In particular, we keep $F\left(\cdot\right)$ as the distribution function for the standard Cauchy, but estimate the propensity score with a probit—that is, we assume that $p(X_i) = \Phi(\eta + \kappa X_i)$. We call this a misspecification in terms of the treatment equation residual, u_i.

Results of these investigations are displayed in Table 3.2. The structure of this table is similar to that of Table 3.1. Table 3.2 presents the RMSB and average variance for the

[37] Although IPW1 does notably worse in terms of variance than IPW2, its performance is not as bad as has been reported in other studies. For instance, Frölich (2004) reports that in a homoscedastic and homogeneous setting IPW1 has an MSE that is between 150% and 1518% higher than that of pair-matching. The good performance of IPW1 documented in Table 3.1 is due to the fact that, in the Normal-Cauchy model, there is a vanishingly small probability of having an observation with a propensity score close to 1. It is propensity scores near 1 that generate extreme weights, and it is extreme weights that lead to large variance of weighted means.

14 estimators in a sample size of 100 under the two types of misspecifications. Covariate misspecification is treated in panels 1 and 3, and distributional misspecification is treated in panels 2 and 4.

The first panel shows that covariate misspecification leads every estimator to become biased in every setting. This is expected and emphasizes the central role of the assumption of selection on observed variables. Unless the unexplained variation in treatment status resembles experimental variation, treatment effects estimators cannot be expected to produce meaningful estimates. These estimators may continue to play a role as descriptive tools, however. The third panel shows that the average variances are always below the SEB, typically by 20% to 30%. Thus, the exclusion of relevant covariates from the propensity score model may lead to precise estimates of the wrong parameter.

We turn next to the results on distributional misspecification, where the DGP continues to have a Cauchy residual on the treatment assignment equation, but the researcher uses a probit model for treatment. The second panel presents results for the bias in this case. In this situation, only pair matching and control function remain unbiased. Double robust is approximately unbiased only in settings of homogeneous treatment effects. The reweighting estimators become biased but are always less biased than the matching estimators. The fourth panel shows that none of the estimators achieve the SEB. Unfortunately, the most robust estimators to misspecification of the propensity score, that is pair matching and control function, are the ones with the largest variance. Ridge matching and IPW3 are closest to the SEB, differing only by 4% to 6%.

IV Problems with Propensity Scores Near Boundaries

The model given in equations (3.10) to (3.12) assumes selection on observed variables. As has been noted by many authors, selection on observed variables is a strong assumption. It is plausible in settings where treatment is randomized conditional on the function of the X_i given in (3.12). However, it may not be plausible otherwise.[38] We feel that practitioners appreciate the importance of this assumption.

However, perhaps less widely appreciated than the importance of the selection on observed variables assumption is the importance of overlap assumptions. As emphasized by Khan and Tamer (2007), the model outlined in equations (3.10) to (3.12)—while quite general and encompassing all of the existing simulation evidence on performance of estimators for ATE and TOT under unconfoundedness of treatment—does not necessarily

[38] We have emphasized the strength of this assumption by writing the selection on observed variables assumption differently than is typical in the literature (see Section I; cf., Imbens (2004)).

admit a $\sqrt{n}$-consistent semiparametric estimator for ATE or TOT. In particular, the standard overlap assumption that $0 < p(X_i) < 1$ is not sufficient to guarantee $\sqrt{n}$-consistency, whereas the strict overlap assumption that $\xi < p(X_i) < 1 - \xi$ for some $\xi > 0$ is. However, the strict overlap assumption can be violated by the model in equations (3.10) to (3.12). For example, Khan and Tamer (2007) note that $\sqrt{n}$-consistency is violated in the special case of X_i and u_i both distributed standard normal, with $\eta = 0$ and $\kappa = 1$. The following proposition sharpens this important result.

Proposition 3.1. *Under the model specified in equations (3.10) to (3.12), with X_i and u_i distributed standard normal, boundedness of the conditional variance of e_i given X_i, and boundedness of the function $m(\cdot)$, $\sqrt{n}$-consistent semiparametric estimators for ATE and TOT are available when $-1 < \kappa < 1$. For $|\kappa| \geq 1$, no $\sqrt{n}$-consistent semiparametric estimator can exist.*

The proof of this result is tedious but elementary and uses bounds on the distribution function of the standard normal distribution to bound the integral directly. We do not include it here, because it is redundant with the integral bounds used to derive the SEB when it is finite. These are given in Appendix II.[39]

Intuitively, when κ grows is magnitude an increasing mass of observations have propensity scores near 0 and 1, leading to fewer and fewer comparable observations. This leads to an effective sample size that is smaller than n, and the discrepancy between the effective sample size and n grows smoothly with κ. This is important, because it implies potentially poor finite sample properties of semiparametric estimators, in contexts where κ is *near* 1. This is confirmed by the simulation results presented in Section V, below.

Assuming both X_i and u_i are distributed continuous, the extent to which the propensity score fluctuates near 0 and 1 is given by the functional form of the density of the propensity score

$$f_{p(X_i)}(q) \quad = \quad \frac{1}{|\kappa|} g\left(\left(F^{-1}(q) - \eta\right)/\kappa\right) \Big/ f(F^{-1}(q)), \tag{3.15}$$

where $F(\cdot)$ and $f(\cdot)$ are the distribution and density functions, respectively, for u_i, and $g(\cdot)$ is the density function for X_i.[40] For q near one (zero), $F^{-1}(q)$ is of extremely large

[39] Formally, the results of the Proposition follow because semiparametric estimators with $\sqrt{n}$-consistency are only available in situations in which the SEB is finite. The functional form of these bounds, given in Section I.E, involves terms akin to the expectation of the inverse of $p(X_i)$ (for ATE) and the expectation of the inverse of $1-p(X_i)$ (for both ATE and TOT). For $\kappa = 1$, the density of the propensity scores is uniform on $[0, 1]$, and for larger values of κ, the density of the propensity scores becomes an upward-facing parabola. The fact that the density has positive height at 0 and 1 implies immediately that the expectations of the inverse of $p(X_i)$ and of $1 - p(X_i)$ are infinite. The only difficult aspect of the proof of the Proposition is to show that these expectations are in fact finite whenever the height of the density at 0 and 1 is zero.

[40] The equation in the display also holds when X_i is a vector. In that case, the density of a linear

magnitude and positive (negative) sign. Thus, the functional form given makes it clear that when η and κ take on modest values, the density of $p(X_i)$ is expected to be zero at one (zero) when the positive (negative) tail of $f(\cdot)$, the density for the residual, is fatter than that of $g(\cdot)$, the density for the covariate. When the tails of the density for the residual are too thin relative to those of the covariate, the density of $p(X_i)$ near zero can take on positive values, in which case the SEB is guaranteed to be infinite and $\sqrt{n}$-consistency is lost.

This is a useful insight, because the behavior of the propensity score density near the boundary can be inferred from data. In fact, many economists already analyze density estimates for the estimated propensity score, separately by treatment status (see, for example, Figure 3.1 of Black and Smith (2004)). As discussed above, we refer to this graphical display as an *overlap plot*. The unconditional density function is simply a weighted average of the two densities presented in an overlap plot. Thus, the behavior of the unconditional density near the boundaries can be informally assessed using a graphical analysis that is already standard in the empirical literature.[41] When the overlap plot shows no mass near the corners, semiparametric estimators enjoy $\sqrt{n}$-consistency. When the overlap plot shows strictly positive height of the density functions at 0 (for ATE) or 1 (for ATE or TOT), no $\sqrt{n}$-consistent semiparametric estimator exists. In the intermediate case, where the overlap plot shows some mass near the corners, but where the height of the density at 0 or 1 is nonetheless zero, $\sqrt{n}$-consistent estimators may or may not be available.[42]

To appreciate the problems with applying standard asymptotics to the semiparametric estimators studied here in situations with propensity scores near the boundaries, we turn now to a sequence of DGPs indexed by κ and inspired by the Proposition. Let the DGP be given by equations (3.10) to (3.12), with X_i, e_i, and u_i each distributed mutually independent and standard normal, with $\gamma = \delta = \psi = \eta = 0$, with κ ranging from 0.25 to 1.75. This DGP has homogeneous treatment effects, homoscedastic residuals of variance 1, and probability of treatment equal to 0.5.

For this DGP, $\gamma = 0$ and IWP2 for TOT is finite sample unbiased, but inefficient. The efficient estimator is the coefficient on treatment in a regression of the outcome on

combination of the vector X_i plays the role of the scalar X_i considered here. Suppose the linear combination has distribution function $G(\cdot)$ and density function $g(\cdot)$. Then the density for the propensity score is as is given in the display, with $\kappa = 1$. Note as well that the density of the propensity score among the treated and control is given by $f_{p(X_i)|T_i=1}(q) = \frac{q}{p} f_{p(X_i)}(q)$ and $f_{p(X_i)|T_i=0}(q) = \frac{1-q}{p} f_{p(X_i)}(q)$, respectively.

[41] Because the behavior of the density at the boundaries is the object of primary interest, it is best to avoid standard kernel density routines in favor of histograms or local linear density estimation (see McCrary (2008) for references).

[42] As the proposition above clarifies, $\sqrt{n}$-consistency is available, despite mass near the corners, when the covariate and treatment equation residuals are distributed standard normal. It is not yet known whether $\sqrt{n}$-consistency is always attainable when there is mass near the corners, but zero height to the density function of $p(X_i)$ in the corners.

a constant and the treatment indicator. It is thus easy to show that the Cramér-Rao bound is 4, regardless of the value of κ. When the SEB is close to the Cramér-Rao bound, there is little cost to using a semiparametric estimator. When there is quite good overlap, such as $\kappa = 0.25$, the SEB is in fact scarcely larger than 4 and there is little cost associated with avoiding parametric assumptions on the outcome equation. However, as problems with overlap worsen, the discrepancy between the SEB and the Cramér-Rao bound diverges. The cost of avoiding parametric assumptions on the outcome equation thus becomes prohibitive as κ increases in magnitude.

To convey a sense of the way in which an infinite SEB would manifest itself in an actual data set, Figure 3.2 shows the evolution of the overlap plot as κ increases. When $\kappa = 1$, the conditional densities are straight lines akin to a supply-demand graph from an undergraduate textbook. For $\kappa < 1$, the values of the conditional densities at the corners are zero. For $\kappa > 1$, the values of the conditional densities at the corners are positive and grow in height as κ increases.

Applying standard asymptotics to this sequence of DGPs suggests that, for $\kappa < 1$, IPW2 and pair matching estimates of the TOT have normalized large sample variances of

$$\begin{aligned}
&(3.16) \qquad & nV_{IPW2} &= \frac{1}{p} + \frac{1}{p} E\left[\frac{p(X_i)^2}{p(1-p(X_i))}\right] > 4, \\
&(3.17) \qquad & nV_{PM} &= nV_{IPW2} + \frac{1}{2}\left(1 + \frac{1}{p} E\left[\frac{p(X_i)}{1-p(X_i)}\right]\right) > 4 + \frac{3}{2}.
\end{aligned}$$

The variance expressions are close to 4 and 4+3/2 for moderate values of κ but are much larger for large values of κ.[43] Indeed, the Proposition implies that both nV_{IPW2} and nV_{PM} diverge as κ approaches 1.[44]

We next examine the accuracy of these large sample predictions by estimating the variance of IPW2 and pair matching for each value of κ.[45] Figure 3.3 presents the estimated standard deviation of these estimators as a function of κ and show that the quality of the large sample predictions depends powerfully on the value of κ.[46] For example, for κ below 0.7, the large sample predicted variances are generally accurate, particularly for IPW2. However, for $\kappa = 0.9$, the large sample predicted variances are markedly above the empirical variances for both estimators and the discrepancy grows rapidly as κ approaches 1, with the large sample variances diverging despite modest empirical variances. Roughly speaking, viewed as a function of κ, the standard deviations of IPW2 and pair matching are both linear to the right of $\kappa = 0.7$, with different slopes. The pattern of the variances

[43] The inequalities follow from Jensen's inequality and from the fact that $p = 0.5$ for these DGPs.

[44] The percent increase of nV_{PM} over nV_{IPW2} is between 37.5 percent (when $\kappa = 0$) and 25 percent (when κ approaches 1) and declines monotonically in the magnitude of κ.

[45] We use 2,000 replications.

[46] Interestingly, large sample predictions appear much more accurate for IPW2 than for pair matching.

is consistent with what would be expected if the variance of pair matching and IPW2 were proportional to the inverse of $n^{c_1+c_2\kappa}$, with possibly different coefficients c_1 and c_2 for the two estimators. Under this functional form restriction on the variances, it is possible to estimate c_1 and c_2 using regression. Define $Y_{g\kappa}$ as $\ln(\hat{V}_{100}/\hat{V}_{500})/\ln(5)$ for $g = 1$ and as $\ln(\hat{V}_{500}/\hat{V}_{1000})/\ln(2)$ for $g = 2$, where $\hat{V}_n$ is the estimated variance for sample size n. Then note that under the functional form restriction on the variances, $Y_{g\kappa} \approx c_1 + c_2\kappa$. Thus, a simple method for estimating c_1 and c_2 is a regression of $Y_{g\kappa}$ on a constant and κ.[47] For both IPW2 and pair matching, we have 26 observations on $Y_{g\kappa}$, 13 for $g = 1$ and 13 for $g = 2$. For IPW2, the regression described has an R-squared of 0.93 and constant and slope coefficients (standard errors) of 1.19 (0.02) and -0.39 (0.02), respectively. For pair matching, the R-squared is 0.94 and the constant and slope coefficients (standard errors) are 1.15 (0.02) and -0.33 (0.02), respectively. We report these results not because we believe that the scaling on the variance is of the form $n^{c_1+c_2\kappa}$, but to emphasize our sense that the correct scaling is a smooth function of κ.[48]

These results create a strong impression that the asymptotic sequences used in the large sample literature may be accurate in settings of good overlap, but are likely inaccurate in settings of poor overlap. The performance of these two estimators does not seem to degrade discontinuously when κ exceeds one, but rather seems to degrade smoothly as κ approaches one.

Failure to satisfy the strict overlap assumption can also lead to bias in semiparametric estimators. The sign and magnitude of the bias will be difficult to infer in empirical work. Consider again the model in equations (3.10) to (3.12), with $\eta = 0$, $\beta = 1$, $\gamma = 0$, and $m(q) = q$. In this DGP, when $\delta = 0$, IPW2 for ATE is finite sample unbiased regardless of the value of κ. When $\delta = 1$, the treatment effect is positively correlated with the propensity score and IPW2 for ATE may be biased. Similarly, when $\delta = -1$, the treatment effect is negatively correlated with the propensity score and IPW2 for ATE may be biased.

Figure 3.4 shows the bias of IPW2 for ATE as a function of κ for $\delta = 0$, $\delta = 1$, and $\delta = -1$. The figure confirms that when $\delta = 0$, large values of κ do not compromise the unbiasedness of IPW2. However, when $\delta \neq 0$, large values of κ lead to bias. Importantly, when overlap is good, IPW2 is unbiased regardless of the value of δ.

[47] Weights improve power since the outcome is more variable for $g = 2$ than for $g = 1$. In particular, the delta method and Wishart approximations suggest that the standard deviation of the outcome is approximately $\sqrt{4/2000}/\ln(5)$ for $g = 1$ and $\sqrt{4/2000}/\ln(2)$ for $g = 2$.

[48] However, it is interesting to note that these regressions can be viewed as minimum chi-square estimates (Ruud 2000). This approach allows for a statistical test of the functional form restriction that the variances are proportional to the inverse of $n^{c_1+c_2\kappa}$. The test takes the form of the minimized quadratic form, or in this case the (weighted) sum of squared residuals. The test statistic is distributed chi-square with 24 degrees of freedom. For IPW2, this test statistic is 28.4 and for pair matching it is 20.5 (95 percent critical value 36.4).

V Results: Boundary Problems

In order to focus attention on how the estimators perform when the strict overlap condition is close to being violated, we turn now to an analysis of a DGP that is a minor modification of that described in Section III, above. Instead of generating u_i as independent draws from the standard Cauchy distribution, we generate u_i as independent draws from the standard normal distribution. We manipulate the parameters η and κ in the treatment equation (3.10) to mimic three designs from the influential study of Frölich (2004). These parameters are summarized below:

Design	η	κ	**Treated-to-Control Ratio**
A	0	0.95	1:1
B	0.3	-0.8	3:2
C	-0.3	0.8	2:3

Figure 3.1B shows the overlap plot implied by these designs. Each of these designs is consistent with standard overlap, but none are consistent with strict overlap. This figure shows that having many control observations per treated observations does not imply the validity of the strict overlap condition. For example, design A is closer to violating the strict overlap assumption than design C is, even though the ratio of treated to control observation is higher in the former than in the latter.

A Simulation Results with Boundary Problems

In Table 3.3 we explore estimator performance in DGPs that are close to violating the strict overlap condition. The structure of the table is identical to that of Table 3.1, but the DGPs correspond to the Normal-Normal model, rather than the Normal-Cauchy model.

The results in the table support several conclusions. First, when $n = 100$, nearly all estimators are biased in all settings. The exceptions are the control function and double robust estimators in homogeneous treatment effect settings. These two estimators impose parametric assumptions on the outcome equation. This allows for extrapolation from the region of common support to the region over which there are treated observations but no controls. Second, although reweighting estimators are biased with $n = 100$, they become unbiased when $n = 500$. This raises the possibility that, for a good finite sample performance, a larger sample size is required for DGPs with poor overlap that is nonetheless technically sufficient to guarantee $\sqrt{n}$-consistency. Third, aside from pair matching, the magnitude of the bias of matching estimators is between two and five times that of the reweighting estimators, and they remain biased even for $n = 500$. Pair matching is biased for $n = 100$ and nearly unbiased for $n = 500$.[49] The third and fourth panel show that

[49] The sign of the bias of the TOT depends on the shape of the outcome equation. An outcome equation

the variance of all estimators is much higher than in the case in which we satisfy the strict overlap assumption, even though none of the designs imply an infinite SEB. For all estimators we reject the null that the variance equals the SEB in every setting. Contrary to the case of strict overlap analyzed in the preceding section, this holds true for both $n = 100$ and $n = 500$. The variance of all the estimators is on average below the SEB.

In sum, in settings of poor overlap, semiparametric estimators of average treatment effects do not perform well for $n = 100$. Once $n = 500$, the pair matching, reweighting, double robust, and control function estimators show acceptable bias, but only IPW1 has bias small enough that we fail to reject the null of zero bias. The variance of semiparametric estimators is hard to assess in settings of poor overlap, since neither the SEB nor other large sample approximations form acceptable benchmarks. However, considering both bias and variance and performance for $n = 100$ and $n = 500$, the best estimators in settings with poor overlap appear to be IPW2, IPW3, and double robust.

B Trimming

In many empirical applications, researchers encounter a subset of observations whose propensity scores do not have common support. Such a finding is expected when the strict overlap condition is violated, although it can also occur in finite samples when strict overlap is satisfied in the population. Confronted by lack of common support, many researchers resort to trimming rules. These sample selection rules involve dropping individuals from the treatment group who have no counterparts in the control group with similar propensity scores (for TOT).[50] Trimming aims at ensuring validity of the common support assumption in the subset of observations that are not trimmed. See Heckman, Ichimura and Todd (1998a), Smith and Todd (2005), and Crump, Hotz, Imbens and Mitnik (2007a) for discussion. There are several trimming methods that have been proposed in the literature. Little is known about their effect on the performance of semiparametric estimators.

As noted by Heckman et al. (1998a), reweighting and matching *at best* correct for bias for the subsample of individuals whose propensity scores have common support. For this reason, trimming is only expected to work in situations of treatment effect homogeneity, simply because the treatment effect can be estimated anywhere on the support of the propensity score. Dropping observations will make the estimator more inefficient but the bias is expected to decrease because we will be estimating the counterfactual mean only in

that is increasing (decreasing) in the propensity score like curve 1 (curve 2) implies that the bias will be more positive (negative) the closer we are to violating the strict overlap condition because we have too many treated observations and too few controls at the right end of the distribution of the propensity score (see appendix). The bias is not related to the overall ratio of treated per controls units in the sample. The bias of all the estimators tends to be of the same order of magnitude in the three designs.

[50] Trimming in the case of estimation of the ATE is similar, but individuals from both the treatment and the control group are deleted.

regions in which both treated and control units are available. However, if the treatment effect is heterogeneous, and more importantly, if the heterogeneity occurs precisely in the part of the support for which we do not have both treated and control observations, then trimming will not be a solution.[51] In those type of situations the researcher might need to redefine the estimand (see Crump et al. 2006) paying a cost in terms of having a result with less external validity or resort to fully parametric models—which will typically only be effective if the full parametric model is correctly specified.

We analyze the effectiveness of the four trimming rules reviewed in Crump et al. (2006):

1. Let $D_i^{ATE} = \mathbf{1}(\widehat{a} < \hat{p}(X_i) < \widehat{b})$ and $D_i^{TOT} = \mathbf{1}(\hat{p}(X_i) < \widehat{b})$ setting $\widehat{b}$ to be the kth largest propensity score in the control group and $\widehat{a}$ to be the kth smallest propensity score in the treatment group. Then we compute the estimators on the subsample for which $D_i^{TOT} = 1$ (or $D_i^{ATE} = 1$). This rule was proposed by Dehejia and Wahba (1999).

2. Heckman et al. (1996, 1998) and Heckman, Ichimura, Smith and Todd (1998) propose discarding observations for which the conditional density of the propensity score is below some threshold. Let $D_{0i}\left(c\right) = \mathbf{1}(\widehat{f}_{\hat{p}(X_i)|T_i=0} < c)$ and $D_{1i}\left(c\right) = \mathbf{1}(\widehat{f}_{\hat{p}(X_i)|T_i=1} < c\ )$ where c is a tuning parameter, and $\widehat{f}_{\hat{p}(X_i)|T_i=1}$ and $\widehat{f}_{\hat{p}(X_i)|T_i=0}$ are kernel density estimates (with Silverman's rule as a bandwidth selector). Then, following Smith and Todd (2005), fix a quantile $q = 0.02$ and consider the J observations with positive densities $\widehat{f}_{\hat{p}(X_i)|T_i=1}$ and $\widehat{f}_{\hat{p}(X_i)|T_i=0}$. Rank all the values of $\widehat{f}_{\hat{p}(X_i)|T_i=1}$ and $\widehat{f}_{\hat{p}(X_i)|T_i=0}$ and drop units with a density less than or equal to c_q, where c_q is the largest real number such that $\frac{1}{2J}\sum_{i=1}^{J}\left[D_{0i}\left(c_q\right) + D_{1i}\left(c_q\right)\right] \leq q$ for the ATE. For the TOT we can proceed in a similar fashion but only using $\widehat{f}_{\hat{p}(X_i)|T_i=1}$.

3. Ho, Imai, King and Stuart (2007) define the common support region as the convex hull of the propensity scores used by pair matching.

4. Finally, Crump et al. (2007a) propose discarding all units with an estimated propensity score outside the interval $[0.1, 0.9]$ for the ATE and $[0, 0.9]$ for the TOT.

In Table 3.4 we study whether, in a DGP that is close to violating the strict overlap assumption, trimming succeeds in reducing the bias. As expected, the double robust and control function estimators stay unbiased in homogeneous settings, but trimming increases the bias of those estimators in heterogeneous settings. Trimming rules 1 and 4 seem to lead to unbiasedness of reweighting and pair matching in settings with a homogeneous treatment effect. These rules also reduce the bias of all the matching estimators. Trimming rule 3 only works with pair matching and to a lesser extent with ridge matching. Trimming rule 2 does not seem to work with $n = 100$. This may not be surprising since this rule requires estimating the conditional density of the propensity score with very few observations.

[51] An alternative to trimming is to compute bounds for the treatment effects. This possibility was advocated by Lechner (2001) in the context of matching estimators of treatment effects.

In Table 3.5 we present the effect of trimming on the variance of the estimators. Rules 1 and 4 reduce the variance of IPW estimators and of local linear and ridge matching. Surprisingly, the variance of the other matching estimators seem to be basically unaffected by any of the trimming rules.

VI Reconciliation with Previous Literature

Previous literature has analyzed the finite sample properties of semiparametric estimators of treatment effects in situations with homogeneous treatment effects and homoscedastic outcome error terms. Frölich (2004) compares the finite sample performance of several matching estimators based on the propensity score and the IPW1 estimator, in simulation settings that highlight the interactions between the non-linearities of the outcome equation and different degrees of overlapping density mass between treated and control observations. Zhao (2004) contrasts the performance of propensity score matching and covariate matching methods using a simulation study that varies the degree of selection on observed variables and the correlation between covariates, the outcome and the treatment indicator. Lunceford and Davidian (2004) compare reweighting, double robust and blocking estimators via a simulation analysis that assesses the effect of different degrees of correlation between regressors in the outcome and treatment equation, emphasizing situations in which there is misspecification of various types. Finally, Freedman and Berk (n.d.) study the costs and benefits of using a semiparametric estimator such as double robust rather than fully parametric estimators. Some of our conclusions are at odds with findings in this previous literature.

Frölich (2004) is the most similar to the present work in terms of the estimators considered and the simulation studies performed. The study reaches the conclusion that ridge matching is often the estimator with smallest MSE. As we showed in sections IV and VI, ridge matching does relatively well, especially among the matching estimators, in terms of variance but was only unbiased in a situation with good overlap and for a moderate sample size of 500 observations. A surprising conclusion of Frölich is that reweighting estimators perform very poorly, usually presenting a larger MSE than pair matching. Even more surprising is that the relative MSE of reweighting does not decline with the sample size in Frölich's DGPs.

Several differences between Frölich and this study account for the discrepancies in the conclusions. First, Frölich only considers the performance of IPW1. As noted above, in many DGPs IPW1 is substantially more variable than IPW2 and IPW3. Second, Frölich computes all estimators using the *true* propensity score instead of the *estimated* propensity score. As noted by Hirano et al. (2003), reweighting performs better when the estimated propensity score is used. Third, Frölich's study evaluates estimators by how well they perform in estimation of a non–standard estimand: the counterfactual mean outcome for the control group, as opposed to the TOT or the ATE, which would be more conventional estimands of interest.[52] However, the most important difference between our study and Frölich is that Frölich's DGPs violate the strict overlap condition and are quite close to exhibiting an infinite SEB. As we have shown, in such a setting nearly all semiparametric estimators acquire difficulties with bias, and MSE may not be the best metric for performance. In particular, our own simulation evidence suggests that in situations with poor overlap, the most biased estimators are also the least variable. For example, in the Normal-Normal model for both $n = 100$ and $n = 500$ and for all four settings, kth nearest-neighbor matching exhibits *both* the worst bias and the best variance of any estimator (Table 3.3).

Figure 3.5 presents overlap plots for the finite sample papers reviewed above. The figure is to be compared to Figure 3.2, which shows the evolution of the conditional densities of the propensity score as we increase κ in our Normal-Normal model. The design of Frölich displayed in Figure 3.5 is quite similar to that of the Normal-Normal model when $\eta = 0$ and $\kappa = 0.9$. As noted above, although the SEB is technically speaking finite, this is a situation in which strict overlap is violated and asymptotic approximations may be poor. This is confirmed by our own simulation results. For the Normal-Normal model, IPW1 and pair matching exhibit similar bias, but IPW1 has notably higher variance. In a MSE metric, pair matching is superior to IPW1 for these DGPs, for both $n = 100$ and $n = 500$.

Failure of strict overlap is also characteristic of the DGPs studied by Zhao. Figure 3.5 displays the conditional densities of the propensity score for the first of the DGPs he uses in his simulation study. Inspection of the figure indicates an extraordinarily serious failure of strict overlap. In this DGP, it is further true that the SEB is infinite: Zhao's DGP is the same as our Normal-Normal model (discussed in Section V) with $\kappa = 2.8$. That the SEB is infinite thus follows from the Proposition of Section IV.

We turn next to the analyses of Lunceford and Davidian (2004) and Freedman and Berk (n.d.). Lunceford and Davidian focus on the performance of reweighting and double robust estimators for ATE. One of their principal conclusions is that a double robust estimator performs well in a broader class of DGPs than IPW estimators. Freedman and Berk also consider a variant of the double robust estimator, but focus on the comparison with parametric OLS models, which are of course best in the sense of being minimum variance among unbiased estimators. Perhaps influenced by this hopeful benchmark, Freedman and Berk express reservations about the utility of reweighting estimators of average treatment effects.

Figure 3.5 displays a representative overlap plot for the DGPs used by Lunceford and Davidian and Freedman and Berk. The figure reveals that the DGPs studied by Lunceford

[52] This tends to amplify MSE differences, since $MSE(\hat{\theta}) = MSE(\hat{E}[Y_i(0)|T_i = 1]) + V(\overline{Y}_1)$.

and Davidian and Freedman and Berk are similar to those of Frölich and Zhao in that they violate strict overlap. Indeed, the displayed DGP of Freedman and Berk is further associated with an infinite SEB. We disagree with Freedman and Berk's characterization of this DGP as "favorable to weighting". In DGPs of the type studied by Freedman and Berk, none of the semiparametric estimators studied here will be effective. We find uncontroversial the overarching point of Freedman and Berk that (correctly specified) parametric models outperform semiparametric estimators.

One further aspect deserves mention in understanding the findings here and in Lunceford and Davidian (2004) and Freedman and Berk (n.d.). To fix ideas, consider DGPs where the outcome equation is linear in the propensity score. In this case, the critique of weighting in Freedman and Berk (n.d.) corresponds closely to what Deaton (1997) has referred to as "the econometric critique of weighting." Deaton is interested in OLS estimation of a parametric relationship between the outcome and a set of covariates, where there are departures from pure random sampling, such as cluster sampling. Specifically he considers the case where $Y_i = X_i\beta + \epsilon_i$ under the usual ideal conditions that deliver consistency of OLS. In the case of stratified or cluster sampling, simple OLS regression continues to have desirable properties. Weighted OLS regression, where the weights reflect, say, the probability of being included in the sample continue to yield consistent estimates of β, but are merely less efficient than the unweighted estimator, a conclusion consistent with the conclusion in Freedman and Berk (n.d.) that "weighting is likely to increase random error in the estimates."

Now consider the case where there is some limited heterogeneity in β. In particular, suppose that $Y_{is} = X_{is}\beta_s + \epsilon_{is}$, where the effect of the covariate on the outcome is no longer fixed but differs by stratum, s. OLS will no longer be consistent for the stratum weighted average of β_s regardless of whether or not weights are used in the regression.[53]Given his context, Deaton suggests that the weighted regression might be used in a formal or informal specification test – the weighted and unweighted regression estimates should be "close" under the null that β is the same for each strata.

These points can be seen most clearly in the following simple example. Let ϵ_i, ν_i, and u_i be standard normal. Define $X_i = \kappa \times u_i$ and consider the following DGP, for a sample size $N = 100$, which is a simplified version of the one we have studied above (equations (3.10) to (3.12)):

$$T_i^\star = X_i + \nu_i;\ T_i = 1\left(T_i^\star > 0\right);\ Y_i = T_i + X_i + \epsilon_i.$$

As before different values of $0 < \kappa < 1$ correspond to cases of strict overlap. In this

[53] A consistent estimator of the strata weighted coefficient can be obtained by performing separate regressions for each stratum and then appropriately weighting the estimated coefficients.

context, three estimators considered by Lunceford and Davidian (2004) and Freedman and Berk (n.d.) are OLS (of the correctly) specified model, double robust, and IPW2 where: (i) OLS is just the simple unweighted regression of the outcome on T and X; (ii) double robust is a weighted regression of the outcome on T and X where $1/\widehat{p}$ and $1/(1-\widehat{p})$ are the weights for treated and untreated observations, respectively and $\widehat{p}$ is the predicted probability of treatment from a simple probit of T on X;[54](iii) and IPW2 is identical to double robust except that it requires a different weighted regression, using the same outcome variable and weights but dropping the covariate.

When strict overlap is satisfied and the treatment effect is homogenous, all three estimators are consistent for ATE and in each case the variance of the estimator depends on the value of κ. Figure 3.6 displays the standard deviation of the three estimators for various values of κ that resulted from a simulation of the above DGP with 20,000 replications. Consistent with the findings of Lunceford and Davidian (2004) and Freedman and Berk (n.d.), the double robust estimator outperforms IPW2 and correctly specified OLS outperforms both. We take away a couple of key points from this demonstration:

1. The extent of the superior performance of double robust relative to IPW depends crucially on κ. Higher values of κ are associated with worse performance of all estimators, with the degradation in IPW2 being the most severe.
2. Consistent with the analysis in Freedman and Berk (n.d.) and Lunceford and Davidian (2004), however, it is also the case that weighting regressions merely "adds noise" to the estimate when the parametric model is correctly specified.
3. It is important to stress that these results refer only to the variance of the estimator and only when strict overlap is satisfied. All three estimators are consistent. For values of $\kappa > 1$ the IPW2 estimator is not properly identified. When $\kappa = 2$ for example, the mean value of the IPW estimates in 20,000 simulations was fully 100% large than its true value.

VII Conclusion

In this paper, we assess the finite sample properties of semiparametric estimators of treatment effects using simulated cross-sectional data sets of size 100 and 500. The estimators we consider are semiparametric in the sense that only the treatment assignment process is parametrically modeled. This perspective on estimation encompasses several popular approaches including reweighting, double robust, control function, and matching,

[54] Previously we implemented the double robust estimator by including the propensity score instead of the covariate X. We do it this way here for ease of exposition.

but rules out maximum likelihood estimation and estimators based on parametric assumptions on the relationship between the outcome of interest and predicting variables. The semiparametric estimators we consider are popular in the empirical literature.

The simulation evidence suggests that when there is good overlap in the distribution of propensity scores for treatment and control units, reweighting estimators are preferred on bias grounds and attain the semiparametric efficiency bound, even for samples of size 100. The double robust estimator can be thought of as regression adjusted reweighting and performs slightly worse than reweighting when there is good overlap, but slightly better when there is poor overlap. Control function estimators perform well only for samples of size 500. Matching estimators perform worse than reweighting if preferences over bias and variance are lexicographic and if good performance for $n = 100$ is required. If there is enough data, then local linear or ridge matching may be competitive with reweighting. The difficulty of the more complicated matching estimators is potentially related to the difficulty of accurate finite sample selection of tuning parameters.[55]

When overlap in the distribution of propensity scores for treatment and control units is close to failing, the semiparametric estimators studied here do not perform well. This difficulty can be inferred from the available large sample results in the literature (Hirano et al. 2003, Abadie and Imbens 2006, Khan and Tamer 2007). We also show that the standard asymptotic arguments used in the large sample literature provide poor approximations to finite sample performance in cases of near failure of overlap. However, our qualitative conclusion is the same as that reached by Khan and Tamer (2007), who note that the semiparametric estimators considered here are on a sound footing only when there is strict overlap in the distribution of propensity scores (see Section II).

In empirical applications, economists confronting problems with overlap often resort to trimming schemes, in which some of the data are discarded after estimation of the propensity score. We simulate the performance of the estimators studied in conjunction with four trimming rules discussed in the literature. None of these procedures yield good performance unless there is homogeneity in treatment effects along the dimension of the propensity score.

What is then to be done in empirical work in which problems with overlap are suspected? First, to assess the quality of overlap, we recommend a careful examination of

[55] If preferences over bias and variance are not lexicographic, then some of the biased matching estimators may be preferred to reweighting. We caution, however, that the data generating processes we consider may not represent those facing the economist in empirical applications. In empirical applications, the bias could be of lesser, or greater, magnitude than suggested here, in which case the economist's preference ranking over estimators could be different than that suggested by a literal interpretation of the simulation evidence. Our own preferences over bias and variance lean towards lexicographic because we have a taste for estimators that minimize the maximum risk over possible data generating processes.

the overlap plot, possibly focused on histograms and possibly involving smoothing using local linear density estimation. Second, if overlap indeed appears to be a problem, we recommend analysis of subsamples based on *covariates* to determine if there are subsamples with good overlap. For example, in some settings, it could occur that problems with overlap stem from one particular subpopulation that is not of particular interest. Analyzing subsamples based on covariates is likely to work better than analyzing subsamples based on quantiles of an estimated propensity score. Third, if there is no obvious subpopulation displaying good overlap, we recommend that the economist consider parametric assumptions on the outcome equation. Semiparametric estimators work well in this context when there is good overlap. When overlap is poor, however, these estimators are highly variable, biased, and subject to nonstandard asymptotics. In settings with poor overlap, the motivation for semiparametric estimation is poor and the most effective methods are likely parametric approaches such as those commonly employed in the older Oaxaca (1973) and Blinder (1973) (1973) literature.

Table 3.1: Bias and Variance of the Estimated Treatment Effect on the Treated (TOT)

Normal-Cauchy Model

	Sample Size	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
A. Simulated Root	100	I. Homog.-Homosk.	5.2	25.1*	42.5*	35.5*	39.5*	39.0*	39.5*	9.2*	12.8*	3.2	4.0	4.9*	2.5	3.0
Mean Squared Bias		II. Heterog.-Homosk.	2.8	44.4*	41.9*	34.9*	39.2*	38.4*	39.3*	7.8*	12.0*	1.4	2.4	3.1	5.0*	2.0
(x 1000)		III. Homog.-Heterosk.	5.0	26.9*	35.3*	26.8*	28.6*	11.2*	13.0*	8.5*	10.7*	3.5	4.1	4.8	4.1	3.8
		IV. Heterog.-Heterosk.	3.3	42.6*	34.0*	25.0*	26.9*	10.9*	13.2*	6.1*	8.3*	2.2	2.1	2.5	6.5*	2.3
		All	4.2	35.8*	38.6*	30.9*	34.0*	28.5*	29.4*	8.0*	11.1*	2.7	3.3	4.0*	4.8*	2.9
		F-stat (no bias)	37.7	3505.5	5276.1	3309.3	4163.6	2144.7	2384.1	183.4	379.0	19.3	30.9	46.0	58.5	18.8
		[p-value]	[0.037]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.734]	[0.155]	[0.004]	[0.000]	[0.760]
	500	I. Homog.-Homosk.	2.4	7.6*	36.0*	30.2*	33.4*	2.9	2.9	3.1	4.9*	2.3	2.3	2.2	1.9	1.7
		II. Heterog.-Homosk.	2.2	7.7*	32.9*	27.8*	31.1*	2.1	1.8	1.7	2.9	1.9	1.9	1.8	2.6	2.0
		III. Homog.-Heterosk.	2.4	6.8*	23.2*	16.2*	17.9*	2.4	2.4	2.6	3.2	2.3	2.3	2.2	2.3	2.4
		IV. Heterog.-Heterosk.	2.3	9.8*	25.0*	17.5*	19.4*	2.4	2.7	3.2	4.3	2.2	2.2	2.4	1.6	2.2
		All	2.3	8.0*	29.8*	23.7*	26.4*	2.5	2.5	2.7	3.9*	2.2	2.2	2.2	2.2	2.1
		F-stat (no bias)	12.1	182.4	3094.5	2043.2	2548.8	17.6	18.1	20.6	45.8	13.7	13.9	13.5	14.0	11.9
		[p-value]	[0.979]	[0.000]	[0.000]	[0.000]	[0.000]	[0.824]	[0.800]	[0.662]	[0.005]	[0.953]	[0.949]	[0.958]	[0.946]	[0.981]
B. Simulated	100	I. Homog.-Homosk.	103.3*	68.9*	53.5*	56.2*	54.5*	82.4*	78.5*	67.4*	64.0*	72.2*	65.8*	65.5*	65.8*	89.3*
Average Variance		II. Heterog.-Homosk.	106.2*	74.9*	54.9*	57.6*	55.7*	84.1*	80.0*	68.9*	65.4*	73.0*	66.9*	66.8*	67.2*	91.4*
(x 1000)		III. Homog.-Heterosk.	119.0*	110.1*	108.0	109.0	108.8	113.7*	113.0*	111.7	111.1	118.4*	112.6*	112.0*	117.6*	115.9*
		IV. Heterog.-Heterosk.	118.1*	113.7*	107.0*	107.8*	107.6*	112.9*	112.1	110.5	110.0	117.6*	111.7	110.9	117.0*	115.5*
		Average (V-SEB)/SEB	0.376	0.083	-0.078	-0.052	-0.067	0.180	0.145	0.050	0.020	0.116	0.040	0.035	0.064	0.247
		F-stat (V = SEB)	418.6	45.4	73.0	30.1	54.5	158.5	116.4	17.5	4.0	60.1	9.7	8.3	19.5	243.7
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	500	I. Homog.-Homosk.	21.1*	12.8	11.2*	11.2*	11.1*	13.5*	12.9	13.0	12.6	13.3*	12.6	12.6	12.6	12.9
		II. Heterog.-Homosk.	21.5*	13.3*	11.6*	11.4*	11.3*	13.6*	13.2*	13.2*	12.8	13.7*	12.9	12.9	13.0	13.3*
		III. Homog.-Heterosk.	24.1*	21.8	21.2	21.5	21.5	22.2	22.0	22.0	21.8	22.6	21.9	21.8	21.9	21.9
		IV. Heterog.-Heterosk.	24.1*	22.3	21.5	21.8	21.8	22.5	22.3	22.3	22.2	23.0	22.3	22.2	22.4	22.3
		Average (V-SEB)/SEB	0.398	0.021	-0.060	-0.055	-0.062	0.048	0.025	0.027	0.009	0.059	0.012	0.010	0.014	0.025
		F-stat (V = SEB)	88.8	1.6	11.3	9.5	12.3	3.3	1.4	1.5	0.7	3.9	0.8	0.7	0.9	1.6
		[p-value]	[0.000]	[0.032]	[0.000]	[0.000]	[0.000]	[0.000]	[0.097]	[0.064]	[0.817]	[0.000]	[0.793]	[0.817]	[0.584]	[0.032]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.2: Bias and Variance of the Estimated Treatment Effect on the Treated (TOT) under Misspecification

Misspecification of the Propensity Score in the Normal-Cauchy Model (sample size 100)

	Misspec. Type	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
A. Simulated Root	Xs	I. Homog.-Homosk.	127.2*	123.2*	142.4*	141.1*	142.6*	155.4*	153.3*	128.7*	130.4*	125.4*	125.6*	125.9*	125.2*	124.9*
Mean Squared Bias		II. Heterog.-Homosk.	126.4*	120.9*	143.1*	141.6*	143.1*	156.3*	154.0*	128.8*	130.5*	125.3*	125.7*	126.2*	123.7*	125.1*
(x 1000)		III. Homog.-Heterosk.	125.5*	122.7*	140.8*	139.3*	140.6*	131.4*	132.3*	127.7*	129.4*	124.6*	124.9*	125.2*	124.5*	124.2*
		IV. Heterog.-Heterosk.	126.1*	120.1*	142.2*	140.7*	142.1*	131.8*	132.8*	128.4*	130.4*	125.5*	125.9*	126.2*	123.2*	124.9*
		All	126.3*	121.7*	142.1*	140.7*	142.1*	144.2*	143.5*	128.4*	130.2*	125.2*	125.5*	125.9*	124.1*	124.8*
		F-stat (no bias)	36397.5	46067.5	69490.6	67500.6	70217.2	56723.0	58158.0	51241.9	54298.3	49436.3	50884.0	51261.2	48656.0	43699.7
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	Dist of u.	I. Homog.-Homosk.	5.4	27.3*	42.6*	37.0*	40.7*	39.5*	41.1*	9.5*	12.9*	22.6*	7.6*	6.5*	2.4	2.1
		II. Heterog.-Homosk.	3.0	47.3*	42.3*	36.3*	40.1*	39.1*	40.3*	8.4*	12.1*	19.5*	4.7*	3.9	7.3*	4.0
		III. Homog.-Heterosk.	5.4	28.7*	36.1*	29.4*	31.8*	12.8*	15.5*	9.3*	11.5*	21.8*	7.9*	7.1*	4.0	3.8
		IV. Heterog.-Heterosk.	3.6	46.1*	34.7*	27.9*	30.3*	12.7*	15.8*	7.1*	9.2*	21.0*	7.9*	6.1*	8.6*	3.1
		All	4.5	38.5*	39.1*	32.9*	36.0*	29.2*	30.8*	8.6*	11.5*	21.3*	7.2*	6.0*	6.1*	3.3
		F-stat (no bias)	42.4	4067.8	5378.4	3728.1	4597.8	2209.3	2587.9	212.0	402.1	744.2	130.1	98.5	99.0	23.8
		[p-value]	[0.012]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.474]
B. Simulated	Xs	I. Homog.-Homosk.	87.4*	55.0*	48.3*	48.8*	47.7*	67.2*	64.3*	55.8*	53.6*	54.9*	53.0*	52.9*	53.8*	64.9*
Average Variance		II. Heterog.-Homosk.	88.0*	59.0*	49.9*	50.2*	49.1*	68.8*	65.6*	57.0*	54.8*	56.0*	54.3*	54.2*	55.0*	66.3*
(x 1000)		III. Homog.-Heterosk.	132.7*	121.7*	121.4*	121.5*	121.3*	125.5*	124.8*	123.2*	122.5*	124.4*	122.6*	122.5*	128.3*	126.3*
		IV. Heterog.-Heterosk.	132.5*	124.6*	121.8*	121.9*	121.7*	125.6*	125.2*	123.5*	123.0*	124.4*	122.9*	122.8*	128.7*	126.6*
		Average (V-SEB)/SEB	-0.066	-0.266	-0.314	-0.311	-0.318	-0.197	-0.216	-0.268	-0.282	-0.270	-0.285	-0.286	-0.261	-0.208
		F-stat (V = SEB)	36.0	1079.7	2031.3	1955.7	2150.8	460.2	580.3	1095.5	1323.2	1186.4	1387.9	1401.2	1258.8	512.1
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	Dist of u.	I. Homog.-Homosk.	103.1*	68.6*	53.1*	55.4*	53.9*	82.9*	78.9*	67.2*	63.8*	93.3*	72.2*	67.3*	68.5*	93.0*
		II. Heterog.-Homosk.	105.9*	74.9*	54.5*	56.7*	55.1*	84.1*	80.0*	68.5*	65.3*	93.8*	73.3*	68.9*	70.1*	96.5*
		III. Homog.-Heterosk.	118.7*	110.2*	107.5*	108.8	108.7	114.1*	113.9*	111.7*	111.2	138.9*	118.7*	114.8*	121.4*	118.7*
		IV. Heterog.-Heterosk.	117.8*	114.4*	106.4*	107.7*	107.4*	113.4*	113.0*	110.6	110.1	145.4*	118.5*	114.0*	121.4*	117.6*
		Average (V-SEB)/SEB	0.373	0.084	-0.084	-0.060	-0.073	0.184	0.150	0.047	0.019	0.383	0.118	0.063	0.105	0.293
		F-stat (V = SEB)	415.7	46.4	77.1	40.0	62.1	162.5	119.1	16.1	3.5	418.4	80.3	27.9	61.0	296.0
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.3: Bias and Variance of the Estimated Treatment Effect on the Treated (TOT)

Normal-Normal Model

	N	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
A. Simulated Root	100	I. Homog.-Homosk.	19.3*	45.1*	88.8*	70.5*	76.7*	54.2*	55.4*	26.9*	32.1*	15.8*	16.7*	16.8*	3.3	7.8*
Mean Squared Bias		II. Heterog.-Homosk.	16.5*	66.3*	87.4*	68.6*	75.1*	53.3*	55.0*	25.4*	30.4*	12.8*	14.8*	15.4*	16.9*	7.6*
(x 1000)		III. Homog.-Heterosk.	14.4*	43.4*	76.1*	55.4*	58.7*	21.6*	25.4*	24.4*	27.5*	15.2*	13.8*	13.9*	3.7	6.6
		IV. Heterog.-Heterosk.	14.6*	67.4*	74.8*	54.2*	57.9*	21.1*	24.3*	23.0*	26.7*	15.0*	12.4*	12.9*	19.5*	5.7
		All	16.3*	56.7*	82.0*	62.6*	67.7*	40.9*	42.8*	24.9*	29.3*	14.7*	14.5*	14.8*	13.2*	7.0*
		F-stat (no bias)	428.2	6163.6	21354.4	11216.0	13439.8	3071.5	3564.4	1385.7	1989.8	267.3	422.9	466.7	335.6	69.9
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	500	I. Homog.-Homosk.	7.8*	17.1*	72.1*	52.8*	58.4*	12.0*	14.2*	14.1*	17.4*	3.1	6.3	7.7*	4.0	6.4*
		II. Heterog.-Homosk.	5.9	16.4*	68.7*	50.1*	55.4*	9.8*	12.1*	11.5*	15.4*	4.7	4.1	4.5	8.5*	3.5
		III. Homog.-Heterosk.	4.1	15.0*	60.3*	36.9*	40.5*	7.3*	8.4*	11.0*	13.1*	4.4	3.8	5.4	2.7	3.5
		IV. Heterog.-Heterosk.	7.3	17.4*	60.7*	37.2*	40.7*	9.6*	9.9*	12.2*	14.1*	4.6	6.7	7.5*	6.4	5.0
		All	6.4*	16.5*	65.7*	44.9*	49.4*	9.8*	11.4*	12.3*	15.1*	4.3	5.4*	6.4*	5.8*	4.7
		F-stat (no bias)	52.8	487.7	12564.8	5392.8	6719.7	167.8	251.3	285.5	461.6	23.6	43.6	71.8	68.1	42.0
		[p-value]	[0.001]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.482]	[0.009]	[0.000]	[0.000]	[0.013]
B. Simulated	100	I. Homog.-Homosk.	139.6*	104.8*	58.0*	66.5*	64.9*	117.1*	108.8*	88.7*	84.6*	166.8*	98.1*	93.1*	92.7*	166.2*
Average Variance		II. Heterog.-Homosk.	143.6*	122.1*	59.2*	68.5*	66.7*	119.0*	111.4*	91.4*	87.1*	162.3*	101.5*	96.0*	96.5*	169.4*
(x 1000)		III. Homog.-Heterosk.	152.0*	141.2*	120.1*	127.9*	128.1*	140.8*	139.3*	133.4*	133.0*	207.5*	149.1*	142.2*	152.8*	162.2*
		IV. Heterog.-Heterosk.	151.7*	156.0*	120.0*	127.1*	126.8*	139.6*	138.6*	132.7*	132.1*	195.2*	149.0*	142.2*	153.6*	161.6*
		Average (V-SEB)/SEB	-0.280	-0.375	-0.570	-0.533	-0.537	-0.369	-0.392	-0.459	-0.472	-0.165	-0.404	-0.432	-0.410	-0.190
		F-stat (V = SEB)	6454.1	8872.2	46732.8	31680.5	33472.8	9986.0	11525.1	17199.1	18762.0	2701.3	11956.3	13919.1	12969.7	4169.5
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	500	I. Homog.-Homosk.	33.7*	21.4*	13.2*	15.0*	14.6*	22.1*	20.2*	19.8*	18.7*	39.8*	24.5*	21.0*	21.4*	21.9*
		II. Heterog.-Homosk.	33.9*	22.8*	13.2*	14.9*	14.5*	22.0*	20.1*	19.7*	18.7*	36.3*	24.2*	20.8*	21.4*	21.6*
		III. Homog.-Heterosk.	35.7*	29.8*	24.2*	26.9*	26.8*	30.6*	29.8*	28.7*	28.4*	44.3*	34.4*	30.6*	31.7*	29.7*
		IV. Heterog.-Heterosk.	36.4*	30.8*	24.5*	27.1*	26.9*	30.9*	29.7*	28.9*	28.6*	86.9*	34.2*	30.7*	31.7*	29.7*
		Average (V-SEB)/SEB	-0.162	-0.379	-0.547	-0.503	-0.510	-0.375	-0.409	-0.423	-0.439	-0.011	-0.322	-0.393	-0.380	-0.387
		F-stat (V = SEB)	627.6	1754.0	7311.1	4590.5	4981.7	1688.7	2165.0	2390.6	2688.4	244.4	1087.3	1904.1	1704.4	1877.6
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.4: Simulated Root Mean Squared Bias (x 1000) of the Estimated Treatment Effect on the Treated (TOT)

Trimming Results in the Normal-Normal Model (Sample size 100)

Trimming	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
Rule 1	I. Homog.-Homosk.	6.2*	34.3*	56.5*	46.5*	51.8*	49.5*	48.6*	11.0*	15.4*	8.9*	5.4*	4.2	2.2	4.1
	II. Heterog.-Homosk.	44.8*	69.2*	23.0*	24.6*	24.3*	53.5*	54.4*	39.7*	35.2*	47.8*	45.0*	46.9*	55.4*	48.3*
	III. Homog.-Heterosk.	5.4	30.7*	49.5*	37.9*	40.7*	17.9*	20.1*	10.9*	14.1*	11.7*	5.4	4.3	4.0	3.8
	IV. Heterog.-Heterosk.	43.5*	66.0*	19.1*	21.8*	19.9*	45.2*	46.6*	37.9*	34.6*	47.5*	44.1*	45.7*	53.3*	47.5*
	All	31.5*	53.1*	40.4*	34.2*	36.5*	43.8*	44.4*	28.5*	26.8*	34.5*	31.7*	32.9*	38.5*	34.0*
	F-stat (no bias)	1750.2	5867.8	4868.6	3309.1	3921.1	4082.7	4373.9	1891.3	1729.6	2662.4	2343.6	2461.9	3404.0	2156.7
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 2	I. Homog.-Homosk.	16.7*	44.6*	89.3*	70.2*	77.2*	53.5*	54.6*	25.7*	31.3*	13.0*	20.0*	23.2*	1.8	5.7
	II. Heterog.-Homosk.	22.2*	64.1*	96.0*	77.0*	83.9*	60.6*	62.2*	30.7*	36.2*	18.3*	26.3*	29.3*	15.0*	11.4*
	III. Homog.-Heterosk.	16.1*	41.8*	80.8*	59.1*	62.7*	23.4*	27.7*	25.9*	29.5*	14.2*	20.9*	23.6*	3.9	5.7
	IV. Heterog.-Heterosk.	21.9*	64.8*	87.3*	66.3*	70.3*	30.4*	33.8*	32.4*	36.2*	21.9*	28.7*	30.9*	15.4*	11.9*
	All	19.4*	54.9*	88.5*	68.4*	73.9*	44.7*	46.8*	28.8*	33.5*	17.2*	24.2*	27.0*	11.0*	9.2*
	F-stat (no bias)	569.1	5531.9	23283.6	12403.4	14937.3	3423.2	3971.0	1708.3	2405.7	344.7	1085.5	1481.9	230.6	114.0
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 3	I. Homog.-Homosk.	2.3	3.3	49.8*	41.6*	47.2*	42.3*	40.3*	5.3*	9.2*	139.4*	34.7*	12.7*	1.8	1.7
	II. Heterog.-Homosk.	50.3*	48.3*	22.3*	27.3*	26.2*	54.5*	54.4*	47.2*	42.6*	155.5*	83.9*	62.0*	35.7*	49.7*
	III. Homog.-Heterosk.	4.2	4.6	42.9*	33.0*	36.2*	14.6*	16.4*	7.1*	9.5*	137.2*	35.4*	14.6*	5.3	4.6
	IV. Heterog.-Heterosk.	48.9*	45.8*	20.5*	27.0*	24.0*	50.7*	51.3*	45.6*	42.4*	151.5*	82.2*	60.8*	34.6*	48.7*
	All	35.1*	33.4*	36.2*	32.8*	34.7*	43.4*	43.3*	33.1*	30.8*	146.1*	63.7*	44.5*	25.0*	34.9*
	F-stat (no bias)	2364.1	2587.8	3980.7	3168.6	3728.1	4519.7	4571.4	2742.5	2449.6	23016.7	7638.3	4646.7	1332.2	3071.5
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 4	I. Homog.-Homosk.	5.6	37.8*	59.3*	48.2*	53.8*	47.6*	46.8*	10.1*	14.9*	7.4*	2.4	3.5	2.3	2.9
	II. Heterog.-Homosk.	38.0*	70.8*	31.9*	27.2*	29.5*	51.8*	52.3*	32.8*	28.6*	41.4*	40.6*	39.4*	45.5*	40.7*
	III. Homog.-Heterosk.	5.4	35.6*	53.2*	40.1*	42.8*	17.6*	19.9*	11.4*	14.8*	5.7	4.2	4.8	3.6	3.1
	IV. Heterog.-Heterosk.	38.5*	69.4*	27.4*	22.5*	21.7*	40.8*	42.4*	32.9*	29.8*	42.5*	41.6*	40.1*	45.6*	42.0*
	All	27.3*	56.0*	45.1*	36.0*	39.0*	41.6*	42.2*	24.4*	23.2*	30.0*	29.2*	28.3*	32.3*	29.3*
	F-stat (no bias)	1345.5	6791.7	6219.9	3780.3	4623.4	3790.0	4049.8	1400.9	1309.3	1931.8	1973.8	1868.4	2387.4	1708.9
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.5: Simulated Average Variance (x 1000) of the Estimated Treatment Effect on the Treated (TOT)

Trimming Results in the Normal-Normal Model (Sample size 100)

Trimming	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
Rule 1	I. Homog.-Homosk.	125.7*	86.7*	61.1*	66.2*	63.5*	102.7*	97.6*	81.6*	77.4*	85.2*	79.0*	81.8*	79.5*	111.3*
	II. Heterog.-Homosk.	128.0*	99.6*	62.8*	67.7*	65.4*	105.2*	100.0*	83.9*	79.4*	87.3*	81.6*	84.2*	81.2*	114.0*
	III. Homog.-Heterosk.	141.9*	134.1*	127.2*	130.0*	130.1*	136.6*	136.1*	132.7*	132.3*	141.5*	134.9*	136.1*	143.1*	141.7*
	IV. Heterog.-Heterosk.	141.9*	144.5*	128.5*	130.9*	130.9*	137.0*	136.8*	133.6*	133.0*	142.9*	136.5*	137.2*	145.1*	141.6*
	Average (V-SEB)/SEB	-0.3	-0.4	-0.5	-0.5	-0.5	-0.4	-0.4	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.4
	F-stat (V = SEB)	9263.8	15514.0	40067.8	32825.1	35463.8	13319.6	14776.0	20859.7	23178.4	17567.5	21230.5	19582.3	20380.1	10186.5
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 2	I. Homog.-Homosk.	149.0*	111.0*	60.6*	70.3*	68.1*	124.0*	114.7*	94.1*	89.3*	170.7*	99.4*	91.0*	96.1*	176.1*
	II. Heterog.-Homosk.	150.5*	128.4*	61.6*	71.6*	69.5*	124.7*	116.7*	96.0*	91.4*	172.2*	102.5*	93.3*	99.8*	176.5*
	III. Homog.-Heterosk.	157.6*	147.0*	124.5*	132.1*	132.4*	145.7*	144.0*	138.1*	137.6*	214.2*	153.8*	143.8*	159.7*	167.3*
	IV. Heterog.-Heterosk.	157.0*	162.7*	124.7*	132.0*	132.1*	145.3*	144.5*	137.9*	137.6*	213.7*	152.9*	143.8*	159.8*	167.6*
	Average (V-SEB)/SEB	-0.2	-0.3	-0.6	-0.5	-0.5	-0.3	-0.4	-0.4	-0.5	-0.1	-0.4	-0.4	-0.4	-0.2
	F-stat (V = SEB)	5497.7	7484.9	42164.2	27865.0	29746.0	8474.7	9822.4	14735.8	16190.7	2447.2	11114.9	14281.3	11368.1	3468.4
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 3	I. Homog.-Homosk.	111.0*	79.3*	59.3*	62.1*	59.2*	88.6*	85.7*	72.8*	68.7*	202.0*	102.6*	77.9*	82.2*	71.9*
	II. Heterog.-Homosk.	112.8*	82.6*	60.9*	63.8*	61.2*	90.7*	88.1*	75.0*	70.9*	202.1*	105.2*	79.6*	84.7*	73.6*
	III. Homog.-Heterosk.	139.7*	135.6*	127.5*	130.0*	129.7*	135.6*	134.9*	133.0*	132.2*	226.8*	158.4*	140.1*	164.8*	132.4*
	IV. Heterog.-Heterosk.	139.3*	137.6*	128.3*	130.3*	130.1*	135.7*	135.3*	133.2*	132.6*	226.7*	157.7*	140.4*	165.0*	132.7*
	Average (V-SEB)/SEB	-0.374	-0.476	-0.550	-0.538	-0.546	-0.452	-0.461	-0.501	-0.514	-0.021	-0.377	-0.475	-0.415	-0.506
	F-stat (V = SEB)	11858.0	20730.0	42775.8	37569.1	41420.8	18083.0	19281.0	26259.2	29626.0	1414.1	10112.0	21478.8	16769.2	26786.4
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 4	I. Homog.-Homosk.	121.3*	84.2*	59.2*	64.0*	61.4*	100.0*	95.1*	79.5*	75.1*	86.9*	78.8*	78.0*	78.4*	103.0*
	II. Heterog.-Homosk.	122.5*	96.6*	59.9*	64.9*	62.5*	101.8*	97.0*	81.1*	76.7*	88.4*	80.6*	79.8*	80.4*	105.3*
	III. Homog.-Heterosk.	140.4*	131.1*	125.5*	128.4*	128.3*	135.4*	134.7*	131.5*	130.9*	141.8*	135.1*	133.8*	142.7*	139.0*
	IV. Heterog.-Heterosk.	139.6*	142.1*	125.7*	128.3*	128.2*	135.4*	134.9*	131.4*	130.9*	142.1*	135.3*	134.1*	143.4*	138.8*
	Average (V-SEB)/SEB	-0.347	-0.454	-0.557	-0.538	-0.545	-0.420	-0.435	-0.486	-0.500	-0.447	-0.481	-0.486	-0.466	-0.398
	F-stat (V = SEB)	10121.3	16728.4	43560.5	35892.8	38878.5	14475.8	15970.7	22559.7	25166.4	17014.9	21910.2	22396.7	21177.8	13108.1
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. “All” is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.A1: Bias and Variance of the Estimated Treatment Effect on the Treated (ATE)

Normal-Cauchy Model

	Sample Size	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
A. Simulated Root	100	I. Homog.-Homosk.	3.4	19.2*	44.8*	37.4*	43.5*	63.0*	4.3*	7.5*	12.6*	2.8	1.9	2.5	2.0	2.0
Mean Squared Bias		II. Heterog.-Homosk.	2.7	30.5*	60.1*	47.1*	54.9*	22.9*	2.9	10.7*	17.3*	1.9	2.7	3.9	2.7	16.5*
(x 1000)		III. Homog.-Heterosk.	4.4	19.1*	43.5*	35.9*	46.6*	366.9*	20.7*	9.6*	11.9*	3.8	3.7	4.0	4.5	4.3
		IV. Heterog.-Heterosk.	4.8	30.9*	59.2*	46.8*	61.5*	330.8*	17.1*	13.6*	17.2*	2.3	3.2	4.5	4.2	13.0*
		All	3.9	25.6*	52.5*	42.1*	52.1*	249.3*	13.7*	10.6*	15.0*	2.8	3.0	3.8*	3.5	10.8*
		F-stat (no bias)	34.8	2126.2	9872.6	6125.0	9339.9	69084.1	421.8	337.2	760.9	20.5	26.2	44.9	34.6	405.2
		[p-value]	[0.071]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.667]	[0.343]	[0.006]	[0.075]	[0.000]
	500	I. Homog.-Homosk.	4.0	5.4*	34.7*	26.7*	29.3*	3.5	3.4	3.1	3.0	3.0	3.4	3.3	3.5	3.4
		II. Heterog.-Homosk.	2.8	8.5*	38.6*	27.7*	30.2*	2.1	1.8	2.2	3.3	2.6	1.9	1.9	2.0	13.6*
		III. Homog.-Heterosk.	4.0	6.2*	42.5*	32.5*	37.2*	6.2*	4.7	4.0	4.1	3.4	3.3	3.2	3.4	3.1
		IV. Heterog.-Heterosk.	4.2	9.5*	51.6*	37.8*	43.0*	5.5*	4.4	4.2	5.4*	3.1	3.0	3.0	3.1	13.2*
		All	3.8	7.6*	42.3*	31.5*	35.4*	4.6*	3.8*	3.5	4.0*	3.1	3.0	2.9	3.1	9.7*
		F-stat (no bias)	33.6	191.7	5766.9	3301.7	4140.2	61.0	43.4	37.7	52.0	26.0	29.6	27.9	31.2	338.2
		[p-value]	[0.092]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.009]	[0.037]	[0.001]	[0.354]	[0.199]	[0.263]	[0.148]	[0.000]
B. Simulated	100	I. Homog.-Homosk.	76.8*	57.9*	48.6*	50.9*	48.8*	75.8*	56.7*	56.5*	53.6	69.3*	56.4*	55.9*	57.4*	53.6*
Average Variance		II. Heterog.-Homosk.	77.6*	60.8*	50.9*	53.2*	51.1*	69.3*	57.4*	57.5*	54.7	77.2*	57.0*	56.5*	58.1*	54.2
(x 1000)		III. Homog.-Heterosk.	130.2*	101.4*	98.9*	105.3*	99.0*	157.0*	101.9*	106.0*	98.2*	110.1*	97.6*	97.3*	100.2*	98.9*
		IV. Heterog.-Heterosk.	130.0*	103.1*	99.8*	106.0*	99.7*	251.4*	101.4*	106.3*	98.7*	118.5*	97.9*	97.5*	100.5*	98.6*
		Average (V-SEB)/SEB	0.406	0.093	-0.011	0.044	-0.009	0.751	0.068	0.092	0.026	0.284	0.044	0.038	0.068	0.025
		F-stat (V = SEB)	418.1	40.1	36.1	46.2	33.4	746.9	28.5	46.2	8.2	256.9	10.1	7.8	21.3	7.3
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	500	I. Homog.-Homosk.	15.3*	11.0	10.2*	10.0*	10.0*	10.9	10.8	10.8	10.7	13.0*	10.9	10.9	10.9	10.2*
		II. Heterog.-Homosk.	15.3*	10.9	10.4*	10.2*	10.2*	10.8	10.7	10.8	10.6	13.9*	10.8	10.8	10.8	10.2*
		III. Homog.-Heterosk.	25.7*	19.1	19.6*	19.0	19.0	19.2*	18.7	18.8	18.6	20.5*	18.6	18.6	18.7	18.6
		IV. Heterog.-Heterosk.	25.0*	18.6	19.4*	18.5	18.5	18.6	18.2	18.2	18.1*	21.2*	18.2	18.2	18.2	18.1
		Average (V-SEB)/SEB	0.383	0.009	-0.004	-0.031	-0.034	0.007	-0.011	-0.008	-0.020	0.174	-0.008	-0.010	-0.006	-0.039
		F-stat (V = SEB)	77.3	1.6	5.5	5.4	6.0	1.7	1.6	1.4	1.8	25.7	1.4	1.4	1.4	3.3
		[p-value]	[0.000]	[0.028]	[0.000]	[0.000]	[0.000]	[0.017]	[0.040]	[0.073]	[0.007]	[0.000]	[0.085]	[0.072]	[0.078]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.A2: Bias and Variance of the Estimated Treatment Effect on the Treated (ATE) under Misspecification

Misspecification of the Propensity Score in the Normal-Cauchy Model (sample size 100)

	Misspec. Type	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
A. Simulated Root	Xs	I. Homog.-Homosk.	122.9*	120.0*	143.6*	141.8*	144.1*	111.5*	120.8*	126.0*	128.9*	121.6*	122.8*	123.2*	122.4*	124.5*
Mean Squared Bias		II. Heterog.-Homosk.	185.1*	182.1*	215.1*	211.3*	215.2*	180.0*	183.1*	189.6*	193.5*	183.0*	185.0*	185.5*	183.6*	186.6*
(x 1000)		III. Homog.-Heterosk.	123.6*	121.5*	142.9*	140.7*	145.1*	265.7*	117.2*	127.6*	129.4*	121.8*	123.5*	123.8*	123.2*	125.1*
		IV. Heterog.-Heterosk.	184.2*	181.5*	211.8*	208.6*	214.5*	220.5*	177.7*	190.1*	192.1*	181.9*	184.0*	184.6*	183.3*	186.3*
		All	157.0*	154.3*	181.8*	178.9*	183.1*	202.5*	152.8*	161.4*	164.1*	155.1*	156.8*	157.3*	156.1*	158.6*
		F-stat (no bias)	61774.7	78783.1	117000	112000	120000	78669.4	78860.3	85738.5	92317.6	76674.7	84121.1	84759.8	81267.2	85421.5
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	Dist of u.	I. Homog.-Homosk.	3.6	19.3*	45.3*	38.2*	44.3*	67.9*	5.3*	7.3*	12.6*	36.1*	2.5	2.3	2.1	1.9
		II. Heterog.-Homosk.	3.0	31.2*	61.2*	48.6*	56.2*	27.5*	4.5*	11.6*	17.3*	46.0*	4.4	4.6*	2.1	16.9*
		III. Homog.-Heterosk.	4.5	19.3*	43.7*	38.4*	48.1*	355.2*	25.7*	11.3*	12.3*	34.9*	6.1*	5.2	4.3	4.2
		IV. Heterog.-Heterosk.	5.1	31.6*	60.1*	50.5*	63.7*	321.2*	21.9*	16.7*	17.9*	47.0*	6.1*	6.8*	5.5	13.4*
		All	4.1	26.1*	53.2*	44.3*	53.6*	242.2*	17.3*	12.2*	15.2*	41.4*	5.0*	5.0*	3.8	11.0*
		F-stat (no bias)	38.5	2187.6	10191.9	6682.6	9863.5	67751.5	653.5	426.4	777.2	2475.9	63.1	68.4	36.5	422.9
		[p-value]	[0.031]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.049]	[0.000]
B. Simulated	Xs	I. Homog.-Homosk.	66.5	49.6*	45.5*	45.7*	44.8*	57.4*	49.4*	49.2*	47.8*	52.9*	48.5*	48.4*	49.6*	48.1*
Average Variance		II. Heterog.-Homosk.	67.8	51.9*	47.0*	47.3*	46.3*	54.0*	50.8*	50.4*	48.9*	56.5*	49.4*	49.3*	50.6*	49.2*
(x 1000)		III. Homog.-Heterosk.	163.7*	123.9*	123.5*	128.8*	122.0*	184.3*	127.5*	130.8*	123.5*	125.8*	121.6*	121.5*	125.0*	125.5*
		IV. Heterog.-Heterosk.	165.7*	126.5*	124.0*	129.5*	122.2*	269.5*	127.5*	131.6*	124.1*	129.4*	122.2*	122.1*	125.6*	125.8*
		Average (V-SEB)/SEB	0.258	-0.046	-0.085	-0.059	-0.098	0.461	-0.040	-0.025	-0.070	-0.006	-0.074	-0.076	-0.050	-0.059
		F-stat (V = SEB)	299.5	352.3	602.7	614.4	647.6	804.9	391.2	422.2	462.3	228.9	418.0	423.2	372.3	445.4
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	Dist of u.	I. Homog.-Homosk.	76.8*	58.4*	48.4*	50.5*	48.6*	75.4*	57.1*	56.4*	53.8	93.0*	63.5*	58.7*	60.4*	53.4*
		II. Heterog.-Homosk.	77.4*	61.4*	50.5*	52.8*	50.9*	69.7*	57.9*	57.4*	54.8	111.2*	64.5*	59.3*	61.3*	54.1
		III. Homog.-Heterosk.	130.2*	102.0*	99.0*	105.2*	98.9*	153.7*	103.7*	106.1*	98.5*	153.4*	107.2*	101.0*	104.3*	98.8*
		IV. Heterog.-Heterosk.	130.0*	103.8*	99.7*	106.3*	99.8*	243.8*	102.7*	106.7*	98.9*	164.0*	107.7*	101.4*	104.8*	98.7*
		Average (V-SEB)/SEB	0.405	0.102	-0.013	0.041	-0.011	0.722	0.081	0.093	0.028	0.767	0.161	0.083	0.118	0.024
		F-stat (V = SEB)	416.2	47.1	38.6	49.6	35.7	732.2	38.5	47.3	8.8	858.5	104.4	32.6	59.0	7.4
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.A3: Bias and Variance of the Estimated Treatment Effect on the Treated (ATE)

Normal-Normal Model

	N	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
A. Simulated Root	100	I. Homog.-Homosk.	9.1*	35.0*	87.1*	67.3*	77.1*	62.3*	10.7*	16.0*	21.9*	15.3*	8.9*	9.7*	3.3	2.5
Mean Squared Bias		II. Heterog.-Homosk.	11.3*	49.3*	114.8*	80.5*	90.8*	22.6*	6.4*	22.6*	28.6*	18.8*	14.2*	15.4*	8.8*	28.6*
(x 1000)		III. Homog.-Heterosk.	8.6*	33.9*	87.9*	72.0*	86.3*	404.1*	31.6*	26.3*	23.7*	14.6*	12.0*	12.1*	4.6	4.6
		IV. Heterog.-Heterosk.	12.3*	50.4*	119.6*	94.3*	112.0*	424.5*	26.5*	34.9*	29.7*	22.5*	14.3*	14.3*	8.1*	26.1*
		All	10.4*	42.9*	103.4*	79.2*	92.4*	294.9*	21.5*	25.9*	26.2*	18.1*	12.5*	13.0*	6.6*	19.6*
		F-stat (no bias)	221.3	4510.7	35319.0	18137.4	25567.4	133000	969.4	1625.8	2008.5	446.9	383.6	465.9	116.7	1220.5
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	500	I. Homog.-Homosk.	3.4	10.9*	65.0*	43.2*	48.1*	4.3	4.1	4.3	5.5*	4.4	3.2	3.8	3.1	2.6
		II. Heterog.-Homosk.	3.3	12.4*	73.0*	41.0*	45.4*	4.2	3.7	4.5	5.4*	6.4*	3.2	4.3	3.0	26.0*
		III. Homog.-Heterosk.	2.8	11.4*	82.1*	58.5*	66.7*	10.6*	9.7*	8.8*	9.8*	4.3	3.1	4.1	2.7	2.6
		IV. Heterog.-Heterosk.	5.2	15.3*	93.4*	60.5*	69.1*	7.9*	7.3*	6.9*	7.6*	8.2	4.7	6.3*	2.0	24.9*
		All	3.8	12.6*	79.1*	51.5*	58.3*	7.3*	6.6*	6.4*	7.3*	6.0*	3.6	4.7*	2.7	18.1*
		F-stat (no bias)	26.4	395.1	17972.6	7399.3	9487.5	123.8	110.1	104.3	142.0	49.4	28.9	56.1	20.1	1091.7
		[p-value]	[0.332]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.002]	[0.222]	[0.000]	[0.690]	[0.000]
B. Simulated	100	I. Homog.-Homosk.	95.9*	76.2*	49.3*	55.7*	52.9*	83.6*	68.2*	66.1*	62.2*	120.7*	76.5*	69.6*	73.6*	59.5*
Average Variance		II. Heterog.-Homosk.	96.3*	83.2*	53.2*	60.3*	58.1*	81.7*	70.8*	67.9*	64.5*	151.7*	77.5*	70.0*	74.6*	60.3*
(x 1000)		III. Homog.-Heterosk.	148.3*	119.3*	107.5*	118.3*	109.2*	117.6*	114.2*	119.8*	108.4*	172.3*	118.6*	111.5*	118.2*	109.1*
		IV. Heterog.-Heterosk.	150.0*	127.2*	109.6*	119.4*	110.2*	271.8*	116.3*	120.5*	109.2*	187.2*	118.8*	112.1*	119.3*	109.1*
		Average (V-SEB)/SEB	-0.169	-0.317	-0.468	-0.410	-0.449	-0.085	-0.377	-0.370	-0.420	0.053	-0.340	-0.386	-0.350	-0.434
		F-stat (V = SEB)	2203.2	4026.1	14828.2	10162.5	11744.5	4744.6	6446.5	7039.7	8573.4	639.4	4620.2	6593.0	5160.8	9592.9
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	500	I. Homog.-Homosk.	22.3*	15.3*	11.1*	11.8*	11.5*	14.2*	13.5*	13.5*	13.1*	29.2*	18.3*	15.2*	15.8*	11.5*
		II. Heterog.-Homosk.	21.6*	16.0*	11.4*	12.3*	12.0*	14.9*	14.0*	13.9*	13.4*	38.9*	17.8*	15.0*	15.7*	11.6*
		III. Homog.-Heterosk.	32.4*	23.7*	22.4*	21.3*	21.4*	23.0*	21.7*	21.7*	21.3*	40.0*	25.8*	23.1*	23.8*	20.8*
		IV. Heterog.-Heterosk.	31.7*	23.9*	22.0*	20.9*	21.0*	22.3*	21.2*	21.1*	20.6*	41.8*	25.6*	22.4*	23.3*	20.1*
		Average (V-SEB)/SEB	-0.090	-0.335	-0.443	-0.444	-0.449	-0.373	-0.407	-0.407	-0.424	0.193	-0.271	-0.363	-0.339	-0.464
		F-stat (V = SEB)	253.2	874.9	2323.6	1896.6	2033.9	1118.3	1356.5	1373.1	1549.7	86.1	457.4	1003.7	832.0	2160.5
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.A4: Simulated Root Mean Squared Bias (x 1000) of the Estimated Treatment Effect on the Treated (ATE)

Trimming Results in the Normal-Normal Model (Sample size 100)

Trimming	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
Rule 1	I. Homog.-Homosk.	5.7*	28.1*	51.7*	43.5*	49.7*	75.3*	6.1*	12.7*	16.5*	71.2*	4.4	3.8	3.3	2.9
	II. Heterog.-Homosk.	23.5*	30.6*	64.1*	49.4*	56.1*	37.0*	19.8*	24.2*	25.2*	111.1*	24.5*	22.2*	27.2*	35.5*
	III. Homog.-Heterosk.	3.8	29.3*	47.9*	39.0*	49.2*	374.2*	22.6*	13.7*	13.8*	69.8*	8.5*	5.7	5.0	4.8
	IV. Heterog.-Heterosk.	18.4*	32.6*	62.4*	49.0*	61.8*	379.8*	13.4*	22.1*	20.1*	109.4*	19.4*	16.6*	21.9*	30.2*
	All	15.3*	30.2*	57.0*	45.4*	54.5*	269.9*	16.7*	18.9*	19.4*	92.6*	16.3*	14.3*	17.7*	23.5*
	F-stat (no bias)	508.0	2230.7	9461.0	5589.7	8332.2	97336	686.3	922.0	1100.6	12451.3	711.3	575.3	863.1	1594.7
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 2	I. Homog.-Homosk.	18.9*	112.5*	97.6*	78.2*	86.6*	61.0*	15.7*	23.9*	29.3*	85.8*	55.9*	41.8*	2.9	2.6
	II. Heterog.-Homosk.	21.6*	173.4*	126.3*	92.5*	101.2*	26.1*	12.4*	29.2*	35.2*	134.7*	78.8*	57.2*	20.9*	31.4*
	III. Homog.-Heterosk.	15.2*	113.7*	95.8*	80.7*	93.9*	399.0*	37.4*	32.4*	28.5*	85.5*	52.9*	38.8*	4.3	4.3
	IV. Heterog.-Heterosk.	24.6*	175.8*	132.3*	108.4*	124.7*	416.0*	35.1*	45.0*	39.2*	136.5*	80.2*	58.6*	17.9*	26.3*
	All	20.4*	147.1*	114.2*	90.8*	102.6*	290.1*	27.5*	33.5*	33.3*	113.4*	68.1*	49.9*	14.0*	20.6*
	F-stat (no bias)	689.5	47809.4	41096.3	22584.5	29204.9	124000	1469.6	2550.6	2945.1	32887.1	14341.6	7340.5	517.5	1253.8
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 3	I. Homog.-Homosk.	5.2*	23.4*	47.1*	39.9*	45.5*	85.6*	6.7*	11.8*	15.4*	17.8*	5.9*	2.7	3.7	3.2
	II. Heterog.-Homosk.	24.5*	23.0*	59.4*	46.9*	53.3*	46.4*	20.4*	26.1*	26.8*	46.1*	29.2*	26.0*	25.1*	34.0*
	III. Homog.-Heterosk.	4.9	25.4*	44.1*	36.0*	46.0*	382.4*	25.5*	12.6*	12.8*	16.3*	11.1*	7.1*	6.2*	5.5
	IV. Heterog.-Heterosk.	18.0*	25.6*	57.2*	44.9*	56.9*	397.4*	16.1*	21.6*	20.0*	40.5*	23.6*	19.9*	18.6*	27.5*
	All	15.6*	24.4*	52.3*	42.1*	50.7*	280.0*	18.5*	19.0*	19.5*	33.0*	19.8*	16.8*	16.0*	22.1*
	F-stat (no bias)	514.0	1395.2	7450.2	4479.7	6758.0	105000	765.5	896.2	1061.9	1810.1	995.5	755.6	680.0	1341.8
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 4	I. Homog.-Homosk.	5.1*	36.3*	57.3*	47.7*	54.4*	62.2*	6.3*	12.8*	16.4*	4.0	2.5	3.8	2.6	2.6
	II. Heterog.-Homosk.	25.2*	43.3*	72.4*	56.4*	64.1*	24.4*	21.7*	27.1*	28.6*	29.6*	25.4*	24.9*	26.7*	34.8*
	III. Homog.-Heterosk.	3.5	37.7*	53.8*	43.7*	55.3*	381.7*	22.7*	15.3*	13.7*	6.0	4.7	4.0	5.1	4.8
	IV. Heterog.-Heterosk.	20.0*	46.0*	71.8*	55.7*	71.3*	388.9*	15.6*	26.4*	23.4*	24.6*	20.1*	19.5*	21.5*	29.7*
	All	16.4*	41.0*	64.4*	51.2*	61.7*	274.5*	17.8*	21.4*	21.4*	19.6*	16.4*	16.1*	17.4*	23.0*
	F-stat (no bias)	613.6	4394.9	12593.3	7379.7	11116.8	115000	835.3	1207.1	1394.1	891.8	802.5	781.6	877.8	1599.7
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.A5: Simulated Average Variance (x 1000) of the Estimated Treatment Effect on the Treated (ATE)

Trimming Results in the Normal-Normal Model (Sample size 100)

Trimming	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
Rule 1	I. Homog.-Homosk.	92.3*	73.1*	58.8*	62.8*	59.5*	89.0*	69.3*	69.2*	65.3*	108.8*	73.9*	70.5*	71.0*	65.8*
	II. Heterog.-Homosk.	93.2*	78.2*	61.5*	65.8*	62.4*	89.0*	71.2*	71.3*	67.0*	131.5*	76.2*	71.9*	72.1*	66.9*
	III. Homog.-Heterosk.	159.2*	127.2*	125.2*	138.6*	124.6*	144.3*	126.5*	138.1*	122.4*	173.0*	124.7*	121.7*	126.3*	125.0*
	IV. Heterog.-Heterosk.	160.6*	132.5*	126.3*	139.7*	125.3*	296.9*	126.1*	138.5*	122.7*	199.1*	125.7*	122.2*	126.5*	124.7*
	Average (V-SEB)/SEB	-0.2	-0.3	-0.4	-0.3	-0.4	0.0	-0.3	-0.3	-0.4	0.0	-0.3	-0.4	-0.3	-0.4
	F-stat (V = SEB)	2314.2	4645.6	8800.8	7041.2	8420.6	3544.6	5783.1	5571.9	6931.8	919.4	4739.7	5695.0	5453.2	6705.2
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 2	I. Homog.-Homosk.	116.0*	78.2*	51.0*	57.9*	56.1*	95.8*	76.8*	71.3*	67.7*	67.0*	58.6*	62.4*	72.9*	64.7*
	II. Heterog.-Homosk.	117.2*	94.9*	55.4*	63.6*	62.7*	96.2*	82.3*	74.4*	71.2*	79.4*	60.2*	63.7*	74.7*	66.2*
	III. Homog.-Heterosk.	179.8*	122.2*	114.1*	125.8*	117.8*	124.3*	128.6*	131.4*	120.7*	109.8*	110.4*	112.6*	126.8*	122.2*
	IV. Heterog.-Heterosk.	180.2*	137.3*	114.7*	126.1*	118.4*	283.0*	131.8*	131.0*	120.2*	118.7*	110.2*	111.9*	126.5*	121.5*
	Average (V-SEB)/SEB	0.0	-0.3	-0.4	-0.4	-0.4	0.0	-0.3	-0.3	-0.4	-0.4	-0.4	-0.4	-0.3	-0.4
	F-stat (V = SEB)	879.0	3293.2	13187.5	8771.6	9335.4	2797.2	3800.4	5115.2	6013.1	7814.3	11278.8	8934.6	5173.4	6987.5
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 3	I. Homog.-Homosk.	94.4*	74.8*	62.8*	67.6*	63.3*	93.7*	72.6*	73.4*	68.6*	101.0*	75.9*	73.2*	74.2*	70.0*
	II. Heterog.-Homosk.	96.2*	79.9*	66.3*	71.2*	66.6*	100.2*	74.6*	75.9*	70.7*	115.1*	78.0*	75.0*	75.8*	71.5*
	III. Homog.-Heterosk.	170.6*	135.4*	134.4*	150.4*	134.1*	143.2*	137.0*	150.6*	132.6*	157.2*	132.9*	131.3*	136.1*	134.2*
	IV. Heterog.-Heterosk.	172.8*	140.8*	136.3*	152.2*	135.7*	300.6*	137.6*	151.6*	133.7*	173.1*	134.0*	132.1*	137.3*	134.5*
	Average (V-SEB)/SEB	-0.105	-0.278	-0.337	-0.270	-0.337	0.055	-0.295	-0.249	-0.323	-0.083	-0.296	-0.310	-0.292	-0.316
	F-stat (V = SEB)	2043.0	4214.1	7047.7	5454.9	6869.8	2835.0	4811.6	4449.4	5771.7	1315.1	4129.2	4817.5	4555.0	5394.8
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 4	I. Homog.-Homosk.	87.3*	68.2*	55.7*	59.4*	56.1*	80.5*	65.8*	66.1*	62.2*	77.8*	66.4*	65.4*	67.4*	63.3*
	II. Heterog.-Homosk.	88.2*	73.9*	58.3*	62.3*	59.1*	78.1*	67.5*	67.8*	63.8*	87.1*	68.0*	66.6*	68.4*	64.5*
	III. Homog.-Heterosk.	154.1*	120.9*	121.3*	134.4*	120.7*	125.2*	121.0*	133.9*	118.5*	129.2*	117.4*	116.8*	121.6*	120.2*
	IV. Heterog.-Heterosk.	154.4*	125.8*	121.1*	133.7*	120.2*	277.1*	120.8*	133.1*	117.8*	136.3*	117.3*	116.4*	121.4*	119.9*
	Average (V-SEB)/SEB	-0.188	-0.348	-0.411	-0.355	-0.410	-0.079	-0.372	-0.332	-0.395	-0.277	-0.381	-0.388	-0.366	-0.386
	F-stat (V = SEB)	2745.3	5673.3	10211.1	8290.0	9902.3	4888.3	6796.4	6447.1	8003.3	3594.1	6781.5	7165.9	6439.9	7546.3
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). Models: Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, RMSB=sqrt{(1/6)(b1+...+b6)} and the AV=(1/6)(v1+…+v6) where bi (i=1,…,6) is the square of the bias and vi (i=1,…,6) is the variance of one of the 6 combinations of the 2 curves and the 3 designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present 2 F-tests and their p-values: (i) H0:bias=0, (ii) H0: V=SEB (see section IV for details). One star means that we reject the null at the 1%.

Table 3.A6: Cramer-Rao and Semiparametric Efficiency Bounds

			Cauchy				Cauchy Misspecified				Normal			
			ATE		TOT		ATE		TOT		ATE		TOT	
Setting	Design	Curve	CRLB	SEB	CRLB	SEB	CRLB	SEB	CRLB	SEB	CRLB	SEB	CRLB	SEB
I	A	1	4.70	4.79	5.32	5.58	4.49	6.37	4.65	8.73	5.81	24.07	7.56	44.14
	A	2	4.51	4.79	5.39	5.58	4.28	6.37	4.51	8.73	5.61	24.07	9.60	44.14
	B	1	5.47	5.70	6.29	6.80	4.54	6.98	4.68	9.45	5.71	9.85	7.44	17.63
	B	2	5.56	5.70	6.75	6.80	4.27	6.98	4.43	9.45	6.34	9.85	9.82	17.63
	C	1	5.37	5.70	6.05	6.39	4.41	6.98	4.57	9.34	5.60	9.85	6.78	10.02
	C	2	4.80	5.70	5.24	6.39	4.19	6.98	4.37	9.34	4.79	9.85	6.21	10.02
II	A	1	4.66	4.81	5.39	5.61	4.58	6.40	4.77	8.79	6.00	24.11	7.93	44.20
	A	2	4.55	4.80	5.44	5.59	4.24	6.38	4.50	8.76	5.68	24.09	9.65	44.19
	B	1	5.50	5.71	6.34	6.82	4.47	7.01	4.71	9.48	5.69	9.88	7.24	17.66
	B	2	5.55	5.70	6.74	6.81	4.44	6.99	4.68	9.47	6.37	9.87	9.82	17.66
	C	1	5.46	5.71	6.14	6.45	4.39	7.01	4.64	9.44	5.77	9.88	6.88	10.08
	C	2	4.64	5.70	4.99	6.40	4.35	6.99	4.60	9.38	4.57	9.86	5.87	10.05
III	A	1	8.48	8.79	9.49	10.14	8.85	10.37	11.81	13.91	9.42	28.07	11.08	49.40
	A	2	8.70	8.79	9.75	10.14	8.30	10.37	10.41	13.91	10.26	28.07	13.30	49.40
	B	1	8.79	9.70	9.75	11.13	8.75	10.98	11.67	14.17	8.72	13.85	10.19	22.32
	B	2	9.22	9.70	10.30	11.13	8.31	10.98	10.37	14.17	9.87	13.85	12.12	22.32
	C	1	9.50	9.70	11.02	11.62	8.57	10.98	11.48	15.27	9.91	13.85	11.60	15.49
	C	2	9.17	9.70	10.02	11.62	8.23	10.98	10.33	15.27	9.62	13.85	11.34	15.49
IV	A	1	8.42	8.81	9.44	10.17	8.90	10.40	12.01	13.96	9.46	28.11	11.37	49.46
	A	2	8.51	8.80	9.43	10.15	8.42	10.38	10.67	13.94	10.07	28.09	12.93	49.45
	B	1	8.77	9.71	9.72	11.14	8.90	11.01	12.07	14.20	8.90	13.88	10.39	22.36
	B	2	9.50	9.70	10.56	11.14	8.49	10.99	10.77	14.19	10.01	13.87	12.33	22.36
	C	1	9.51	9.71	11.22	11.67	8.82	11.01	11.88	15.37	9.88	13.88	11.72	15.55
	C	2	9.32	9.70	10.26	11.62	8.25	10.99	10.43	15.31	9.87	13.86	11.68	15.52

NOTES: See section II.E and Appendix II for details on the computation of the SEB. The CRLB was computed assuming full knowledge of the parametric model.

Figure 3.1A: Overlap Plots, by design (Normal-Cauchy model)

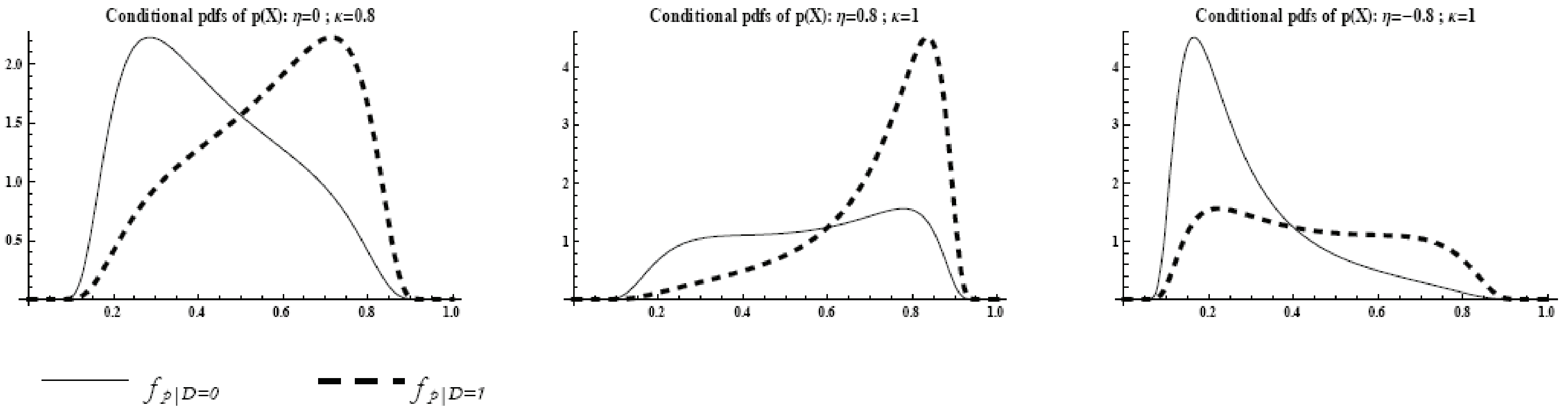


Figure 3.1B: Overlap Plots, by design (Normal-Normal model)

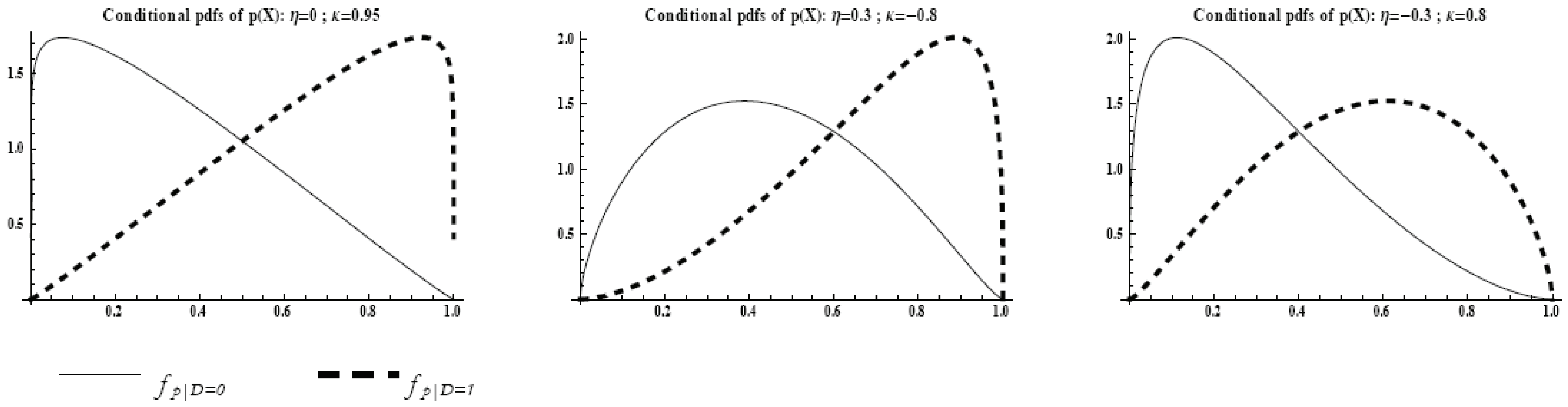
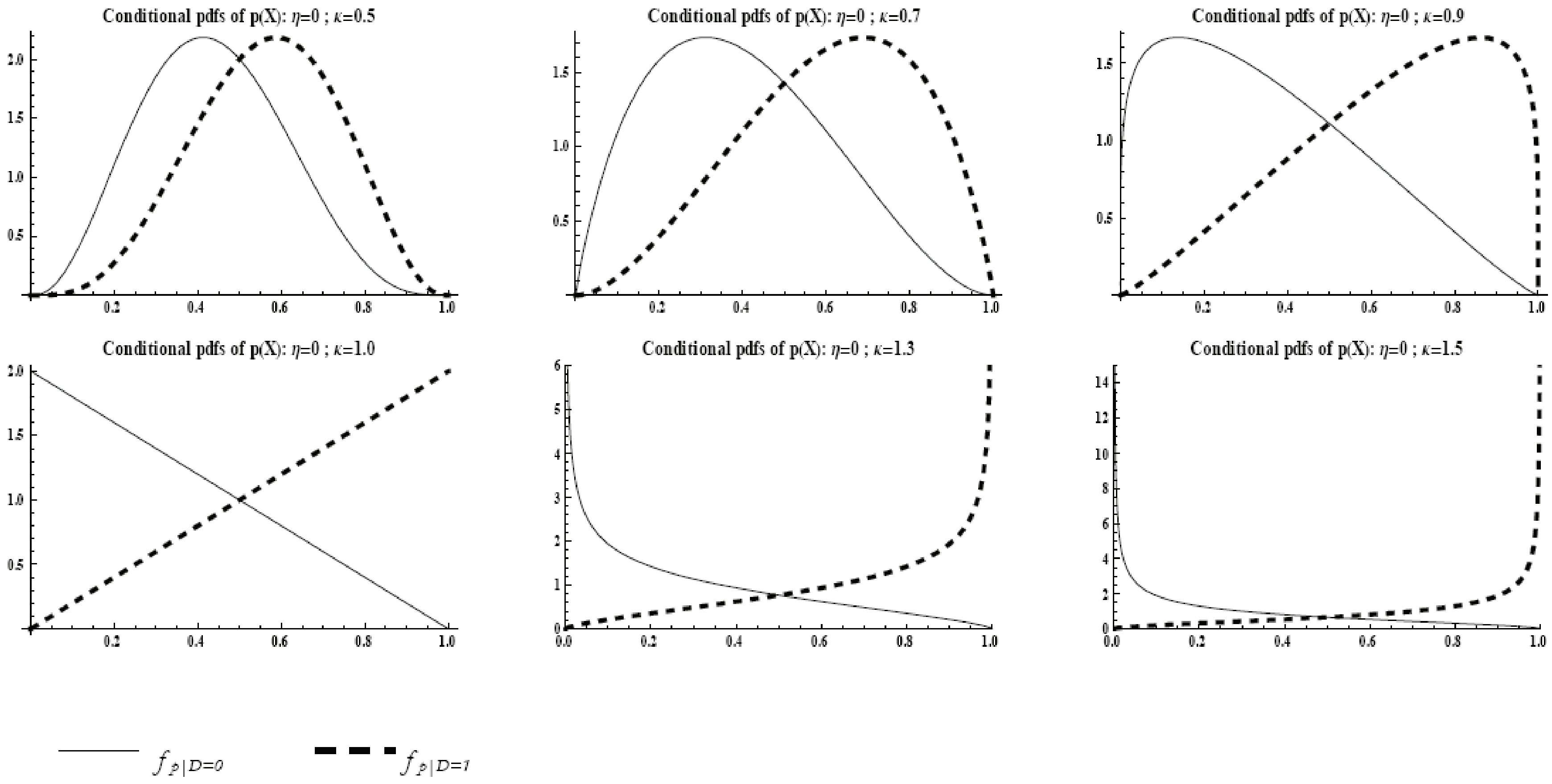


Figure 3.2: Overlap Plots, by design (Normal-Normal model)



Conditional pdfs of p(X): η=0 ; κ=0.5

Conditional pdfs of p(X): η=0 ; κ=0.7

Conditional pdfs of p(X): η=0 ; κ=0.9

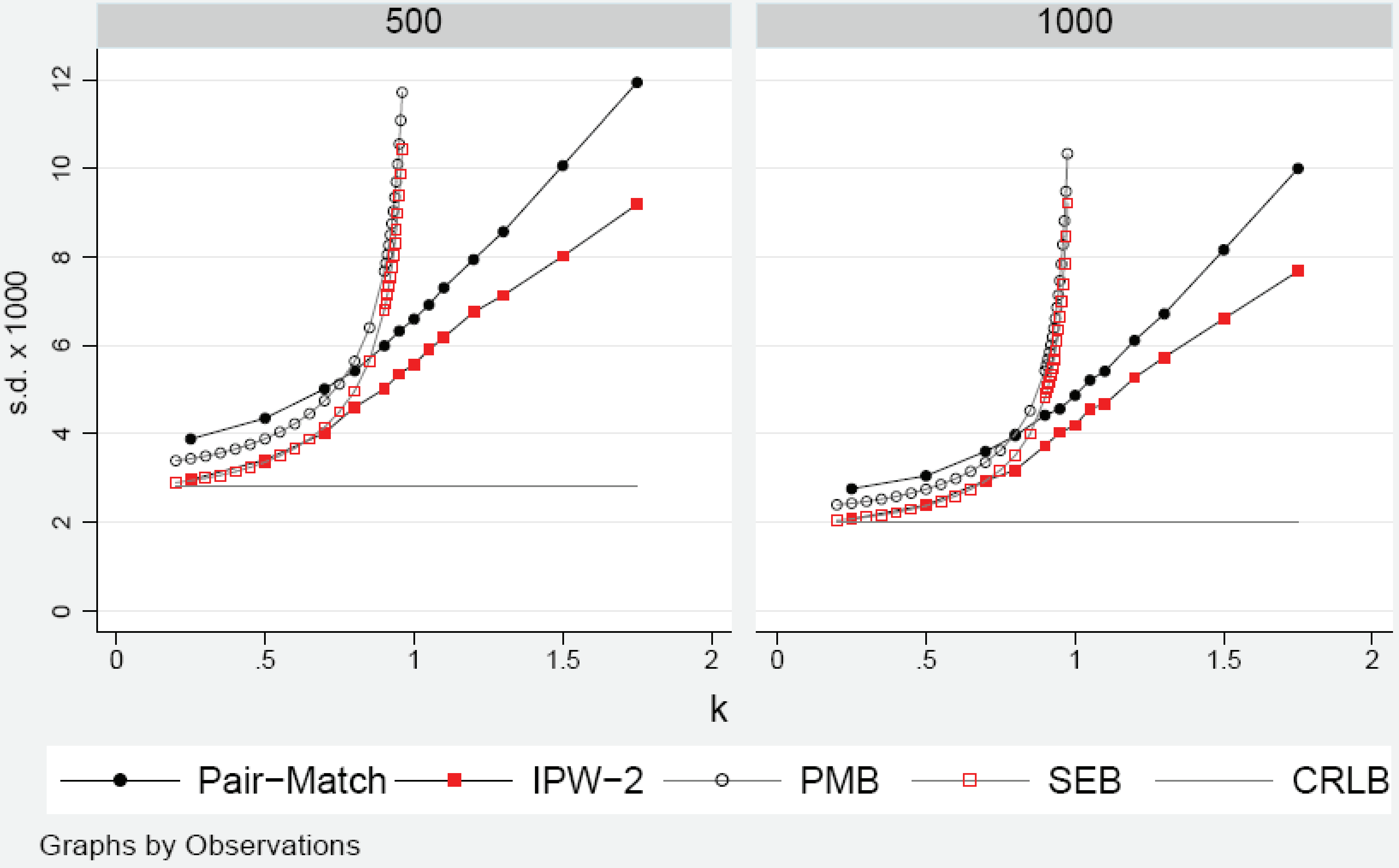
Conditional pdfs of p(X): η=0 ; κ=1.0

Conditional pdfs of p(X): η=0 ; κ=1.3

Conditional pdfs of p(X): η=0 ; κ=1.5

Figure 3.3: Breakdown of Standard Asymptotics as k grows

Actual and Expected s.d. of IPW2 & PM in Normal−Normal model



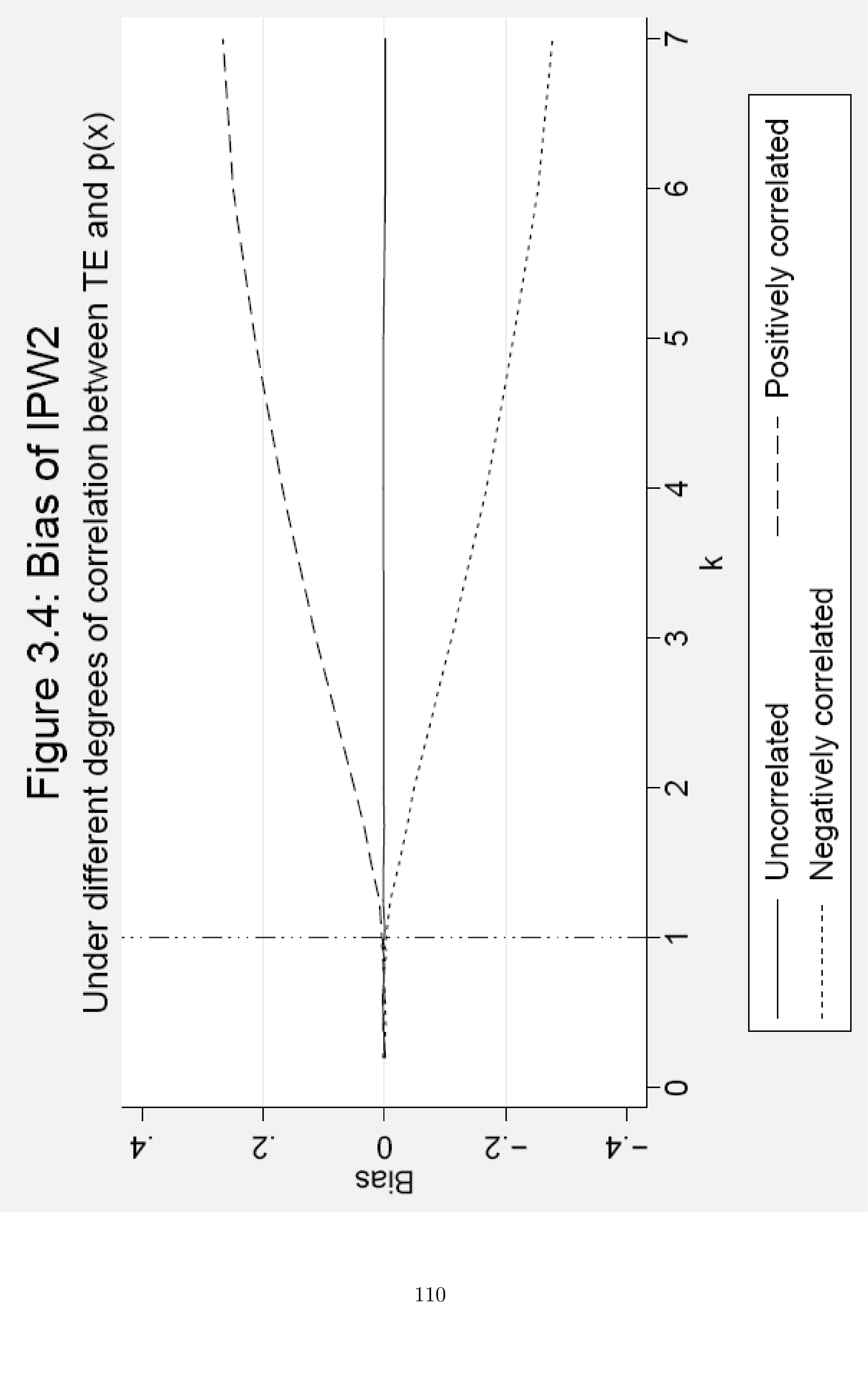
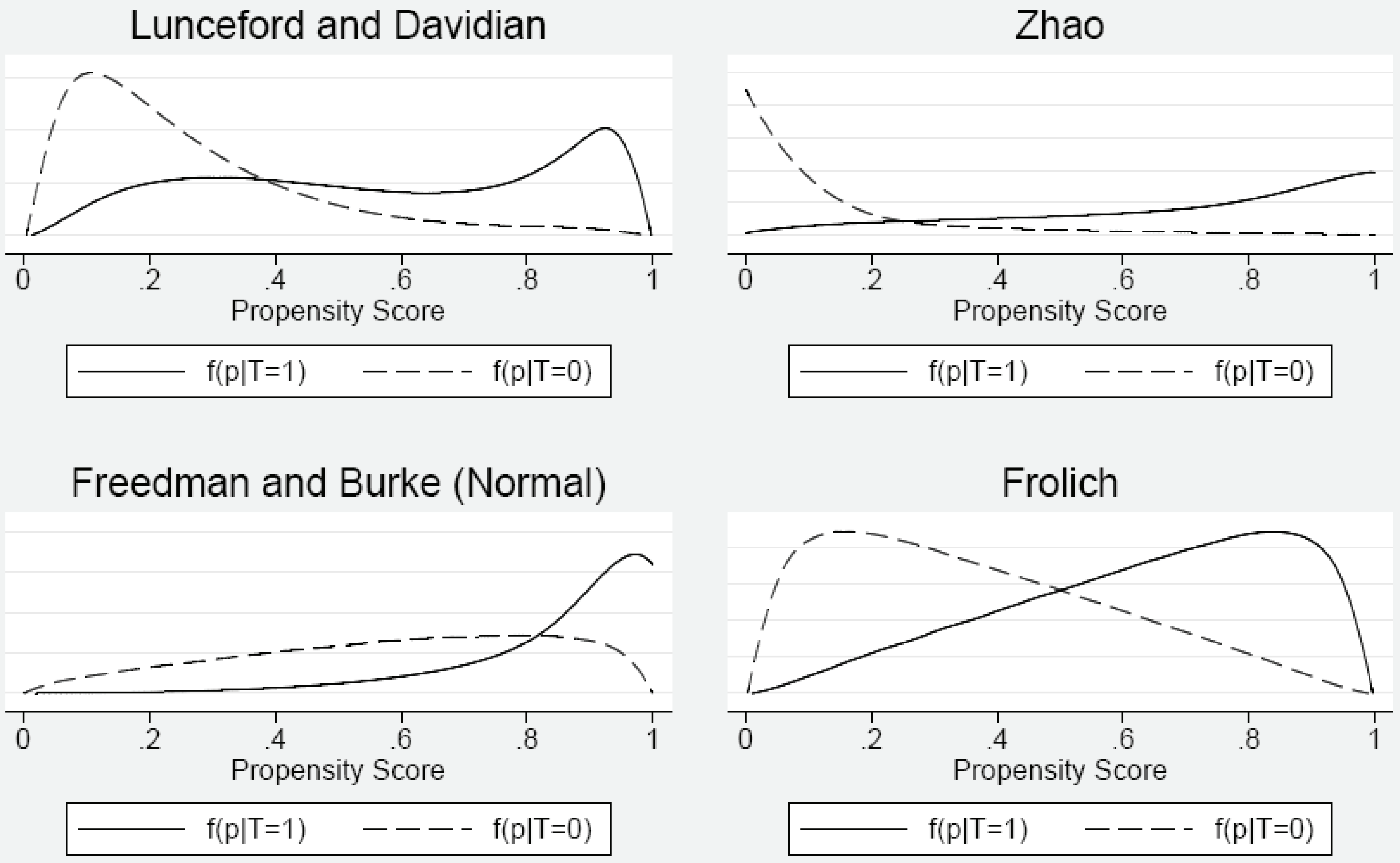


Figure 3.4: Bias of IPW2

Under different degrees of correlation between TE and p(x)

Figure 3.5: Overlap Plots in Related Literature

Examples of Failure of Strict Overlap



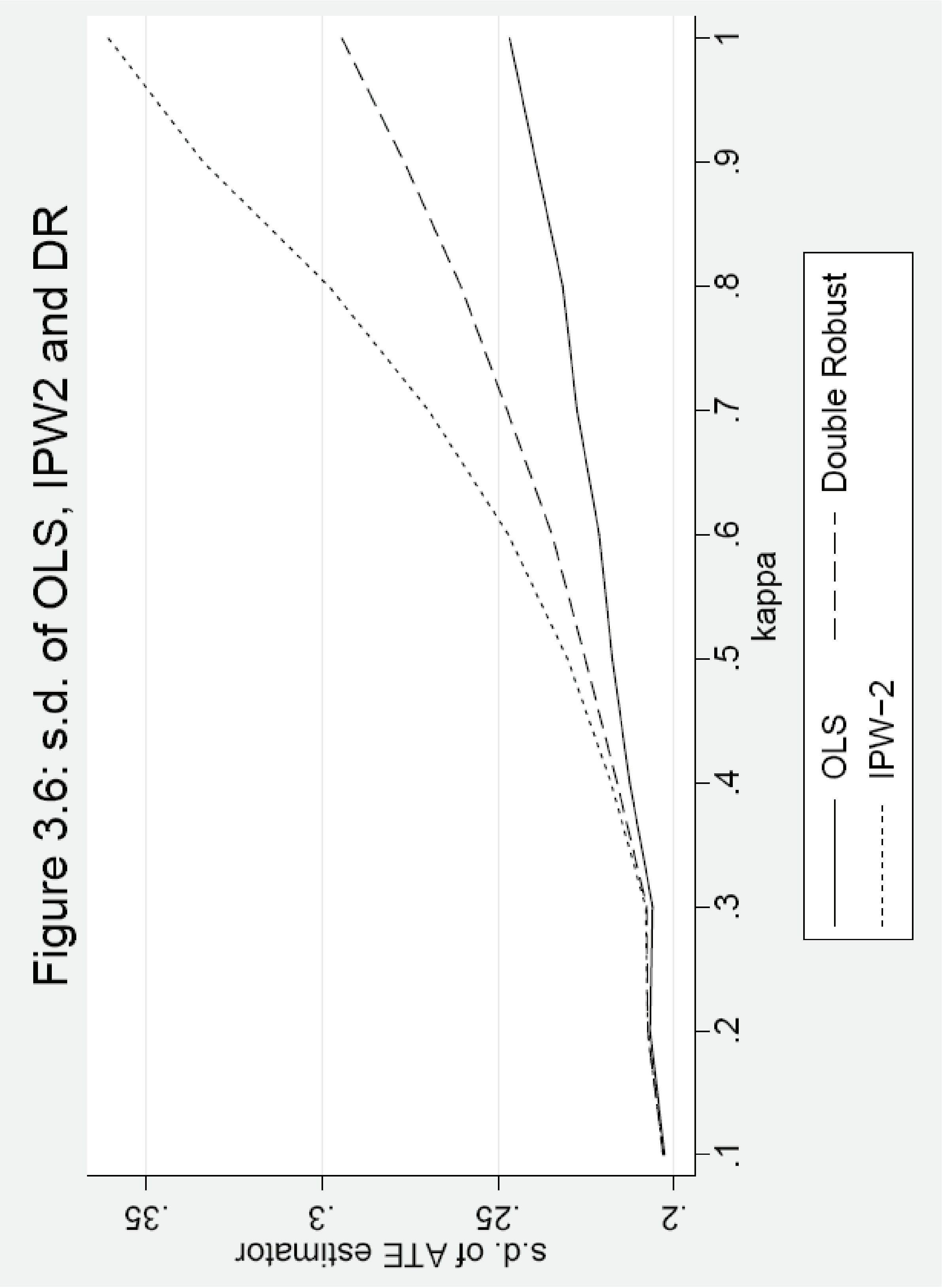


Figure 3.6: s.d. of OLS, IPW2 and DR

Appendix I. IPW3 for TOT

Lunceford and Davidian (2004) derive the IPW3 estimator for the case of ATE. Here we show the derivation of an IPW3 estimator for the TOT. For simplicity of notation in the displays, define $\pi = \hat{p}$ and $\hat{p}_i = \hat{p}(X_i)$.

$$\begin{aligned}
\hat{\theta}_{IPW1} &= \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\frac{1}{\pi}Y_i \equiv \overline{Y}(1) - \hat{\nu}_{0,i\hat{p}w1}, \\
\hat{\theta}_{IPW2} &= \sum_{i=1}^{n}\frac{T_i}{\pi}Y_i - \left(\sum_{i=1}^{n}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\frac{1}{\pi}\right)^{-1}\sum_{i=1}^{n}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\frac{1}{\pi}Y_i \equiv \overline{Y}(1) - \hat{\nu}_{0,i\hat{p}w2}.
\end{aligned}$$

We can combine these two estimators by optimally weighting $\hat{\nu}_{0,i\hat{p}w1}$ and $\hat{\nu}_{0,i\hat{p}w2}$ as follows:

Step 1: Nest $\hat{\nu}_{0,IPW1}$ and $\hat{\nu}_{0,IPW2}$ in one moment condition and solve for $\hat{\nu}_0(\eta_0)$

Step 2: Find η_0^* that minimizes the large sample variance of $\hat{\nu}_0(\eta_0)$

Step 3: Find $\hat{\nu}_{0IPW3} = \theta(\eta_0^*)$.

Step 1: First Moment Condition

The moment condition that yields $\hat{\nu}_{0,IPW2}$ as a solution is

$$\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\,(Y_i-\nu_0)}{1-\hat{p}_i} = 0.$$

In order to nest $\hat{\nu}_{0,IPW1}$, introduce a fake parameter η_0

$$\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\,(Y_i-\nu_0)}{1-\hat{p}_i} - \eta_0 B_i = 0,$$

and then find B_i such that when $\eta_0 = \nu_0$ the solution of the moment condition is $\hat{\nu}_{0,i\hat{p}w1}$. That is

$$\underbrace{\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)Y_i}{1-\hat{p}_i}}_{n\ \hat{\nu}_{0,i\hat{p}w1}} - \sum_{i=1}^{n}\left[\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\nu_0 + \nu_0 B_i\right] = 0.$$

So we want

$$\sum_{i=1}^{n}\left[\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\nu_0 + \nu_0 B_i\right] = n\ \nu_0$$

or

$$\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}+B_i=1\;\Rightarrow\;B_i=1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}.$$

Thus,

$$\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i-\nu_0\right)}{1-\hat{p}_i}-\eta_0\left(1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)=0 \tag{3.18}$$

solves for $\hat{\nu}_{0,IPW1}$ when $\eta_0=\nu_0$ and it solves for $\hat{\nu}_{0,IPW2}$ when $\eta_0=0$. The solution to (3.18) is:

$$\tilde{\nu}_0\left(\eta_0\right)=\left[\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right]^{-1}\sum_{i=1}^{n}\left[\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)Y_i}{1-\hat{p}_i}-\eta_0\left(1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)\right]. \tag{3.19}$$

Step 2: Find η_0^* that minimizes the variance of (3.19)

Subtract ν_0 from both sides to get:

$$\begin{aligned}\tilde{\nu}_0-\nu_0 &= \left[\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right]^{-1}\sum_{i=1}^{n}\left[\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)Y_i}{1-\hat{p}_i}-\eta_0\left(1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)\right]\\ &\quad -\underbrace{\left[\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right]^{-1}\left[\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right]}_{1}\nu_0\\ &= \left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right]^{-1}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i-\nu_0\right)}{1-\hat{p}_i}-\eta_0\left(1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)\right].\end{aligned}$$

Thus,

$$\sqrt{n}\left(\tilde{\nu}_0-\nu_0\right)=\left[\frac{1}{n}\sum_{i=1}^{n}C_i\right]^{-1}\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left\{A_i-\eta_0B_i\right\}$$

where,

$$A_i=\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i-\nu_0\right)}{1-\hat{p}_i}\;;\;B_i=1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i};\;C_i=\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\;and\,E\left[B_i\right]=0;\;E\left[C_i\right]=1$$

Using a LLN and continuity of probability limits, $\left[\frac{1}{n}\sum_{i=1}^{n}C_i\right]^{-1}\xrightarrow{p}1$ and since $E\left[A_i-\eta_0B_i\right]=0$ using a CLT $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left\{A_i-\eta_0B_i\right\}\xrightarrow{d}N\left(0,V\left[A_i-\eta_0B_i\right]\right)$. Consequently,

$$\sqrt{n}\left(\tilde{\mu}_0-\mu_0\right)\xrightarrow{d}N\left(0,V\left[A_i-\eta_0B_i\right]\right).$$

Minimizing the variance with respect to $\eta_0$is the same as finding the least squares estimator of η_0 in a regression of A_i on B_i. That is,

$$\eta_0^* = Cov\left[A_i, B_i\right]\left(V\left[B_i\right]\right)^{-1},$$

where since $E\left[B_i\right] = 0$,

$$\begin{aligned} V\left[B_i\right] &= E\left[\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)^2\right], \\ Cov\left[A_i, B_i\right] &= E\left[\left(\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i - \nu_0\right)}{1-\hat{p}_i}\right)\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)\right]. \end{aligned}$$

Thus,

$$\eta_0 = \left(E\left[\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)^2\right]\right)^{-1} E\left[\left(\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i - \nu_0\right)}{1-\hat{p}_i}\right)\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)\right],$$

which can be rewritten as

$$E\left[\left(\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i - \nu_0\right)}{1-\hat{p}_i}\right)\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right) + \eta_0^*\left[\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)^2\right]\right] = 0.$$

This suggests a second moment condition

$$(3.20) \quad \sum_{i=1}^{n}\left(\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i - \nu_0\right)}{1-\hat{p}_i}\right)\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right) + \eta_0\left[\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)^2\right] = 0.$$

Step 3. Solve system (3.18) and (3.20)

Using (3.20)

$$\eta_0 = \left[\sum_{i=1}^{n}\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)^2\right]^{-1}\left[\sum_{i=1}^{n}\left(\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i - \nu_0\right)}{1-\hat{p}_i}\right)\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)\right].$$

Thus,

$$\begin{aligned} \eta_0\left[\sum_{i=1}^{n}\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)\right] &= \underbrace{\left[\sum_{i=1}^{n}\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)^2\right]^{-1}\left[\sum_{i=1}^{n}\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)\right]}_{C_0} \times \\ &\quad \times\left[\sum_{i=1}^{n}\left(\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i - \nu_0\right)}{1-\hat{p}_i}\right)\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)\right] \\ &= C_0\sum_{i=1}^{n}\left(\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i - \nu_0\right)}{1-\hat{p}_i}\right)\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right) \\ &= \sum_{i=1}^{n}\left(\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i - \nu_0\right)}{1-\hat{p}_i}\right)\left(1 - \frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)C_0. \end{aligned}$$

Substituting in (3.18)

$$\begin{aligned}\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i-\nu_0\right)}{1-\hat{p}_i}-\left(\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i-\nu_0\right)}{1-\hat{p}_i}\right)\left(1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)C_0 &= 0,\\ \sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)\left(Y_i-\nu_0\right)}{1-\hat{p}_i}\left(1-\left(1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)C_0\right) &= 0.\end{aligned}$$

Therefore

$$\begin{aligned}\hat{\nu}_{0,IPW3}=&\left[\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\left(1-\left(1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)C_0\right)\right]^{-1}\times\\ &\times\sum_{i=1}^{n}\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)Y_i}{1-\hat{p}_i}\left(1-\left(1-\frac{1}{\pi}\frac{\hat{p}_i(1-T_i)}{1-\hat{p}_i}\right)C_0\right).\end{aligned}$$

Appendix II. Derivation of SEB for proposed DGPs

This appendix derives the SEBs of Hahn (1998) for the DGPs we study. As noted in the text, we assume that both X_i and e_i follow independent standard normal distributions. For u_i, we consider two possible distributions, standard Cauchy and standard normal. For the standard Cauchy case, we have $F(u)=\frac{1}{\pi}\arctan(X_i)\frac{1}{2}$ and for the standard normal case, we have $F(u)=\Phi(u)$.

Theorems 1 and 2 of Hahn (1998) provide SEBs for ATE and TOT. We next compute the terms of these bounds for the case of unknown propensity score for the DGPs specified in equations (3.10) to (3.12) The functional form of these bounds are given in the text in equations (3.6) to (3.7).

Let σ^2 denote the variance of e_i. Then we have

$$\begin{aligned}Y_i^1 &= \beta+(\gamma+\delta)\,m(p(X_i))+(1+\psi p(X_i))\,e_i,\\ E\left[Y_i^1|X_i\right] &= \beta+(\gamma+\delta)\,m(p(X_i)),\\ \sigma_1^2(X_i) &= E\left[\left(Y_i^1-E\left[Y_i^1|X_i\right]\right)^2|X_i\right]=\left(1+\psi p(X_i)\right)^2\sigma^2,\\ Y_i^0 &= \gamma m(p(X_i))+(1-\psi(1-p(X_i)))e_i,\\ E\left[Y_i^0|X_i\right] &= \gamma m(p(X_i)),\\ \sigma_0^2(X_i) &= E\left[\left(Y_i^0-E\left[Y_i^0|X_i\right]\right)^2|X_i\right]=\left(1-\psi(1-p(X_i))\right)^2\sigma^2,\\ \tau\left(X_i\right) &= \beta+\delta m\left(p\left(X_i\right)\right).\end{aligned}$$

In order to compute the SEBs we need to find the expectation of the following functions:

1. Expressions involving the variance of the treated (first term of the SEB):

$$\begin{aligned}\frac{\sigma_1^2(X_i)}{p(X_i)} &= \frac{\left(1+\psi p(X_i)\right)^2\sigma^2}{p(X_i)} = \sigma^2\left(\frac{1}{p(X_i)} + 2\psi + \psi^2 p\left(X_i\right)\right), \\ \sigma_1^2(X_i)p(X_i) &= \left(1+\psi p(X_i)\right)^2\sigma^2 p\left(X_i\right) = \sigma^2\left(\psi^2 p\left(X_i\right)^3 + 2\psi p(X_i)^2 + p\left(X_i\right)\right).\end{aligned}$$

2. Expressions involving the variance of the controls (second term of the SEB)

$$\begin{aligned}\frac{\sigma_0^2(X_i)}{1-p(X_i)} &= \sigma^2\frac{\left(1-\psi(1-p(X_i))\right)^2}{1-p(X_i)} \\ &= \sigma^2\left(\frac{1}{1-p(X_i)} + \psi^2(1-p(X_i)) - 2\psi\right), \\ \frac{\sigma_0^2(X_i)p(X_i)^2}{1-p(X_i)} &= \sigma^2\frac{\left(1-\psi(1-p(X_i))\right)^2 p(X_i)^2}{1-p(X_i)} \\ &= \sigma^2\left(\frac{p(X_i)^2}{1-p(X_i)} + \psi^2(1-p(X_i))p(X_i)^2 - 2\psi p(X_i)^2\right).\end{aligned}$$

3. Thus the first 2 terms of the ATE and the TOT are

$$\begin{aligned}\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1-p(X_i)} &= \sigma^2\left(\frac{1}{p\left(X_i\right)} + \frac{1}{1-p\left(X_i\right)} + \psi^2\right), \\ \sigma_1^2(X_i)p(X_i) + \frac{\sigma_0^2(X_i)p(X_i)^2}{(1-p(X_i))} &= \sigma^2\left(p\left(X_i\right) + \frac{p\left(X_i\right)^2}{1-p\left(X_i\right)} + \psi^2 p\left(X_i\right).^2\right)\end{aligned}$$

4. Expressions capturing the heterogeneity of the treatment (third term of the SEB)

$$\begin{aligned}(\tau(X_i)-\alpha)^2 &= (\beta-\alpha)^2 + \delta^2\left[m\left(p\left(X_i\right)\right)\right]^2 + \\ &\qquad +2\left(\beta-\alpha\right)\delta m\left(p\left(X_i\right)\right), \\ p(X_i)(\tau(X_i)-\theta)^2 &= (\beta-\theta)^2 p\left(X_i\right) + \delta^2\left[m\left(p\left(X_i\right)\right)\right]^2 p\left(X_i\right) + \\ &\qquad +2\left(\beta-\theta\right)\delta m\left(p\left(X_i\right)\right)p\left(X_i\right), \\ p(X_i)^2(\tau(X_i)-\theta)^2 &= (\beta-\theta)^2 p\left(X_i\right)^2 + \delta^2\left[m\left(p\left(X_i\right)\right)\right]^2 p\left(X_i\right)^2 + \\ &\qquad +2\left(\beta-\theta\right)\delta m\left(p\left(X_i\right)\right)p\left(X_i\right)^2.\end{aligned}$$

Therefore computation of the SEB involves computation of the integrals

1. For $k=-1,1,2$

$$A_1\left(k\right) = E\left[p\left(X_i\right)^k\right] = \int_{-\infty}^{\infty}\left[F_u\left(\eta+\kappa X_i\right)\right]^k\ f_X\left(x\right)dx.$$

2. For $k = 0, 2$

$$A_2(k) = E\left[\frac{p(X_i)^k}{1 - p(X_i)}\right] = \int_{-\infty}^{\infty} \frac{\left[F_u(\eta + \kappa X_i)\right]^k}{1 - F_u(\eta + \kappa X_i)}\; f_X(x)\, dx.$$

3. For $h = 1, 2$; $k = 0, 1, 2$; $j = 1, 2$

$$\begin{aligned} A_3(h,k) &= E\left[m_j\left(p(X_i)\right)^h p(X_i)^k\right] \\ &= \int_{-\infty}^{\infty} \left[m_j\left(F_u(\eta + \kappa X_i)\right)^h\right]^h \left[F_u(\eta + \kappa X_i)\right]^k\; f_X(x)\, dx. \end{aligned}$$

Aside from $A_1(-1)$ and $A_2(0)$, these integrals are readily computed using mathematical software or simulation. In the case where u_i is distributed standard normal and κ is close to 1, the 2 integrals listed are highly difficult to compute with any accuracy. For this case, we use an approach suggested to us by Matias Cattaneo. Consider $A_1(-1)$ and $A_2(0)$ in the case with u_i distributed standard normal. We have, for $c = 1/|\kappa| > 1$,

$$\begin{aligned} E\left[\frac{1}{p(X_i)}\right] &= \int_{-\infty}^{\infty} \frac{\phi(x)}{\Phi(\eta + \kappa x)} dx = c\int_{-\infty}^{0} \frac{\phi(c(t-\eta))}{\Phi(t)} dt + c\int_{0}^{\infty} \frac{\phi(c(t-\eta))}{\Phi(t)} dt \\ &= c\int_{0}^{\infty} \frac{\phi(c(t+\eta))}{1-\Phi(t)} dt + c\int_{0}^{\infty} \frac{\phi(c(t-\eta))}{\Phi(t)} dt \qquad (3.21) \\ E\left[\frac{1}{1-p(X_i)}\right] &= \int_{-\infty}^{\infty} \frac{\phi(x)}{1-\Phi(\eta + \kappa x)} dx = c\int_{0}^{\infty} \frac{\phi(c(t-\eta))}{1-\Phi(t)} dt + c\int_{-\infty}^{0} \frac{\phi(c(t-\eta))}{1-\Phi(t)} dt \\ &= c\int_{0}^{\infty} \frac{\phi(c(t-\eta))}{1-\Phi(t)} dt + c\int_{0}^{\infty} \frac{\phi(c(t+\eta))}{\Phi(t)} dt. \qquad (3.22) \end{aligned}$$

The second integral in (3.21) and (3.22) is easy to simulate accurately because the numerator has rapidly declining tails and the denominator is always between 0.5 and 1. The first integral in both expressions is very difficult to simulate because the denominator is near zero for much of the domain. To handle this problem, we break the first integral into 2 pieces

$$c\int_{0}^{\infty} \frac{\phi(c(t+b))}{1-\Phi(t)} dt = c\int_{0}^{a} \frac{\phi(c(t+b))}{1-\Phi(t)} dt + c\int_{a}^{\infty} \frac{\phi(c(t+b))}{1-\Phi(t)} dt, \tag{3.23}$$

for $b \in \{-\eta, \eta\}$ and a a moderate number such as 5. It is easy to simulate $\int_0^a \frac{\phi(c(t+b))}{1-\Phi(t)} dt$, and we can directly bound $\int_a^\infty \frac{\phi(c(t+b))}{1-\Phi(t)} dt$ using the inequality

$$\frac{t}{1+t^2}\phi(t) < 1 - \Phi(t) < \frac{1}{t}\phi(t), \tag{3.24}$$

which is valid for any $t > 0$ and is highly accurate for any t above 4.

Applying the inequality, we have

$$\int_a^\infty \frac{\phi(c(t+b))}{1-\Phi(t)}dt \quad > \quad \int_a^\infty t\frac{\phi(c(t+b))}{\phi(t)}dt \tag{3.25}$$

$$> \quad \int_a^\infty t\exp\left(-\frac{1}{2}\left(c^2(t+b)^2-t^2\right)\right)dt \equiv LB. \tag{3.26}$$

Completing squares, we have

$$LB \quad = \quad \int_a^\infty t\exp\left[-\frac{1}{2}\left(c^2-1\right)\left(t^2+\frac{c^2b^2}{(c^2-1)}+2\frac{c^2b}{(c^2-1)}t\right)\right]dt \tag{3.27}$$

$$= \quad \int_a^\infty t\exp\left[-\frac{1}{2}\frac{1}{\sigma^2}\left(t^2-2\mu t+\mu^2\right)+\frac{\mu\left(b+\mu\right)}{2\sigma^2}\right]dt \tag{3.28}$$

$$= \quad \exp\left[\frac{\mu\left(b+\mu\right)}{2\sigma^2}\right]\int_a^\infty t\exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right]dt \tag{3.29}$$

$$= \quad \exp\left[\frac{\mu\left(b+\mu\right)}{2\sigma^2}\right]\sqrt{2\pi\sigma^2}\int_d^\infty (\sigma t+\mu)\phi\left(t\right)dt \tag{3.30}$$

$$= \quad \exp\left[\frac{\mu\left(b+\mu\right)}{2\sigma^2}\right]\sqrt{2\pi\sigma^2}\left[\sigma\phi\left(d\right)+\mu\left[1-\Phi\left(d\right)\right]\right], \tag{3.31}$$

where $\mu = -c^2b/(c^2-1)$, $\sigma^2 = 1/(c^2-1)$, and $d = \left(a-\mu\right)/\sigma$.

For the upper bound, apply the inequality again to obtain

$$\int_a^\infty \frac{\phi(c(t+b))}{1-\Phi(t)}dt \quad < \quad \int_a^\infty \frac{1+t^2}{t}\frac{\phi(c(t+b))}{\phi(t)}dt \tag{3.32}$$

$$< \quad \int_a^\infty \frac{1}{t}\frac{\phi(c(t+b))}{\phi(t)}dt+\int_a^\infty t\frac{\phi(c(t+b))}{\phi(t)}dt \tag{3.33}$$

$$< \quad \frac{1}{a}\int_a^\infty \frac{\phi(c(t+b))}{\phi(t)}dt+LB \equiv UB. \tag{3.34}$$

It is easy to show that

$$\int_a^\infty \frac{\phi(c(t+b))}{\phi(t)}dt \quad = \quad \int_a^\infty \frac{\phi(c(t+b))}{\phi(t)}dt = \int_a^\infty \exp\left(-\frac{1}{2}\left(c^2(t+b)^2-t^2\right)\right)dt \tag{3.35}$$

$$= \quad \exp\left[\frac{\mu\left(b+\mu\right)}{2\sigma^2}\right]\sqrt{2\pi\sigma^2}\int_a^\infty \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right]dt \tag{3.36}$$

$$= \quad \exp\left[\frac{\mu\left(b+\mu\right)}{2\sigma^2}\right]\sqrt{2\pi\sigma^2}\left(1-\Phi\left(d\right)\right). \tag{3.37}$$

In summary, we have

$$LB < \int_a^\infty \frac{\phi\left(c(t+b)\right)}{1-\Phi(t)}dt < UB \tag{3.38}$$

where,

$$LB = \exp\left[\frac{\mu(b+\mu)}{2\sigma^2}\right]\sqrt{2\pi\sigma^2}\left[\sigma\phi(d)+\mu\left[1-\Phi(d)\right]\right], \tag{3.39}$$

$$UB = LB+\exp\left[\frac{\mu(b+\mu)}{2\sigma^2}\right]\frac{\sqrt{2\pi\sigma^2}}{a}\left(1-\Phi(d)\right), \tag{3.40}$$

and

$$\mu=\frac{-c^2b}{(c^2-1)};\ \sigma^2=\frac{1}{(c^2-1)};d=\frac{a-\mu}{\sigma};\ c=\frac{1}{\kappa};\ b\in\{-\eta,\eta\}\,. \tag{3.41}$$

For example, for $\eta = -0.3, \kappa = 0.8$ and $a = 7$ the lower bound for this integral is 0.0000515 and the upper bound is 0.0000526. For $\eta = 0, \kappa = 0.95$ and $a = 7$ the lower bound is 0.656062 and the upper bound is 0.667721.

Appendix Figure 3.1 graphs the upper and lower bounds on the integral above, for $\eta = 0$, $a = 7$, and κ ranging from 0.9 to 0.999. This graph makes two points. First, the bounds are extremely accurate globally in κ. Second, the integral in question is an amazingly convex function of κ. Highly similar patterns hold for other values of η.

Bibliography

Abadie, Alberto and Guido W. Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, January 2006, *74* (1), 235–267.

——— and ——— , "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, forthcoming 2008.

Black, Dan A. and Jeffrey A. Smith, "How Robust is the Evidence on the Effects of College Quality? Evidence From Matching," *Journal of Econometrics*, July-August 2004, *121* (1-2), 99–124.

Blinder, Alan S., "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, Fall 1973, *8*, 436–455.

Card, David, "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 1996, *64*, 957–979.

Card, David E., "The Causal Effect of Education on Earnings," in Orley Ashenfelter and David E. Card, eds., *The Handbook of Labor Economics*, Vol. 3A, Amsterdam: Elsevier, 1999.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi, "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics*, forthcoming 2008.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik, "Dealing with Limited Overlap in Estimation of Average Treatment Effects," Unpublished manuscript, UCLA 2007.

——— , ——— , ——— , and ——— , "Nonparametric Tests for Treatment Effect Heterogeneity," *Review of Economics and Statistics*, forthcoming 2007.

Deaton, Angus, *The Analysis of Household Surveys : A Microeconomic Approach to Development Policy*, Washington, D.C.: World Bank, 1997.

Dehejia, Rajeev H. and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," in "Econometric Methods for Program Evaluation," Cambridge: Rajeev H. Dehejia, Ph.D. Dissertation, Harvard University, 1997, chapter 1.

——— and ———, "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, December 1999, *94* (448), 1053–1062.

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

Fan, Jianqing, "Local Linear Regression Smoothers and Their Minimax Efficiencies," *Annals of Statistics*, March 1993, *21* (1), 196–216.

Freedman, David A. and Richard A. Berk, "Weighting Regressions by Propensity Scores," *Evaluation Review*, *32* (4).

Frölich, Markus, "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, February 2004, *86* (1), 77–90.

Hahn, Jinyong, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, March 1998, *66* (2), 315–331.

Haviland, Amelia M. and Daniel S. Nagin, "Causal Inferences with Group Based Trajectory Models," *Psychometrika*, September 2005, *70* (3), 1–22.

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

——— and R. Robb, "Alternative Methods for Evaluating the Impact of Interventions," in James J. Heckman and R. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press Cambridge 1985.

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

———, ———, and ———, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, April 1998, *65* (2), 261–294.

———, ———, Jeffrey Smith, and Petra Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, September 1998, *66* (5), 1017–1098.

———, Sergio Urzua, and Edward Vytlacil, "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, August 2006, *88* (3), 389–432.

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

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart, "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, August 2007, *15*, 199–236.

Horvitz, D. and D. Thompson, "A Generalization of Sampling Without Replacement from a Finite Population," *Journal of the American Statistical Association*, 1952, *47*, 663–685.

Imbens, Guido W., "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, February 2004, *86* (1), 4–29.

Johnston, Jack and John E. DiNardo, *Econometric Methods*, McGraw-Hill, 1996.

Jose, Jeffrey Smith Galdo and Dan Black, "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data," Unpublished manuscript, University of Michigan 2007.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman, "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment," *Quarterly Journal of Economics*, May 2001, *116* (2), 607–654.

Kent, David and Rodney Hayward, "Subgroup analyses in clinical trials.," *New England Journal of Medicine*, Mar 2008, *358* (11), 1199.

Khan, Shakeeb and Elie Tamer, "Irregular Identification, Support Conditions, and Inverse Weight Estimation," Unpublished manuscript, Northwestern University 2007.

Lechner, Michael, "A Note on the Common Support Problem in Applied Evaluation Studies," Discussion Paper N2001-01, Universität St. Gallen 2001.

——, "A Note on Endogenous Control Variables in Evaluation Studies," Discussion Paper N2005-16, Universität St. Gallen 2005.

Loader, Clive R., "Bandwidth Selection: Classical or Plug-In?," *The Annals of Statistics*, Apr 1999, *27* (2), 415–438.

Lunceford, Jared K. and Marie Davidian, "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study," *Statistics in Medicine*, 15 October 2004, *23* (19), 2937–2960.

McCrary, Justin, "The Effect of Court-Ordered Hiring Quotas on the Composition and Quality of Police," *American Economic Review*, March 2007, *97* (4), 318–353.

——, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, February 2008, *142* (2).

Muirhead, Robb J., *Aspects of Multivariate Statistical Theory*, Hoboken: John Wiley and Sons, 2005.

Newey, Whitney, "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, April-June 1990, *5* (2), 99–135.

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

Robins, James M. and Andrea Rotnitzky, "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, March 1995, *90* (429), 122–129.

——, ——, and Lue Ping Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, September 1994, *89* (427), 846–866.

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

Ruud, Paul A., *An Introduction to Classical Econometric Theory*, New York: Oxford University Press, 2000.

Seifert, Burkhardt and Theo Gasser, "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*, June 2000, *9* (2), 338–360.

Smith, Jeffrey A. and Petra Todd, "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, September 2005, *125* (1–2), 305–353.

Stone, Mervyn, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society, Series B*, 1974, *36* (2), 111–147.

Todd, Petra, "Matching Estimators," in P. Newman, M. Milgate, and J. Eatwell, eds., *The New Palgrave—A Dictionary of Economics*, Vol. forthcoming, New York: Macmillan, 2007.

Wishart, John, "The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population," *Biometrika*, July 1928, *20A* (1/2), 32–52.

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

Zhao, Zong, "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence," *Review of Economics and Statistics*, February 2004, *86* (1), 91–107.

——, "Sensitivity of Propensity Score Method to the Specifications," *Economics Letters*, 2008, *98*, 309–319.

Chapter 4

A Sequential Method of Moments Variance Estimator of Weighting Estimators of Average Treatment Effects[1]

Inverse probability weighting estimation (IPW) has been widely used in economics to generate counterfactual distributions (e.g. DiNardo et al. 1996, Biewen 2001, Barsky, Bound, Charles and Lupton 2002, Bailey and Collins 2006, among others) and to estimate average treatment effects (e.g. Dehejia and Wahba 1997, McCrary 2007, Busso and Kline 2008 and Levinsohn, McLaren and Zuma 2008, and others). Despite their popularity, obtaining valid standard errors of IPW estimators of average treatment effects is complicated because they belong to the family of two-step non-linear estimators. The first step usually involves estimating a parametric binary choice model in order to estimate the conditional probability of treatment, typically estimated by maximum likelihood. The second step involves the estimation of the treatment effect using weights which are function of the probability of treatment computed in the first stage. This second stage can be implemented via a weighted least squares regression.

I show that IPW estimators of average treatment effects can be cast as a simple sequential method of moments (SQMM) estimator whose asymptotic distribution is derived in Newey (1984).[2] This result is not entirely new. Wooldridge (2007) studies IPW estimation for general missing data problems, of which treatment effect estimation is a special case. He provides a general expression for the asymptotic variance of this family of estimators. Hirano and Imbens (2002) propose an estimator of the variance of IPW average treatment effect estimators that is, in essence, very similar to the one proposed in this paper. Their result, however, apply to cases in which the weights involved in the estimation do not add up to one and it is only considers a logit first stage.

This paper has three contributions. First, I derive an SQMM variance estimator of average treatment effect estimators using a general parametric first stage, and allowing

[1] I would like to thank John DiNardo, Jesse Gregory, Patrick Kline, Justin McCrary, Serena Ng and Jeff Smith for useful comments and corrections. Any errors are my own.

[2] More generally, the estimator will be of the class of two-step extremum estimators.

for the weights to add up to one, which is in practice the most usual implementation of IPW estimation. It is therefore more general than Hirano and Imbens (2002), but a special case of Wooldridge (2007). Second, I note that the SQMM variance estimator can be used to test not only hypotheses about the treatment effects for a given outcome but also hypotheses involving multiple outcomes and hypotheses about different estimands. Third, I show using Monte Carlo simulations that tests that use the SQMM variance estimator have good finite sample size and power compared to competing inference strategies.

Since the influential work by Rosenbaum and Rubin (1983) there has been substantial interest in estimators of average treatment effects that use the propensity score, or conditional probability of treatment, to construct a balanced sample of treated and control units in order to estimate treatment effects. Examples of these estimators are matching, blocking, and IPW estimators. The latter has three advantages over the first two. First, IPW does not require the selection of tuning parameters (i.e. number of neighbors, bandwidth, or block size selection). Secondly, Hirano et al. (2003) show that IPW estimators that utilize a non-parametric estimate of the propensity score achieve the semiparametric efficiency bound (SEB). [3] Finally, Busso, McCrary and DiNardo (2008) find in a Monte Carlo study that IPW estimators are unbiased in small samples and their variance is very close to the semiparametric efficiency bound, whereas matching and blocking tend to be biased in small samples.

In order to be consistent for an average treatment effect, IPW estimators require that the propensity score converges in probability to the true conditional probability of treatment. Typically researchers assume a flexible parametric functional form on the propensity score model. Usually a probit, a logit or a linear probability model is assumed, and polynomials and interaction terms of the covariates that determine treatment are used as explanatory variables.[4]

Estimation of the variance of IPW average treatment effects estimators is complicated by the fact that we need to take into account that the propensity score, typically unknown, has to be estimated. This implies that the weights involved in the computation of the IPW estimator have sampling variability on their own. Some empirical applications have ignored altogether the sampling variability of the propensity score (e.g. McCrary 2007, Levinsohn, McLaren and Zuma 2008). Hirano et al. (2003) show that reweighting using the true

[3] There is no similar result for matching estimators. The only exception is the nearest-neighbor matching estimator that Abadie and Imbens (2006) show does not achieve the SEB when the number of neighbors is finite.

[4] Robins and Rotnitzky (1995) propose a doubly robust estimator that can mitigate the selection bias that arises from a misspecified propensity score model. This method requires specification of two models. First, the researcher needs to specify the propensity score equation that characterizes the selection mechanism. Second, the outcome equation model, describing the population response to treatment, has to be specified. The key feature of this estimator is that it remains consistent if one of the two models is correctly specified.

propensity score does not achieve the SEB, so in principle a strategy that ignores the fact that the propensity score is estimated should lead in large enough samples to conservative inference. However, their result is only valid in large samples. No formal result exists to guarantee conservative inference in finite samples so it remains an open question whether or not we need to correct the estimator of the variance of IPW estimators. Acknowledging that the propensity score is being estimated may lead to better inference.

Imbens (2004) discusses three alternatives to compute the variance of any semiparametric estimator of treatment effects that is based on the propensity score. One possibility is to construct an estimate of the SEB using kernel methods or series. A second method is to use bootstrapping as done in Busso and Kline (2008). A third possibility, only available when the propensity score is estimated parametrically, is to calculate the contribution of the sample variability of the propensity score to the variance of the average treatment effects estimators.

This paper presents a simple approach for constructing an estimator of the variance of reweighting estimators of average treatment effects using the fact that reweighting estimators are two-step method of moments estimators for which a general simple form of the asymptotic variance already exists (Newey 1984). This approach has several advantages. First, it is easy to implement and therefore inference about the average treatment effects on a given outcome becomes costless. Second, the approach allows the researcher to test hypotheses that are usually ignored but potentially of interest. For example, it is simple to test the equality of the treatment effects on multiple outcomes. It is also easy to test the equality of the ATE and the TOT for a given outcome; a test that can be interpreted as a test of the homogeneity of the treatment effect.

I perform a series of Monte Carlo simulations to study the finite sample properties of these tests. I find that ignoring the fact that the weights are estimated can lead, in small samples, to severe over-rejection of the null of no treatment effect. The SQMM variance estimator solves this problem, a result that is robust in different data generating processes. I also compare the size and power of tests using the SQMM variance to those obtained when using other a priori valid inference procedures such as percentile and percentile-t bootstrap. The percentile bootstrap does not perform well. Tests based on the percentile-t bootstrap method using a bootstrap SQMM variance have very similar size and power properties to the ones obtained using the asymptotic SQMM variance. I interpret this as an indication that the bootstrap percentile-t method is not providing any refinement to the asymptotic variance, which is indicative that the SQMM variance estimator is providing a good enough approximation to the true variance of the treatment effect estimator, even in samples of size 100.

The remainder of the paper is organized as follows. Section I introduces some basic notation and briefly discusses identification of average treatment effects. Section II presents inverse probability weighting estimators. In section III I derive a sequential method of moments estimator of the variance of reweighting estimators of average treatment effects. In section IV I first describe the data generating process used in the Monte Carlo simulations. Then I discuss alternative inference procedures. Finally, I investigate the size and power properties of different hypotheses tests using the SQMM variance estimator, comparing the results to those obtained when the sample variability of the propensity score is ignored, and with those obtained when a valid bootstrap method is used.

I Notation and Identification

Let $Y_i(1)$ be the outcome unit i would obtain under treatment, $Y_i(0)$ the outcome that it would obtain with no treatment and let T_i denote the binary treatment variable. The pair $(Y_i(0), Y_i(1))$ is never observed jointly; instead we observe $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$. In other words, we observe $Y_i(1)$ for the treated units but $Y_i(0)$ is missing for them and, similarly, we observe only $Y_i(0)$ for the control units. For each individual, we observe a vector of K covariates X_i that are not influenced by treatment. There are a total of n observations in the sample, where n_1 observations are treated and the remainder n_0 do not receive treatment. The data $(X_i, Y_i, T_i)_{i=1}^n$ are taken to be independent across i. We are interested in estimating the average treatment effect $\theta_{ATE} = E[Y_{\,i}\,(1) - Y_i\,(0)\,]$ and the average treatment effect on the treated $\theta_{TOT} = E[\,Y_i\,(1) - Y_i\,(0)|\,T_i = 1]$. Semiparametric estimators of treatment effects require two assumptions for identification.

The first assumption is the conditional independence assumption (CIA) which requires that there only exists selection on observed variables. This means that treatment is randomized given X_i, or that $(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i|X_i$.[5] The CIA is the cornerstone of estimation of treatment effects because it allows the construction of counterfactual means for the outcome of the treated units using information on the control observations.

If X_i includes many covariates then the estimation of average treatment effects involves estimating a very high dimensional conditional expectation for which different tuning parameters might be required for different outcomes. We can simplify estimation considerably by conditioning on a scalar balancing score instead of conditioning on X_i. Let the conditional probability of treatment, or propensity score, be $p(X_{\,i}) \equiv P(T_i = 1|X_i)$ and denote the unconditional treatment probability by p. Theorem 3 of Rosenbaum and Rubin (1983)

[5] This assumption is also known as strong ignorability of treatment assignment, unconfoundedness, or exogeneity. For further discussion see Heckman and Robb (1985), Rosenbaum and Rubin (1983), and Imbens (2004).

shows that if the CIA holds when conditioning on X_i then it will also hold when conditioning on $p\left(X_i\right)$; that is $(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i | p\left(X_i\right)$.[6]

The second assumption required for identification of θ_{TOT} and θ_{ATE} is that of strict overlap which assumes that there are no observations with very high/low values of the propensity score; that is, $\xi < p(x) < 1-\xi$ for almost every x in the support of X_i, for some $\xi > 0$. The strict overlap assumption is stronger than the *standard overlap* assumption that $0 < p(x) < 1$ for almost every x in the support of X_i. Standard overlap, which is often referred to as the common-support condition, states that no value of the covariates can deterministically predict receipt (absence) of treatment. The failure of this condition would allow for the possibility that some observations with particular configurations of covariates would only be capable of being observed in the treatment (no treatment) state, therefore preventing the construction of valid controls.[7] Khan and Tamer (2007) notice that strict overlap guarantees that there are *enough* valid controls for every treated observation, therefore guaranteeing $\sqrt{n}-$consistency.[8]

Imbens (2004) presents a simple proof that these assumptions are sufficient to guarantee identification of θ_{TOT} and θ_{ATE}. The CIA guarantees that we can estimate a counterfactual mean for each outcome of interest, using the observed data: $E\left[Y_i\left(d\right)\middle| X_i\right] = E\left[Y_i\middle| X_i, T_i = d\right]$ for $d = 0, 1$. The standard overlap assumption ensures that we can compute the average of $\theta\left(X_i\right) = E\left[Y_i\middle| X_i, T_i = 1\right] - E\left[Y_i\middle| X_i, T_i = 0\right]$ over the appropriate distribution of X_i. Khan and Tamer (2007) add that strict overlap is sufficient for the estimators based on $\theta\left(X_i\right)$ to be $\sqrt{n}-$consistent.

II Inverse Probability Weighting Estimators

IPW estimators are based on the idea that we can recover expectations of $(Y_i(0), Y_i(1))$ by properly reweighting the data in a manner that balances the distribution of covariates across treated and untreated units. This is accomplished by upweighting control (treated) observations that look like treated (control) units, based upon their covariates which are "summarized" in their probability of treatment $p\left(X_i\right)$. Once the distribution of covariates is balanced across treatment and control groups a simple comparison of weighted means will, under the two assumptions made thus far, identify θ_{TOT} and θ_{ATE}.[9]

[6] Heckman and Hotz (1989) point out that propensity score approaches do not completely bypass the curse of dimensionality because the function $p\left(X_i\right)$ is typically unknown and has to be estimated.

[7] For a discussion of the meaning and implications of the overlap assumption see Imbens (2004).

[8] For further discussion of the implications of violating this assumption see Khan and Tamer (2007) and Busso et al. (2008).

[9] This estimator was proposed in the statistic literature by Horvitz and Thompson (1952). It was first used in economics by DiNardo et al. (1996) in a cross-sectional context to compute counterfactual distributions; and later extended to panel settings by Abadie (2005).

IPW estimators are motivated by the following three equalities:[10]

$$\begin{aligned} E\left[Y_i\left(1\right)\right] &= E\left[\frac{T_iY_i}{p\left(X_i\right)}\right], \\ E\left[Y_i\left(0\right)\right] &= E\left[\frac{(1-T_i)Y_i}{1-p\left(X_i\right)}\right], \\ E\left[Y_i\left(0\right)\middle| T_i=1\right] &= E\left[\frac{p\left(X_i\right)}{1-p\left(X_i\right)}\frac{1-p}{p}Y_i\middle| T_i=0\right], \end{aligned}$$

which imply that we can rewrite the estimands as:

$$\begin{aligned} \theta_{ATE} &= E\left[\frac{T_iY_i}{p\left(X_i\right)}-\frac{(1-T_i)Y_i}{1-p\left(X_i\right)}\right], \\ \theta_{TOT} &= E\left[Y_i\middle| T_i=1\right]-E\left[\frac{p\left(X_i\right)}{1-p\left(X_i\right)}\frac{1-p}{p}Y_i\middle| T_i=0\right]. \end{aligned}$$

If $p\left(X_i\right)$ and p are known, we can compute the estimators as the sample analogues of the estimands

$$\hat{\theta}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left[C_1\frac{T_iY_i}{p\left(X_i\right)}-C_0\frac{(1-T_i)Y_i}{1-p\left(X_i\right)}\right], \tag{4.1}$$

$$\hat{\theta}_{TOT} = \frac{1}{n_1}\sum_{i=1}^{n}T_iY_i-\frac{1}{n_0}\sum_{i=1}^{n}C\frac{(1-T_i)\,p\left(X_i\right)}{1-p\left(X_i\right)}\frac{1-p}{p}Y_i, \tag{4.2}$$

where, for the moment, $C=C_1=C_0=1$. These estimators have the problem that, in any given sample, the weights do not add up to one. We can force the weights to add up to one by letting $C=\left(\frac{1}{n_0}\sum_{i=1}\frac{(1-T_i)p(X_i)}{1-p(X_i)}\frac{1-p}{p}\right)^{-1}$, $C_1=\left(\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{p(X_i)}\right)^{-1}$and $C_2=\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-p(X_i)}\right)^{-1}$. I only consider these last IPW estimators because they tend to show better finite sample properties.

In practice $p\left(X_i\right)$ is usually unknown and has to be estimated. Therefore, we proceed in two steps. In the first step we obtain an estimate of the propensity score, $\hat{p}\left(X_i\right)$. In a second step, we use $\hat{p}\left(X_i\right)$ to compute the estimators (4.1) and (4.2).

The first step requires the estimation of a parametric model that will yield a consistent estimator of $p\left(X_i\right)$ provided that we condition on a set of X_i that guarantees satisfaction of the CIA. Thus, researchers usually include in X_i polynomials and interaction of the covariates that make the model flexible enough to make feasible that the CIA holds true. I will assume from now on that the parametric model is such that $\hat{p}\left(X_i\right)$ is consistent for $p\left(X_i\right)$.

[10] These equalities can be shown using the CIA and the strict overlap assumptions in conjunction with iterated expectations. See Appendix I.

Consider a selection mechanism $T_i^* = \kappa X_i + u_i$ where κ is a coefficient vector of dimension K, and the error term u_i is independent of X_i and assumed to have a known symmetric and twice-differentiable distribution $F(\cdot)$. An individual receives treatment if $T_i^* > 0$ so the treatment dummy is $T_i = 1(T_i^* > 0)$. Therefore the probability of receiving treatment is given by $p(X_i) \equiv P(T_i = 1| X_i) = F(\kappa X_i)$. We can then estimate $p(X_i)$ by maximum likelihood: typically a probit or a logit binary choice model is used. The ML estimator $\hat{\kappa}$ solves a sample moment condition that is the sample average of the K individual score vectors; that is, $\hat{\kappa}$ solves,

$$\text{(4.3)} \qquad \bar{g}_1(\kappa) = \frac{1}{n}\sum_{i=1}^{n} v_i \left[T_i - F(\kappa X_i)\right] X_i' = 0,$$

where $\nu_i' = \frac{f(\kappa X_i)}{F(\kappa X_i)[1-F(\kappa X_i)]}$ and $f(\cdot)$ is the density of u_i. Using $\hat{\kappa}$ we can predict the probability of treatment for each observation, $\hat{p}(X_i) = F(\hat{\kappa} X_i)$.

Suppose for simplicity that we are interested in the effect of a treatment on one outcome Y_i only. The generalization to the case of L outcomes is straightforward. The second step of the estimation procedure is to compute the estimators of the ATE and the TOT. Estimators (4.1) and (4.2) can be computed by estimating the coefficient on the treatment indicator in a weighted least squares regression of the outcome on a constant and a treatment indicator using weights $\hat{W}_i^t = T_i \hat{w}_0(i) + (1-T_i)\hat{w}_1(i)$, with $t = ATE, TOT$ and the weights W_i^t defined as

$$\begin{aligned} \text{(4.4)}\ \ \hat{W}_i^{ATE} &= T_i \left(\frac{1}{n}\sum_{i=1}^{n} \frac{T_i}{\hat{p}(X_i)}\right)^{-1} \frac{1}{\hat{p}(X_i)} + (1-T_i)\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1-T_i}{1-\hat{p}(X_i)}\right)^{-1} \frac{1}{1-\hat{p}(X_i)}, \\ \text{(4.5)}\ \ \hat{W}_i^{TOT} &= T_i \left(\frac{1}{n}\sum_{i=1}^{n} T_i\right)^{-1} + (1-T_i)\left(\frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)\hat{p}(X_i)}{1-\hat{p}(X_i)}\right)^{-1} \frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}. \end{aligned}$$

In other words, if we let $\tau_t = [\alpha_t\ \theta_t]'$ for $t = TOT,\ ATE$ and $Z_i = [1\ T_i]$, the estimator of the treatment effects will solve the following sample moment condition:

$$\bar{g}_2^t(\hat{\kappa}, \tau_t) = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \tau_t Z_i)\, \hat{W}_i^t(\kappa)\, Z_i' = 0 \qquad \text{for } t = ATE, TOT.$$

If we are interested in L outcomes then $\dim\left(\bar{g}_2^t\right) = 2L$.

In order to compute the standard error of any of these treatment effect estimators, we need to take into account that the second step uses weights that themselves have sampling variability. It is easy to do this using a sequential method of moments framework. Wooldridge (2007) derives the asymptotic distribution of a general inverse probability weighted M-estimator for general missing data problems. The results that follow are a special case of those presented in Wooldridge (2007).

III Sequential Method of Moments Variance Estimator

We can cast the previous model as an exactly identified sequential method of moments estimation problem. Method of moments exploits population moment conditions of the form $E\left[g\left(X_i,\beta\right)\right]=0$ if and only if $\beta=\beta_0$. This motivates sample moments $\bar{g}\left(\beta\right)=\frac{1}{n}\sum_{i=1}^{n}g\left(X_i,\beta\right)=0$. Let the vector of parameters be $\beta=\left[\kappa\ \tau_{ATE}\ \tau_{TOT}\right]$ and the moment vectors be $\bar{g}\left(\beta\right)=\left[\bar{g}_1,\bar{g}_2^{TOT},\bar{g}_2^{ATE}\right]$. Newey (1984) noticed that if $\bar{g}\left(\beta\right)$ has a recursive structure, we can solve for $\hat{\beta}$ sequentially.[11] First, partition the estimating equation

$$\bar{g}\left(X_i,\beta\right)=\left[\begin{array}{c}\bar{g}_1\left(X_i,\kappa\right)\\ \bar{g}_2^{TOT}\left(X_i,\kappa,\tau_{TOT}\right)\\ \bar{g}_2^{ATE}\left(X_i,\kappa,\tau_{ATE}\right)\end{array}\right]. \tag{4.6}$$

Notice that $\dim\left(\beta\right)=\dim\left(\bar{g}\left(\beta\right)\right)=K+4L$ so the model is exactly identified and we can find $\hat{\beta}$ that sets $\bar{g}\left(\beta\right)$ exactly equal to zero. We first solve $\bar{g}_1\left(X_i,\kappa\right)=0$ to get an estimator $\hat{\kappa}$. In the second step, we plug in $\hat{\kappa}$ and $\hat{p}\left(X_i\right)$ in $\bar{g}_2^t\left(X_i,\kappa,\theta_t\right)=0$ for $t=ATE,\ TOT$ to obtain estimates of the ATE and of the TOT.

The conditional independence assumption, strict overlap and the assumption of a well specified propensity score model imply that $E\left[g\left(X_i,\beta_0\right)\right]=0$. We need to make three additional assumptions to derive the stated distribution of the method of moments estimator. First, assume that

$$g\left(X_i,\beta\right)=g\left(X_i,\beta'\right)\text{ iff }\beta=\beta'.$$

Second, the $(K+4L)\times(K+4L)$ matrix $G_0=E\left[\left.\frac{\partial g(X_i,\beta)}{\partial\beta'}\right|_{\beta_0}\right]$ exists, is finite and has rank $(K+4L)$. Finally, assume that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}g\left(X_i,\beta\right)\right)\xrightarrow{d}N\left(0,S\right),$$

where $S=E\left[g\left(X_i,\beta_0\right)g\left(X_i,\beta_0\right)'\right]$ is finite. A necessary condition for these last two assumptions to hold is that $p\left(X_i\right)$ is never equal to zero or to one (which is guaranteed by strict overlap). Then under these regularity conditions it can be shown that

$$\sqrt{N}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N\left(0,G_0'^{-1}SG_0^{-1}\right).$$

It is important to recognize that estimation error in the first stage carries over to the second stage in a very specific way. Because $\bar{g}_1\left(X_i,\kappa\right)$ only depends on κ but not on τ_t,

[11] A similar result was obtained by Murphy and Topel (1985) and Pagan (1986).

the matrix G is lower triangular

$$G = \begin{bmatrix} G_{\kappa}^{1} & 0 & 0 \\ G_{\kappa}^{2,TOT} & G_{\tau}^{2,TOT} & 0 \\ G_{\kappa}^{2,ATE} & 0 & G_{\tau}^{2,ATE} \end{bmatrix} \text{ with } G_{h}^{j,t} = E\left[\frac{\partial g_{j}^{t}\left(X_{i},\beta\right)}{\partial h}\right] \text{ , } h = \left\{\kappa, \tau^{TOT}, \tau^{ATE}\right\}.$$

The elements $G_{\kappa}^{2,t}$ and $G_{\kappa}^{2,t}$ capture the effect of the first stage estimation on the variance of the parameters estimated in the second step. In Appendix II, I present the expressions for $G_{h}^{j,t}$ for the cases in which the selection equation error follows a normal, logistic, and a Cauchy distribution and for a linear probability model.

This sequential method of moments (SQMM) variance estimator allows us to test several null hypotheses of interest. Obviously, we can test for no average treatment effect on a given outcome Y, $H_0 : \theta_{ATE}\left(Y\right) = 0$ and a zero average treatment effect among those treated $H_0 : \theta_{TOT}\left(Y\right) = 0$. More interestingly, we can use it to test hypotheses that are somewhat difficult to test using other variance estimators (e.g. bootstrap). First, we can perform a test of homogeneity of the treatment effect for outcome Y, that is $H_0 : \theta_{TOT}\left(Y\right) = \theta_{ATE}\left(Y\right)$. Second, we can implement a test of equality of treatment effects across outcomes; that is $H_0 : \theta_{TOT}\left(Y\right) = \theta_{TOT}\left(Y'\right)$ or $H_0 : \theta_{ATE}\left(Y\right) = \theta_{ATE}\left(Y'\right)$. Finally, we can also test for the complete absence of treatment effects by testing that the treatment effect is zero across all outcomes $H_0 : \theta_{TOT}\left(Y_1\right) = \cdots = \theta_{TOT}\left(Y_L\right) = 0$.[12]

IV Monte Carlo Simulation Results

A Data Generating Process

I study the finite sample size and power properties in data generating processes (DGPs) that are all special cases of the latent index model

$$\begin{aligned} T_i^* &= \kappa_0 + \kappa_1 X_{1i} + \kappa_2 X_{2i} + \kappa_3 X_{3i} + u_i, \\ T_i &= 1\left(T_i^* > 0\right), \\ Y_i &= \theta T_i + m(p(X_i)) + \varepsilon_i, \end{aligned}$$

where u_i and ε_i are independent of X_{ji} and of each other, $m(\cdot)$ is a curve to be discussed below, and $p(X_i)$ is the propensity score implied by the model. Covariates X_{ji} are taken to be independently distributed standard normal. My focus is on cross-sectional contexts, so ε_i is independent across i, but potentially heteroscedastic. In particular, let e_i be an

[12] The code to implement the SQMM variance estimator in Stata using different choices of the parametric first stage is available in `http://sitemaker.umich.edu/matiasb/`

independent and identically distributed standard normal sequence and then define

$$\varepsilon_i = \psi\left(e_i p(X_i) + e_i T_i\right) + (1-\psi)e_i. \tag{4.7}$$

In these DGPs the ATE and the TOT both equal θ. Heteroscedasticity is controlled by the parameter ψ in equation (4.7). When $\psi = 0$, we obtain homoscedasticity. When $\psi \neq 0$, the residual variance depends on treatment as well as on the propensity score. The function $m(\cdot)$ manipulates the non-linearity of the outcome equation that is common to both treated and non-treated observations. The precise equations used for these two regression functions are summarized below:[13]

Curve	**Formula**	**Description**
1	$m_1(q) = 0.15 + 0.7q$	Linear
2	$m_2(q) = 0.2 + \sqrt{1-q} - 0.6(0.9-q)^2$	Nonlinear

As discussed in Busso et al. (2008) the choice of distribution for u_i is relevant to both the finite and large sample performance of average treatment effect estimators. In particular, they show that for these DPGs not to violate the strict overlap assumption, it is required that the tails of the distribution of u_i are fatter than the tails of the distribution of the index κX_i. I consider three different distributional assumptions for the treatment assignment equation error, u_i: Cauchy, standard normal and logistic.[14]

The parameters κ_j for $j = 0, 1, 2, 3$ manipulate the different degrees of overlap between the conditional densities of $p\left(X_i\right)$, namely $f_{p(X)|T=1}\left(q\right)$ and $f_{p(X)|T=0}\left(q\right)$. The larger κ_1, κ_2 and κ_3, the closer we are to violating strict overlap. The index κX_i is normal with mean κ_0 and variance $\kappa_1^2 + \kappa_2^2 + \kappa_3^2$. If u_i follows a Cauchy distribution, strict overlap is guaranteed to be satisfied for any value of κ_j since there is no normal distribution whose tails are fatter than those of the Cauchy. If u_i is logistic or normal, large values of κ_j will produce a violation of strict overlap. The values of κ_j also manipulate the different ratios of treated to control units n_1/n_0. For any given κ_j with $j = 1, 2, 3$, if $\kappa_0 = 0$ then the ratio $n_1/n_0 = 1$, if $\kappa_0 > 0$ then the ratio $n_1/n_0 > 1$ and if $\kappa_0 < 0$ the ratio $n_1/n_0 < 1$. For any given κ_0 the larger κ_j (with $j = 1, 2, 3$) the larger will be the maximum propensity score observed in the sample.

[13] The two regression functions we consider, $m(\cdot)$, correspond to the functional forms used by Frölich (2004) and by Busso et al. (2008).

[14] The baseline case (fully described below) assumes u_i follows a Cauchy distribution. In principle, I could have let u_i follow a normal distribution with parameters selected in a manner that they allow for good overlap. In such a case, because in the normal case the parameters that manipulate overlap also change the ratio of treated to control observations in the designs, I would not be able to explore designs as the ones we can study when u_i is distributed Cauchy. However, it should be noted that when assuming a normal or a logistic distribution with different ratios of treated to control units results were basically the same as in the baseline case.

To study the finite sample size and power properties of the estimators, I select a baseline DGP and then change the parameters one at a time in order to assess the robustness of the results. The baseline DGP assumes a homoscedastic outcome error ($\psi = 0$), no treatment effect ($\theta = 0$), a ratio $n_1/n_0 > 1$ ($\kappa_j = 1$ for $j = 0, 1, 2, 3$); good overlap (u_i follows a Cauchy distribution), and a linear outcome equation ($m_1(q)$). [15]

B Competing Inference Procedures

I study the finite sample performance of the SQMM variance estimator when used in a two-sided $t-$test to test the null hypothesis that $H_0 : \theta_t = \theta_t^0$ against the alternative that $H_1 : \theta_t \neq \theta_t^0$, for t=ATE, TOT. I compare it to five alternative inference methods: (i) a percentile-t bootstrap that uses the SQMM variance, (ii) a percentile bootstrap that does not utilize any variance estimator, (iii) a t-test that uses a homoscedastic ordinary least squares variance, (iv) a $t-$test that relies on the heteroscedastic robust (Eicker-White) variance and (v) a t-test that uses the so called HC$_3$ variance estimator. Strategies (i) and (ii) acknowledge that the weights involved in the estimation of the TOT have sampling variability. The last three methods will help us to assess the cost, in terms of size distortion, of ignoring that the weights have been estimated.

I use the paired bootstrap to test the null that $H_0 : \theta_t = \theta_0$ against the alternative that $H_1 : \theta_t \neq \theta_0$ for θ_t for $t = ATE, TOT$. In order to fix the ratio of treated to control observations, I resample with replacement $\{Y_i, X_i\}$ separately from $T_i = 1$ and from $T_i = 0$.[16] In each bootstrap repetition the propensity score model is re-estimated, then it is used to compute a new set of weights and to get estimates $\hat{\theta}_t^*$, as well as an estimate of the SQMM variance for both treatment effect estimators $\hat{V}_t^*$. First consider a percentile-t method that should provide an asymptotic refinement. Consider the test statistic $t = \left|\hat{\theta}_t - \theta_0\right| / \sqrt{\hat{V}_t}$. I perform B bootstrap replications that produce B test statistics $t_1^*, ..., t_B^*$ where $t_b^* = \left|\hat{\theta}_t^* - \hat{\theta}_t\right| / \sqrt{\hat{V}_t^*}$. Then I compute a bootstrap $p-$value as $\frac{1}{B}\sum_{b=1}^{B} 1\left(t_b^* > t\right)$. If we wanted to avoid altogether the calculation of the SQMM covariance matrix, then the only valid bootstrap method would be one that does not use any estimate of the variance of the treatment effect at all. One possibility is to use the percentile bootstrap: for a level

[15] Busso et al. (2008) study the finite sample properties of IPW estimators in DGPs similar to the ones assumed here. They compare IPW estimators to other available semiparametric estimators such as kernel, local-linear, ridge, nearest-neighbor, and pair matching, double robust estimators, among others. They find that IPW is approximately unbiased and semiparametrically efficient.

[16] It is worth noting that one potential problem of using the bootstrap in our context is that in any given draw we could get all (or most of) the observations with $T = 1$ or all observations with $T = 0$. In such a situation the bootstrap draw will fail because we would be unable to compute the counterfactual mean for one group. This is avoided if we resample separately from the treated and untreated groups.

In our baseline specification, which assumes a Cauchy error term for the selection equation and a relatively large number of observations in both groups, resampling from $\{Y_i, X_i, T_i\}$ does not cause any bootstrap replication to fail. The results it yields are the very similar to the ones shown in Table 4.1.

α, we find the lower $\alpha/2$ and the upper $\alpha/2$ quantiles of the bootstrap estimates $\hat{\theta}^*_{t1}, ..., \hat{\theta}^*_{tB}$ and reject the null if θ_0 falls outside that region.[17]

Recall that the second step is a weighted regression of the outcome of interest Y_i on the treatment indicator T_i and a constant term. An alternative strategy to produce inference would be to proceed as if the weights involved in that second step have no sampling variability. The homoscedastic OLS variance would be inconsistent, even if the weights were known, because by weighting the data according to (4.4) and (4.5) we are, by construction, introducing heteroscedasticity. A standard Eicker-White covariance matrix would be more appropriate. MacKinnon and White (1985) show that the standard Eicker-White covariance matrix tends to perform badly in small samples. Since we are interested in the size and power properties in finite samples I also compare the performance of the HC_3 covariance matrix that estimates the i-th element of the variance of the error term in the outcome equation by $\hat{e}_i^2 / \left(1 - \hat{h}_i\right)^2$ where $\hat{e}_i$ is the WLS residual and $\hat{h}_i$ is the leverage of observation i.[18]

The number of Monte Carlo replications R is 100,000 for the t-tests based on the SQMM, OLS, Eicker-White and HC_3 covariance matrices. Due to computational constraints, the number of replications for the bootstrap-based procedures is 4,000 and the number of bootstrap replications is 1,000. I am interested in studying the size of the test under alternative testing strategies. Let the true size of a given test be s. Each Monte Carlo replication will generate a test statistic that either exceeds or does not exceed the nominal critical value $t_{\alpha/2}$. These can be thought as Bernoulli trials so, using the normal approximation to the Binomial, a 95% confidence interval must cover 2 × 1.96 standard errors or $3.92\sqrt{s\left(1-s\right)/R}$. If the size of the test is 0.05 and $R = 100,000$, the length of the 95% confidence interval of the true size is 0.0028; if $R = 4,000$ the length of the confidence interval is 0.0135.

[17] Notice that the bootstrap requires homoscedasticity. However, to estimate the ATE and the TOT we are reweighting the data and therefore introducing heteroscedasticity. Thus, it is possible that the bootstrap has bad size properties in this context.

[18] The Eicker-White heteroscedastic robust variance estimator (HC) utilizes an estimator of the variance of the error term Ω, that is given by $\hat{e}_i^2$ in the diagonal and zeroes in the off-diagonal elements. The HC estimator is consistent both under homoscedasticity and heteroscedasticity. However, this variance estimator might not perform well in small samples. Recall that the squared of the OLS residual $\hat{e}_i^2$ is not un unbiased estimate of the squared error term e_i^2. To see this, consider a linear model with constant variance σ^2. It is easy to show that $E\left[\hat{e}_i^2\right] = \sigma^2(1 - \hat{h}_i) \leq \sigma^2$, where $0 \leq \hat{h}_i \leq 1$ is the i-th diagonal element of $P_X = X(X'X)^{-1}X'$. Thus, in the linear case, we can divide $\hat{e}_i^2$ by $1 - \hat{h}_i$ which would yield an unbiased estimate of σ^2 if the error terms were actually homoscedastic. This motivates the HC_2 variance estimator that uses $\hat{e}_i^2/(1 - \hat{h}_i)$ as an estimator of the diagonal elements of Ω. MacKinnon and White (1985) propose to use the Jacknife. They show that the resulting i-th diagonal element of $\hat{\Omega}$ is very well approximated by $\hat{e}_i^2/(1 - \hat{h}_i)^2$. Intuitively, when the error term is heteroscedastic, observations with large variances will tend to influence the parameter estimates more than observations with small variances and will therefore tend to have residuals that are too small.

C Finite Sample Size and Power

Table 4.1 compares the size properties of a t-test of the null of $\theta_{TOT} = 0$ against the alternative that $\theta_{TOT} \neq 0$.[19] The rows present the simulated empirical size for the baseline DGP, with samples sizes 50, 100 and 500, evaluating different nominal test sizes. The t-test based on the SQMM variance estimator achieves a size that is close to the nominal size even for samples with 50 and 100 observations. This test tends to slightly under-reject the true null when the sample size is 500. The percentile-t bootstrap methods based on the SQMM variance tends to also perform well, slightly overrejecting in smaller samples. The percentile bootstrap rejects the true null too often in small samples. For a nominal size of 0.05, the empirical size is 0.128 for a sample with 50 observations, and it is 0.075 for a sample of 100 units.

Ignoring that the weights are estimated lead to very poor inference. As expected the t-test that uses the homoscedastic least squares variance estimator has a large size distortion that does not disappear as the sample size increases. This is partly caused by the fact that the weights, even in a situation in which they are known, introduce heteroscedasticity rendering the homoscedastic LS variance estimator inconsistent. A t-test based on the Eicker-White variance estimator does not perform well in small samples either: for a sample with 100 observations and a nominal size of 0.05, the size of such a t-test is 0.095. This size distortion is not caused by the known bad finite-sample performance of the Eicker-White covariance matrix since the the t-test that uses the HC_3 variance estimator has a similar size distortion. This problem however, tends to disappear as the sample size increases and the precision of the estimates of the propensity score (and thus the weights) improves. Indeed, note that with a sample size of 500 both the test based on Eicker-White and the one based on the HC_3 estimator tend to only slightly over-reject the true null. Hirano et al. (2003) show that using the known propensity score instead of the estimated one leads to conservative inference. The simulation results suggest that, in finite samples, a variance estimator that ignores the sampling variability of the weights might in fact be too small.

Figures 4.1-4.4 explore the effect that different features of the DGP have on the size of the tests involving θ_{TOT}.[20] The figures plot the difference between the empirical size and the target nominal size, for different values of the nominal size of the t-test. They all compare a test that uses the SQMM variance estimator with one that uses the HC_3 variance estimator. In Figure 4.1 we can observe that, as expected, when the sample size increases from 40 to 500 the empirical size of the tests tend to the nominal size. The size distortion when using the SQMM estimator tend to be really small when compared to the HC_3 estimator, specially in small samples with $n < 500$.

[19] Results for the ATE are similar and can be found in the appendix tables.

[20] The results for the θ_{ATE} are very similar. Figures are available in an appendix not included in the paper that can be found in `http://sitemaker.umich.edu/matiasb/`

Figure 4.2 shows the effect of changing κ_0, while keeping the value of $\kappa_1 = \kappa_2 = \kappa_3 = 1$ as in the baseline DGP. Instead of plotting different values of κ_0 on the x-axis, I plot values of the treated-to-control number of observations, the ratio n_1/n_0, which has a one-to-one mapping to κ_0 (for a given $\kappa_j, j = 1, 2, 3$) and provides a more intuitive interpretation. As the ratio n_1/n_0 increases it becomes harder to estimate the TOT because there are fewer controls to build counterfactuals for the treated units. The top panel of Figure 4.2 plots a DGP with homoscedastic errors in the outcome equation. The test that uses the SQMM variance estimate tends to underreject when the ratio n_1/n_0 is too small and it overrejects when the ratio is too large. The same pattern is observed when using the HC_3 variance but for every value of n_1/n_0 the size distortion is larger in the latter. If the outcome error is heteroscedastic ($\psi = 2$) then the size distortion of the test with SQMM variance disappears whereas the test using the HC_3 variance tends to always overreject. The reason for the better performance of the SQMM-based test in a situation with heteroscedastic errors, is that the proposed SQMM variance estimator is agnostic regarding the heteroscedasticity of the error term and it is therefore expected to perform better (when the sample size is small) in a context in which the DGP is heteroscedastic.

Figure 4.3 studies the effect of changing κ_j for $j = 1, 2, 3$, while keeping $\kappa_0 = 1$ as well as all the other features of the baseline DGP constant. Again, in the interest of providing a clearer interpretation of the results, instead of plotting values of κ_j on the x-axis, I plot values of the maximum value of the propensity score among the controls. This is an important quantity involved in the computation of the weights because as the $\max\{p(X_i)|\, T_i = 0\}$ approaches one, we are closer to violating the strict overlap assumption and therefore to violating the assumptions for convergence in distribution of the SQMM estimator of the ATE and TOT. Note, however, that because u is Cauchy it is always the case that strict overlap is satisfied. The top panel of the figure plots the case of homoscedastic outcome errors. The t-test tends to reject slightly more (relatively to the baseline case) as the $\max\{p(X_i)|\, T_i = 0\}$ approaches one. In a DGP with heteroscedastic errors in the outcome equation, the t-test based on the SQMM variance tends to have good size properties.

Figure 4.4 presents the effect of violating the strict overlap assumption. I now assume that u is standard normal, $\kappa_0 = 1$ and the outcome error is homoscedastic. Recall that the variance of the index κX_i is $\kappa_1^2 + \kappa_2^2 + \kappa_3^2$ and let $\kappa_j = \nu$ for $j = 1, 2, 3$. Since X_i are standard normal then κX_i is normal with mean zero and variance $3\nu^2$. So the tails of the index will be fatter than tails of u when $\nu > \sqrt{1/3} \approx 0.577$. Busso et al. (2008) point out that the breakdown of the asymptotics happens continuously even before strict overlap is actually violated. The figure shows that when κ_j is low, the empirical size of a test that uses the SQMM variance is close to the nominal size. As κ increases the test tends to overreject the true null of no treatment effect. This pattern is the opposite of the expected one. As κ increases, a larger number of observations will have a propensity score close to one. This should increase the variance of the treatment effect leading to inability to reject the null. However, as noted by Busso et al. (2008), in contexts with poor overlap the point estimate of the treatment effect will be severely biased which could more than compensate for the increase in the variance, thereby leading to rejection of the null.[21]

I turn now to the computation of a power function of the t-test that uses the SQMM variance estimator. I compare it to the simulated power obtained using a bootstrap percentile-t method, the only other inference procedure that presented relatively good size properties. Due to computational constraints the power function for the percentile-t bootstrap was computed in a sparser grid for θ_{TOT} and using 2,000 Monte Carlo repetitions whereas the power function of the t-test based on the SQMM variance was calculated using 10,000 repetitions. Figure 4.5 plots the power functions for different values of the true θ_{TOT} in the baseline DGP. Both power functions are similar for values of θ_{TOT} less than one. When the true $\theta_{TOT} = 1$ the power of both methods is approximately 0.81. For values greater than one the SQMM t-test has more power specially at lower nominal sizes. The fact that both size and power are so similar between these two procedures is an indication that the bootstrap percentile-t method is not providing any refinement to the asymptotic variance, which indicates that the SQMM variance estimator is providing a good approximation to the true variance of $\hat{\theta}_{TOT}$.[22]

As I mentioned in section III, we can also use the SQMM variance estimator to test hypotheses that are difficult to test using other variance estimators. Some examples are: a test of homogeneity of the treatment effect for outcome y, that is $H_0 : \theta_{TOT}(Y) = \theta_{ATE}(Y)$; a test of equality of treatment effect across outcomes, that is $H_0 : \theta_{TOT}(Y) = \theta_{TOT}(Y')$ or $H_0 : \theta_{ATE}(Y) = \theta_{ATE}(Y')$ for $Y \neq Y'$; and a test for the complete absence of any treatment effect by testing the joint null that the treatment effect is zero across all outcomes $H_0 : \theta_{TOT}(Y_1) = \cdots = \theta_{TOT}(Y_L) = 0$. All these tests require the computation of the covariance between treatment effect estimators for different outcomes and/or estimands, which is easily done using sequential method of moments. Table 4.2 presents the size and power of a Wald test of the null $H_0 : \theta_{TOT}(Y_1) = \theta_{TOT}(Y_2) = 0$ against the alternative that the treatment has some effect on at least one outcome. I use the baseline DGP where, as before, Y_1 is a function of $m_1(p(X_i))$ and I let Y_2 be a function of $m_2(p(X_i))$.[23] The size of the test is very close to the nominal size although the test tends to over-reject for

[21] I find that if the error tern u in the selection equation follows a logistic or a normal distribution with good overlap, the results in terms of size do not differ substantially from the ones found using the baseline DGP (that assumes u follows a Cauchy distribution).

[22] In results not shown I tested whether the distribution of $\hat{\theta}_{TOT}$ is normal. I cannot reject symmetry but I can reject zero excess kurtosis. In a qq-plot it can be observed that the the distribution of $\hat{\theta}_{TOT}$ differs from the normal distribution at the far ends of the tails.

[23] Results for the ATE are presented in appendix Table 4.2. In the case of the ATE the test tends to under-reject the null of no treatment when the null is true. The power of the test is similar to the power observed for nulls involving the TOT.

low nominal sizes. The test has power against the alternative that there is a treatment effect on both outcomes. In the case of the true alternative that $\theta_{TOT}(Y_1) = \theta_{TOT}(Y_2) = 1$ the test rejects the null of no treatment 0.785 percent of the time.

V Conclusion

In this paper I propose a simple sequential method of moments variance estimator of inverse probability weighting estimators of average treatment effects. The proposed estimator has several advantages over competing inference strategies. First, it is easy to compute (i.e. faster than the bootstrap). Second, it allows to simply test hypotheses involving different outcomes and estimands because it directly provides an expression for the covariance of the treatment effects across estimands and outcomes. Third, Monte Carlo simulations suggest that the size and power of tests based on the SQMM variance perform well against other competing inference strategy. If anything inference based on the SQMM variance estimator tend to be slightly conservative.

I assess the finite samples size and power of t-tests of no treatment effect using the SQMM variance estimator and find that it performs well even in samples of size 100. The empirical size is close to the nominal size. This is a finding that is robust in settings with homoscedastic and with heteroscedastic errors in the outcome equation. It is also robust to different samples sizes. In contexts in which the ratio of treated to control units is too high t-tests that use the SQMM variance tend to slightly overreject. The same pattern was observed in situations in which the observed maximum estimated propensity score was close to being one and in settings in which the strict overlap assumption, necessary for identification, was violated. In all these DGPs, using the SQMM variance yields better size than alternative estimators that ignore the fact that the weights were estimated.

I also compare inference using SQMM to inference procedures using the bootstrap. The percentile bootstrap tends to reject too often, specially in very small samples. The percentile-t bootstrap implemented using the SQMM variance estimator has similar size and power to the tests using the asymptotic SQMM variance. This is an indication that the bootstrap percentile-t does not provide any refinement to the asymptotic variance, which indicates that the SQMM variance estimator is a good enough approximation of the true variance of the IPW treatment effect estimators.

Table 4.1: Finite sample test size (TOT)

t-test of H_0*: TOT=0*

Sample Size	Nominal Size	Sequential MM	Percentile-t Bootstrap	Percentile Bootstrap	OLS homoscedastic	Eicker-White	HC_3
		[1]	[2]	[3]	[4]	[5]	[6]
50	0.01	0.017	0.017	0.063	0.128	0.078	0.072
	0.05	0.055	0.062	0.128	0.237	0.158	0.147
	0.10	0.102	0.131	0.187	0.317	0.226	0.212
	0.15	0.148	0.199	0.243	0.379	0.284	0.269
	0.20	0.197	0.258	0.285	0.432	0.336	0.322
100	0.01	0.013	0.013	0.021	0.102	0.033	0.031
	0.05	0.049	0.065	0.075	0.209	0.095	0.091
	0.10	0.095	0.129	0.136	0.288	0.155	0.149
	0.15	0.143	0.188	0.184	0.352	0.211	0.204
	0.20	0.192	0.242	0.237	0.407	0.264	0.256
500	0.01	0.009	0.017	0.020	0.083	0.013	0.013
	0.05	0.044	0.058	0.060	0.185	0.058	0.057
	0.10	0.091	0.104	0.105	0.267	0.110	0.109
	0.15	0.138	0.166	0.164	0.330	0.160	0.158
	0.20	0.186	0.229	0.225	0.386	0.209	0.207

Note: The table shows test size in Monte Carlo experiments for the baseline DGP (see section V.A). Results for colums 1,4,5,6 are based on 100,000 replications. Results for columns 2 and 3 are based on 4,000 replications and each replication uses 1,000 bootstrap repetitions.

Table 4.2: Finite sample test size and power of joint test of no TOT

Wald test of $H_0: TOT_1 = TOT_2 = 0$

		Nominal size		
	True θ	0.01	0.05	0.1
Size	**0.0**	0.020	0.052	0.084
Power	0.1	0.024	0.060	0.094
	0.2	0.037	0.087	0.132
	0.3	0.062	0.135	0.192
	0.4	0.103	0.202	0.276
	0.5	0.162	0.291	0.379
	0.6	0.237	0.393	0.488
	0.7	0.328	0.500	0.598
	0.8	0.428	0.608	0.697
	0.9	0.532	0.703	0.782
	1.0	0.630	0.785	0.849
	1.2	0.790	0.895	0.934
	1.4	0.892	0.954	0.974

Note: The table shows test size and power in Monte Carlo experiments for the baseline DGP (see section V.A) based on 100,000 replications.

Table 4.A1: Finite sample test size (ATE)

t-test of $H_0: ATE=0$

Sample Size	Nominal Size	Sequential MM	Percentile-t Bootstrap	Percentile Bootstrap	OLS homoscedastic	Eicker-White	HC_3
		[1]	[2]	[3]	[4]	[5]	[6]
50	0.01	0.012	0.012	0.030	0.080	0.042	0.037
	0.05	0.043	0.063	0.083	0.172	0.109	0.100
	0.10	0.082	0.123	0.147	0.249	0.174	0.162
	0.15	0.124	0.180	0.203	0.309	0.230	0.217
	0.20	0.167	0.242	0.255	0.364	0.283	0.269
100	0.01	0.009	0.013	0.014	0.057	0.020	0.018
	0.05	0.039	0.066	0.063	0.143	0.072	0.068
	0.10	0.080	0.134	0.117	0.218	0.127	0.122
	0.15	0.124	0.186	0.169	0.279	0.182	0.175
	0.20	0.170	0.246	0.224	0.335	0.235	0.228
500	0.01	0.007	0.018	0.018	0.044	0.011	0.011
	0.05	0.039	0.062	0.063	0.124	0.051	0.051
	0.10	0.081	0.118	0.112	0.197	0.101	0.100
	0.15	0.127	0.185	0.176	0.259	0.150	0.149
	0.20	0.174	0.234	0.226	0.315	0.200	0.199

Note: The table shows test sizes in Monte Carlo experiments for the baseline DGP (see section V.A). Results for colums 1,4,5,6 are based on 100,000 replications. Results for columns 2 and 3 are based on 4,000 replications and each replication uses 1,000 bootstrap repetitions.

Table 4.A2: Finite sample test size of joint test of no ATE

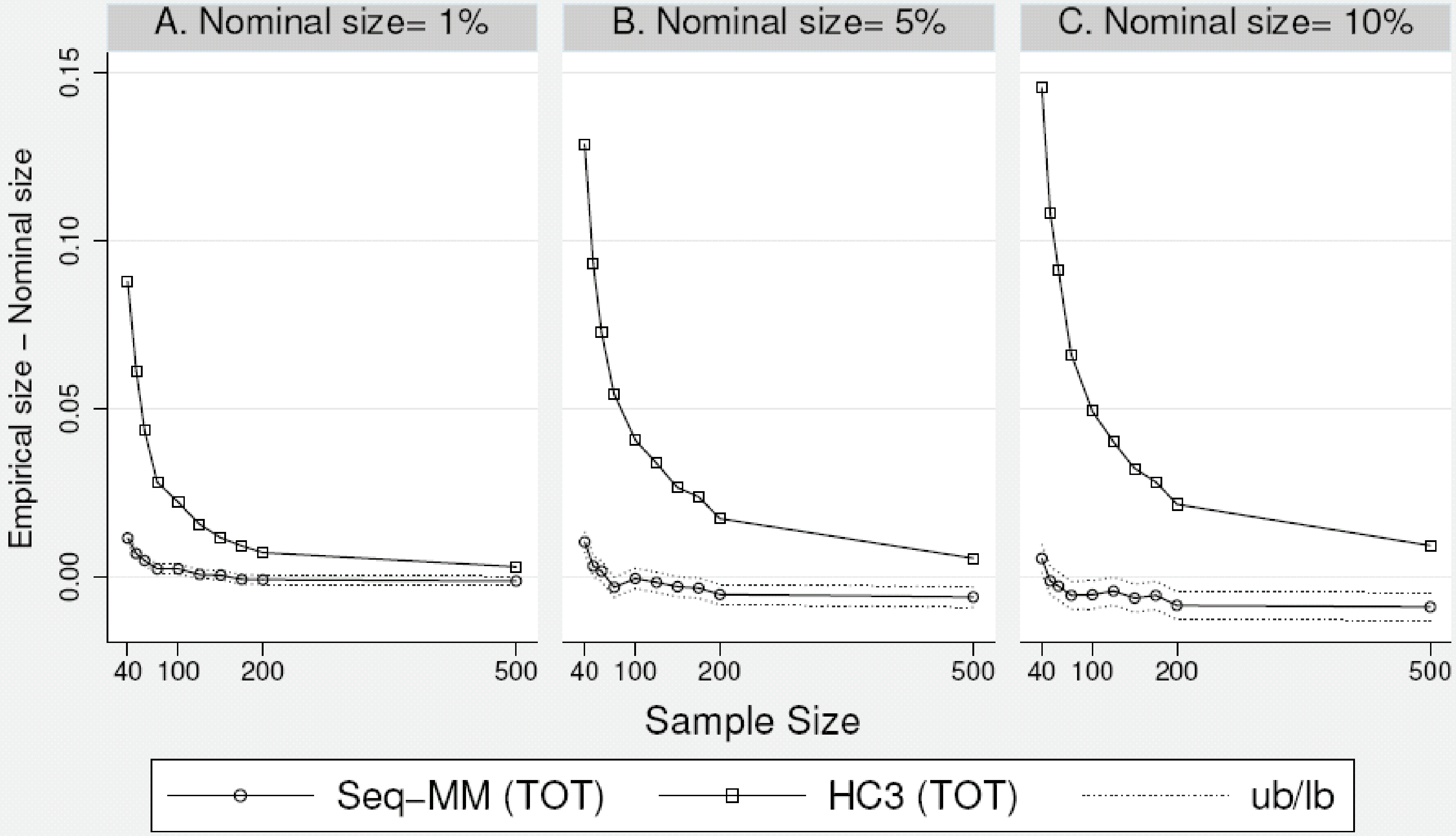
Wald test of $H_0: ATE_1 = ATE_2 = 0$

		Nominal size		
	True θ	0.01	0.05	0.1
Size	**0.0**	0.009	0.031	0.055
Power	0.1	0.013	0.040	0.069
	0.2	0.026	0.070	0.112
	0.3	0.053	0.125	0.187
	0.4	0.097	0.208	0.290
	0.5	0.167	0.317	0.416
	0.6	0.261	0.442	0.548
	0.7	0.377	0.572	0.675
	0.8	0.499	0.693	0.777
	0.9	0.620	0.789	0.857
	1.0	0.728	0.864	0.914
	1.2	0.877	0.951	0.973
	1.4	0.952	0.984	0.992

Note: The table shows test size and power in Monte Carlo experiments for the baseline DGP (see section V.A) based on 100,000 replications.

Figure 4.1: Size of test in DGPs with different sample sizes

Finite sample deviations from nominal size of t−test, H0: TOT=0

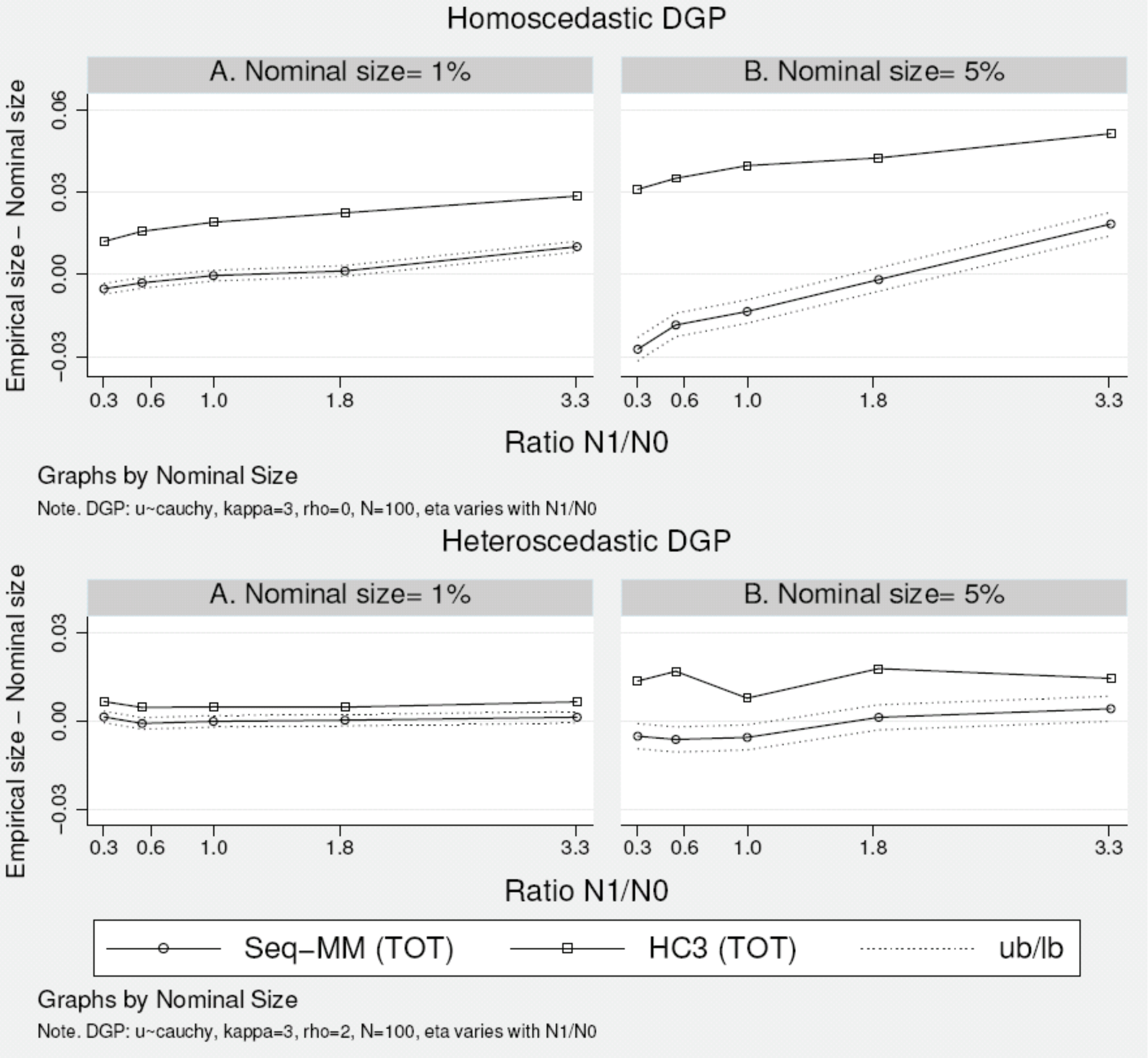


Graphs by Nominal Size

Note. DGP: u~cauchy, kappa=3, eta=1, rho=0, N varies

Figure 4.2: Size of test in DGPs with different N1/N0 ratios

Finite sample deviations from nominal size of t−test, H0: TOT=0

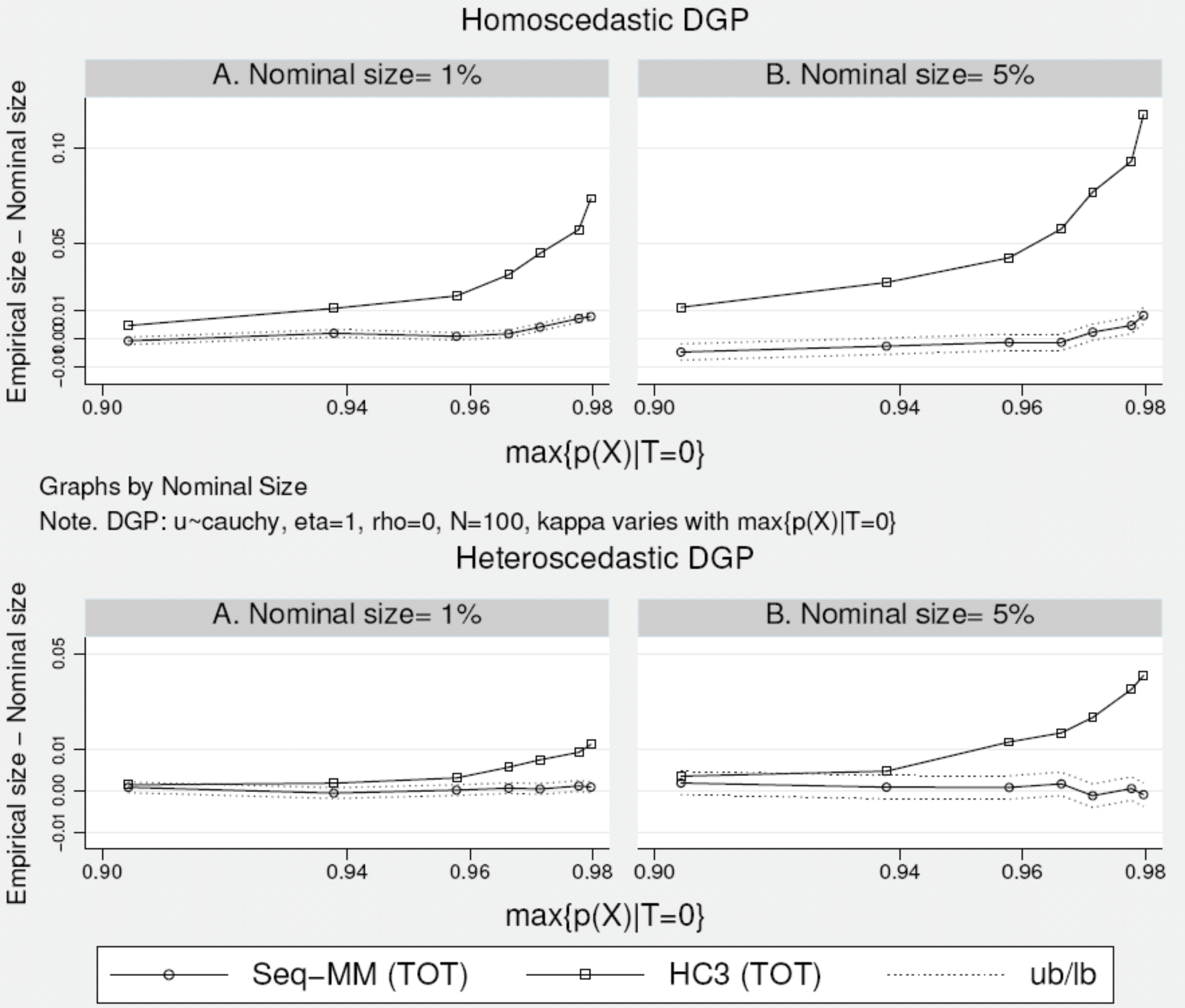


Note. DGP: u~cauchy, kappa=3, rho=0, N=100, eta varies with N1/N0

Note. DGP: u~cauchy, kappa=3, rho=2, N=100, eta varies with N1/N0

Figure 4.3: Size of test in DGPs with different max{p(X)|T=0}

Finite sample deviations from nominal size of t–test, H0: TOT=0



Graphs by Nominal Size

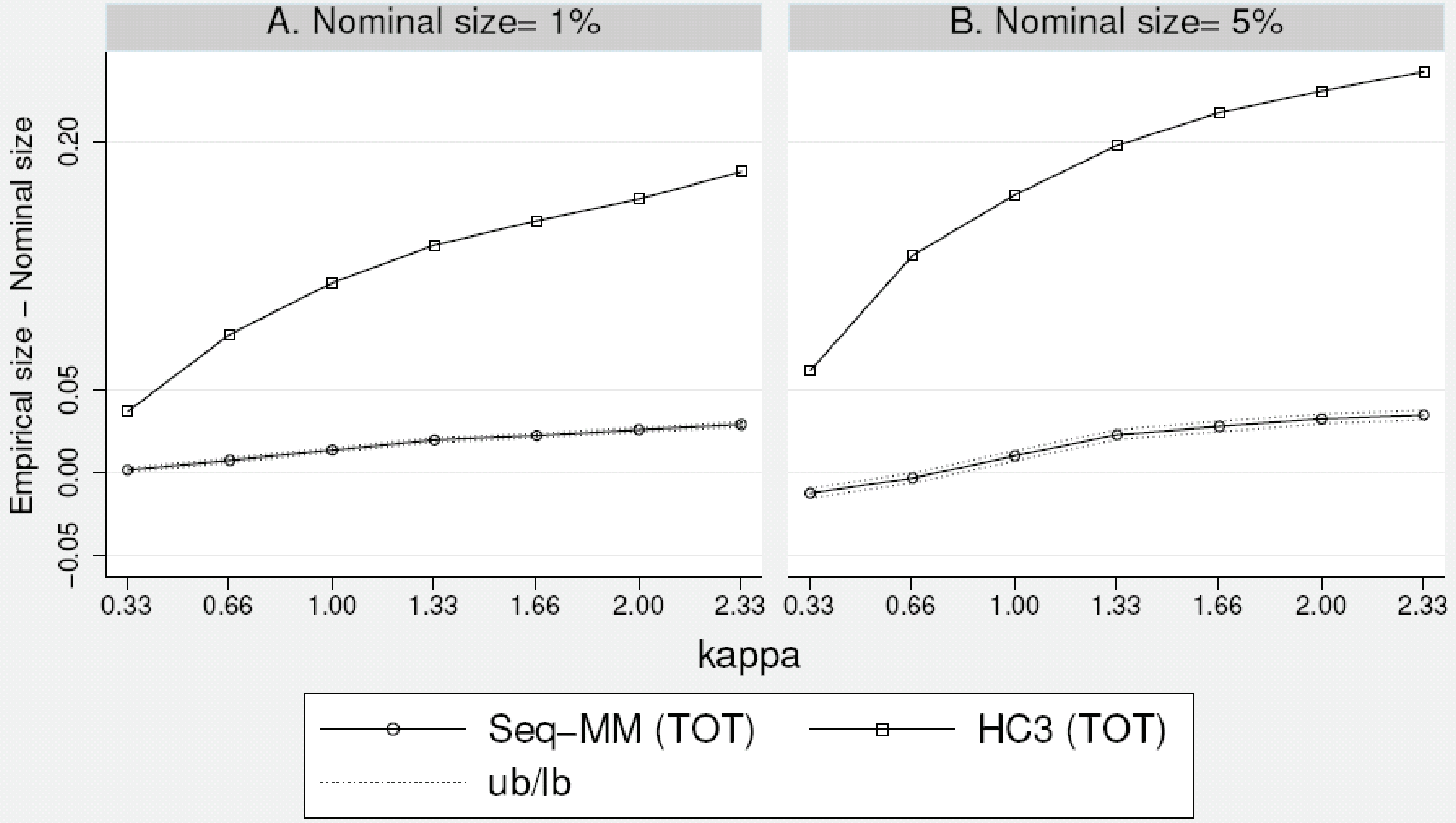
Note. DGP: u~cauchy, eta=1, rho=0, N=100, kappa varies with max{p(X)|T=0}

Graphs by Nominal Size

Note. DGP: u~cauchy, eta=1, rho=2, N=100, kappa varies with max{p(X)|T=0}

Figure 4.4: Size of test in DGPs with different degrees of violation of overlap

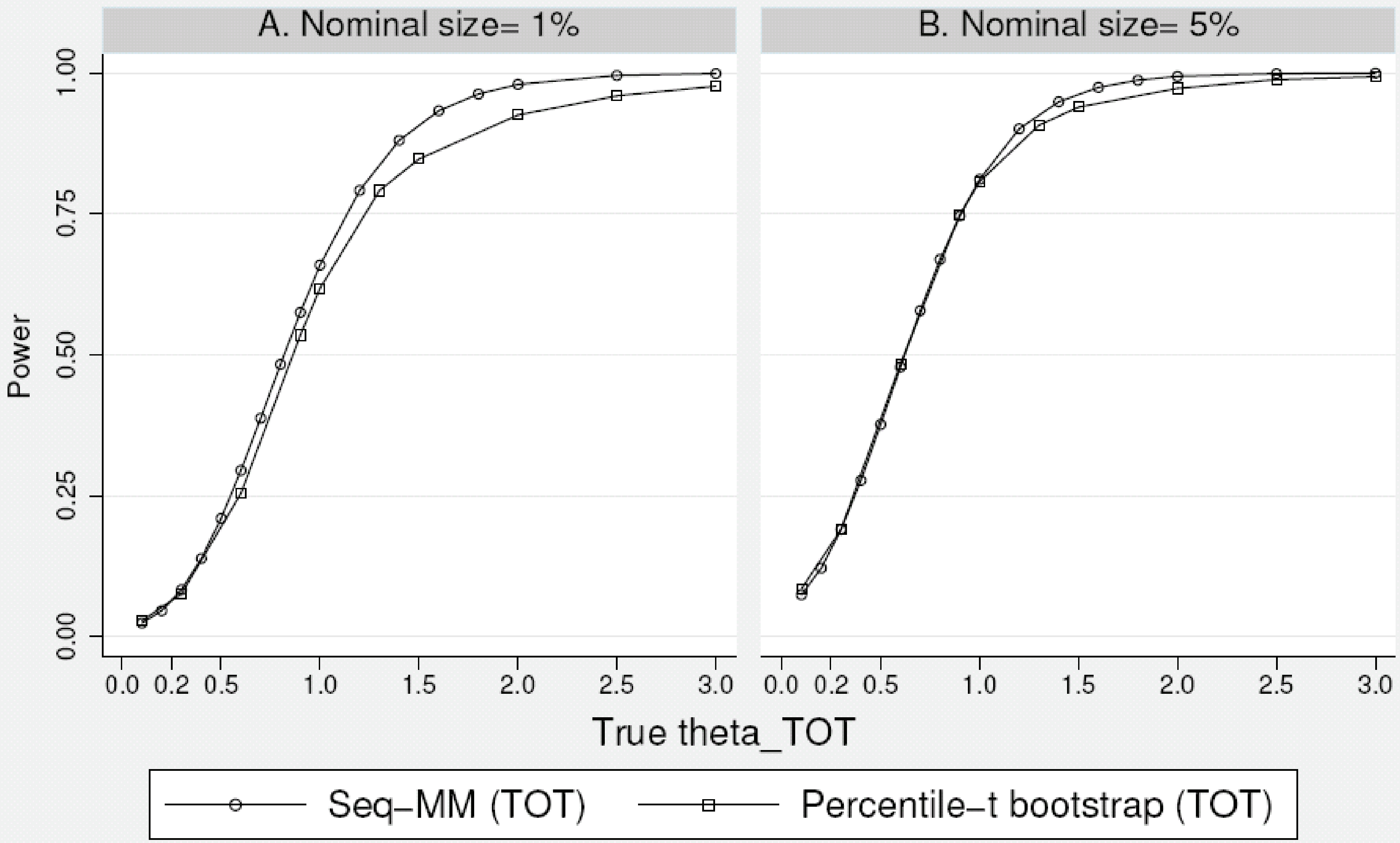
Finite sample deviations from nominal size of t–test, H0: TOT=0



Graphs by Nominal Size

Note. DGP: u~stand.normal, eta=1, rho=0, N=100

Figure 4.5: Power of t–test for H0: TOT=0 against H1: TOT=\=0



Graphs by Nominal Size

Note. DGP: u~cauchy, eta=1, rho=0, N=100

Appendix I. Moments for identification of TOT and ATE

Assume that (1) $Y_i(0) \perp T_i \mid X_i$ and (2)$\xi < p(X_i) < 1-\xi$ for some $\xi > 0$. Then as shown in Imbens (2004)

$$\begin{aligned}
E\left[\frac{T_iY_i}{p(X_i)}\right] &= E\left[\frac{T_iY_i(1)}{p(X_i)}\right] \\
&= E\left[E\left[\left.\frac{T_iY_i(1)}{p(X_i)}\right| X_i\right]\right] \\
&= E\left[\frac{1}{p(X_i)}E\left[T_iY_i(1)|\, X_i\right]\right] \\
&= E\left[\frac{1}{p(X_i)}E\left[T_i|\, X_i\right]E\left[Y_i(1)|X_i\right]\right] \\
&= E\left[\frac{1}{p(X_i)}p(X_i)E\left[Y_i(1)|X_i\right]\right] \\
&= E\left[E\left[Y_i(1)|X_i\right]\right] \\
&= E\left[Y_i(1)\right],
\end{aligned}$$

where the fourth equality is because of (1). Similarly, we can show that $E[Y_i(0)] = E\left[\frac{(1-T_i)Y_i}{1-p(X_i)}\right]$. For the TOT, as shown first in Dehejia and Wahba (1997) and then in Busso and Kline (2008) we have a similar result:

$$\begin{aligned}
E\left[Y_i(0)|\; T_i=1\right] &= E\left[\; E\left[Y_i(0)|\; T_i=1, X_i\right] \mid T_i=1\right] \\
&= E\left[\; E\left[Y_i(0)|\; T_i=0, X_i\right] \mid T_i=1\right] \\
&= \int E\left[Y_i(0)|\; T_i=0, X_i\right]\; dF\left(X \mid T_i=1\right) \\
&= \int E\left[Y_i(0)|\; T_i=0, X_i\right]\; dF\left(X \mid T_i=0\right)\frac{dF\left(X \mid T_i=1\right)}{dF\left(X \mid T_i=0\right)} \\
&= \int w(X)\, E\left[Y_i(0)|\; T_i=0, X_i\right]\; dF\left(X \mid T_i=0\right) \\
&= E\left[w(X_i)\; Y_i(0)|\; T_i=0, X_i\right],
\end{aligned}$$

where the second equality is by (1), the fourth equality is by (2) and

$$w(X_i) = \frac{dF\left(X \mid T_i=1\right)}{dF\left(X \mid T_i=0\right)} = \frac{P\left(T_i=1|\; X_i\right)}{1-P\left(T_i=1|\; X_i\right)}\frac{1-P\left(T_i=1\right)}{P\left(T_i=1\right)} = \frac{p(X_i)}{1-p(X_i)}\frac{1-p}{p}.$$

Appendix II: Derivation of elements of G

Below we show the values of $F(\cdot)$, $F'(\cdot)$ and $F''(\cdot)$ for the normal, logistic, Cauchy and linear probability models. In the binary choice model $V = \kappa X$.

Model	$F(V)$	$F'(V)$	$F^{''}(V)$
Probit	$\Phi(V)$	$\phi(V)$	$-\phi(V)V$
Logit	$\frac{\exp(V)}{1+\exp(V)}$	$F(V)[1-F(V)]$	$F'(V)[1-2F(V)]$
Cauchy	$\frac{1}{\pi}\arctan[V]+\frac{1}{2}$	$\frac{1}{\pi}\frac{1}{1+V^2}$	$-\frac{1}{\pi}\frac{2}{(1+V^2)^2}V$
LPrM	V	1	0

Derivative of $g_1(\kappa)$ with respect to κ which is just the Hessian of the likelihood function of the binary choice parametric model

$$\hat{G}^1_\kappa = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{F''(\kappa X_i)[T_i - F(\kappa X_i)]}{F(\kappa X_i)[1-F(\kappa X_i)]} - \left(\frac{F(\kappa X_i)[T_i - F(\kappa X_i)]}{F(\kappa X_i)[1-F(\kappa X_i)]}\right)^2\right] X_i X_i'.$$

Derivatives g_2 with respect to κ.

$$\hat{G}^{2,t}_\kappa = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta^t Z_i\right)\frac{\partial W_i^t(\kappa)}{\partial\kappa} Z_i' \text{ for } t = ATE, TOT;$$

where,

$$\begin{aligned}
\frac{\partial W_i^{TOT}(\kappa)}{\partial\kappa} &= \frac{(1-T_i) f(\kappa X_i)}{[1-F(\kappa X_i)]^2} C X_i' - \frac{(1-T_i) F(X_i)}{1-F(X_i)} C^2 \left(\frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i) f(\kappa X_i) X_i}{[1-F(\kappa X_i)]^2}\right), \\
\frac{\partial W_i^{ATE}(\kappa)}{\partial\kappa} &= -\frac{T_i f(\kappa X_i)}{[F(\kappa X_i)]^2} C_1 X_i' + \frac{(1-T_i) f(\kappa X_i)}{[1-F(X_i)]^2} C_2 X_i' + \\
&\quad + \frac{T_i C_1^2}{F(X_i)}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{T_i f(\kappa X_i) X_i}{[F(X_i)]^2}\right) - \frac{(1-T_i) C_2^2}{1-F(X_i)}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i) f(\kappa X_i) X_i}{[1-F(X_i)]^2}\right),
\end{aligned}$$

where $F(\cdot)$ is the cdf and $f(\cdot)$ is the pdf of the error term in the selection equation.

Derivatives of g_2 with respect to τ.

$$\hat{G}^{2,t}_\tau = \frac{1}{n}\sum_{i=1}^{n} W_i^t(\kappa) Z_i Z_i'.$$

Bibliography

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 2005, *72*, 1–19.

____ and Guido W. Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, January 2006, *74* (1), 235–267.

Bailey, Martha J. and William J. Collins, "The Wage Gains of African-American Women in the 1940s," *The Journal of Economic History*, September 2006, *66* (3), 737–777.

Barsky, Robert, John Bound, Kerwin Charles, and Joseph Lupton, "Accounting for the Black-White Wealth Gap: A Nonparametric Approach," *Journal of the American Statistical Association*, 2002, *97* (459), 663–673.

Biewen, Martin, "Measuring the Effects of Socio–Economic Variables on the Income Distribution: An Application to the East German Transition Process," *The Review of Economics and Statistics*, February 2001, *83* (1), 185–202.

Busso, Matias and Patrick Kline, "Do Local Economic Development Programs Work? Evidence from the Federal Empowerment Zone Program," *Unpublished manuscript, University of Michigan*, 2008.

____ , Justin McCrary, and John DiNardo, "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," *Unpublished manuscript, University of Michigan and University of Californa–Berkeley*, 2008.

Dehejia, Rajeev H. and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," in "Econometric Methods for Program Evaluation," Cambridge: Rajeev H. Dehejia, Ph.D. Dissertation, Harvard University, 1997, chapter 1.

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

Frölich, Markus, "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, February 2004, *86* (1), 77–90.

Heckman, James J. and Joseph Hotz, "Alternative Methods for Evaluating the Impact of Training Programs," *Journal of the American Statistical Association*, 1989.

____ and R. Robb, "Alternative Methods for Evaluating the Impact of Interventions," in James J. Heckman and R. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press Cambridge 1985.

Hirano, Keisuke and Guido Imbens, “Estimation of Causal Effects Using Propensity Score Reweighting: An Application to Data on Right Heart Catherization,” *Health Services and Outcomes Research Methodology*, 2002, *1*, 259–278.

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

Horvitz, D. and D. Thompson, “A Generalization of Sampling Without Replacement from a Finite Population,” *Journal of the American Statistical Association*, 1952, *47*, 663–685.

Imbens, Guido W., “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *Review of Economics and Statistics*, February 2004, *86* (1), 4–29.

Khan, Shakeeb and Elie Tamer, “Irregular Identification, Support Conditions and Inverse Weight Estimation,” Unpublished Working Paper, Department of Economics, Duke University, Durham, NC August 2007.

Levinsohn, James, Zöe McLaren, and Khangelani Zuma, “HIV Status and Labor Market Participation in South Africa,” *Unpublished manuscript, University of Michigan*, 2008.

MacKinnon, James G. and Halbert White, “Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties,” *Journal of Econometrics*, September 1985, *29* (3), 305–325.

McCrary, Justin, “The Effect of Court-Ordered Hiring Quotas on the Composition and Quality of Police,” *American Economic Review*, March 2007, *97* (4), 318–353.

Murphy, Kevin M. and Robert H. Topel, “Estimation and Inference in Two-Step Econometric Models,” *Journal of Business and Economic Statistics*, October 1985, *3* (4), 370–379.

Newey, Whitney, “A Method of Moments Interpretation of Sequential Estimators,” *Economic Letters*, 1984, *14*, 201–206.

Pagan, Adrian, “Two Stage and Related Estimators and Their Applications,” *The Review of Economic Studies*, August 1986, *53* (4), 517–538.

Robins, James M. and Andrea Rotnitzky, “Semiparametric Efficiency in Multivariate Regression Models With Missing Data,” *Journal of the American Statistical Association*, March 1995, *90* (429), 122–129.

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

Wooldridge, Jeffrey M., “Inverse Probability Weighted M-Estimators for General Missing Data Problems,” *Journal of Econometrics*, March 2007, *141*, 1281–1301.

Chapter 5

Conclusion

This dissertation studies semiparametric methods for the evaluation of social programs. The first essay, with Patrick Kline, evaluates Round I of the federal urban Empowerment Zone (EZ) program, which constitutes one of the largest standardized federal interventions in impoverished urban American neighborhoods since President Johnson's Model Cities program. The EZ program is a series of spatially targeted tax incentives and block grants designed to encourage economic, physical, and social investment in the neediest urban and rural areas in the United States. We use four decades of Census data on urban neighborhoods in conjunction with information on the proposed boundaries of rejected EZs to assess the impact of Round I EZ designation on local labor and housing market outcomes over the period 1994-2000. We use a semiparametric difference-in-differences estimator to estimate the effect of the program on neighborhoods that received EZ status.

Our comparison of EZ neighborhoods to rejected and future EZ tracts in other cities strongly suggests that EZ designation substantially improved local labor and housing market conditions in EZ neighborhoods. The implications of these findings for the study of local economic development policies are manifold. First, it appears that the combination of tax credits and grants can be effective at stimulating local labor demand in areas with very low labor force participation rates. That this can occur without large changes in average earnings suggests either that labor force participation in such neighborhoods is very responsive to wages or that job proximity itself affects participation perhaps due to reductions in the cost of learning about vacancies or the cost of commuting to work. Second, in the case of the EZs, the impact of these demand subsidies does not seem to have been captured by the relatively well off; economic development and poverty reduction seem to have accompanied one another in the manner originally hoped for by proponents of the program. Indeed, our use of disaggregate Census tabulations suggests that even young high school dropouts experienced improved labor market prospects as a result of the program. Third, while the treated communities appear to have avoided large scale gentrification over the period examined in this study, policymakers should consider carefully the potential impact of demand side interventions on the local cost of living. Given that the vast majority of

EZ residents rent their homes, small changes in the cost of zone living can be expected to impose large burdens on the roughly two thirds of the EZ population who do not work. Tradeoffs of this sort should be taken into account when attempting to determine the incidence of the EZ subsidies. If authorities wish to use EZs as anti-poverty programs they may wish to consider combining housing assistance or incentives for the development of mixed income housing as complements to demand side subsidies.

Though our results appear to corroborate the findings of the Abt study, we cannot, with our data, ascertain whether the employment gains of local residents are the result of job growth or the substitution of local workers for outside workers. A detailed analysis of matched employer-employee data might yield insights into whether the scale or substitution effects are responsible for generating the local employment gains observed. More research is also needed to determine whether any job creation that is occurring is due to existing firms expanding, new firms being born, or outside firms relocating.

Finally, this evaluation has only examined the first six years of the EZ program. Very little is known about the dynamics of neighborhood interventions. The decisions of residents, developers, and landlords that lead to neighborhood gentrification and turnover may respond to changes in housing values and rents with a lag. Moreover, as the program comes to a close, firms may move out of zones or close up altogether, reversing any employment gains in the process. Understanding these issues is key to determining the long run winners and losers of EZ designation.

The second essay, with John DiNardo and Justin McCrary, explores the finite sample properties of several semiparametric estimators of average treatment effects. The estimators we consider are semiparametric in the sense that only the treatment assignment process is parametrically modeled. This perspective on estimation encompasses several popular approaches including reweighting, double robust, control function, and matching, but rules out maximum likelihood estimation and estimators based on parametric assumptions on the relationship between the outcome of interest and predicting variables. The semiparametric estimators we consider predominate in the empirical literature. We assess their finite sample properties using simulated cross-sectional data sets of size 100 and 500.

The simulation evidence suggests that when there is good overlap in the distribution of propensity scores for treatment and control units, reweighting estimators are preferred on bias grounds and attain the semiparametric efficiency bound, even for samples of size 100. The double robust estimator can be thought of as regression adjusted reweighting and performs slightly worse than reweighting when there is good overlap, but slightly better when there is poor overlap. Control function estimators perform well only for samples of size 500. Matching estimators perform worse than reweighting if preferences over bias and variance are lexicographic and if good performance for $n = 100$ is required. If there is enough data, then local linear or ridge matching may be competitive with reweighting. The difficulty of the more complicated matching estimators is potentially related to the difficulty of accurate finite sample selection of tuning parameters.[1]

When overlap in the distribution of propensity scores for treatment and control units is close to failing, the semiparametric estimators studied here do not perform well. This difficulty can be inferred from the available large sample results in the literature (Hirano et al. 2003, Abadie and Imbens 2006, Khan and Tamer 2007). We also show that the standard asymptotic arguments used in the large sample literature provide poor approximations to finite sample performance in cases of near failure of overlap. However, our qualitative conclusion is the same as that reached by Khan and Tamer (2007), who note that the semiparametric estimators considered here are on a sound footing only when there is strict overlap in the distribution of propensity scores.

In empirical applications, economists confronting problems with overlap often resort to trimming schemes, in which some of the data are discarded after estimation of the propensity score. We simulate the performance of the estimators studied in conjunction with four trimming rules discussed in the literature. None of these procedures yield good performance unless there is homogeneity in treatment effects along the dimension of the propensity score.

What is then to be done in empirical work in which problems with overlap are suspected? First, to assess the quality of overlap, we recommend a careful examination of the overlap plot, possibly focused on histograms and possibly involving smoothing using local linear density estimation. Second, if overlap indeed appears to be a problem, we recommend analysis of subsamples based on *covariates* to determine if there are subsamples with good overlap. For example, in some settings, it could occur that problems with overlap stem from one particular subpopulation that is not of particular interest. Analyzing subsamples based on covariates is likely to work better than analyzing subsamples based on quantiles of an estimated propensity score. Third, if there is no obvious subpopulation displaying good overlap, we recommend that the economist consider parametric assumptions on the outcome equation. Semiparametric estimators work well in this context when there is good overlap. When overlap is poor, however, these estimators are highly variable, biased, and subject to nonstandard asymptotics. In settings with poor overlap, the motivation for semiparametric estimation is poor and the most effective methods are likely

[1] If preferences over bias and variance are not lexicographic, then some of the biased matching estimators may be preferred to reweighting. We caution, however, that the data generating processes we consider may not represent those facing the economist in empirical applications. In empirical applications, the bias could be of lesser, or greater, magnitude than suggested here, in which case the economist's preference ranking over estimators could be different than that suggested by a literal interpretation of the simulation evidence. Our own preferences over bias and variance lean towards lexicographic because we have a taste for estimators that minimize the maximum risk over possible data generating processes.

parametric approaches such as those commonly employed in the older Oaxaca (1973) and Blinder (1973) (1973) literature.

In the third essay I propose a sequential method of moments variance estimator of IPW estimators of average treatment effects. IPW estimators are becoming increasingly popular to compute average treatment effects. Obtaining valid standard errors for these estimators, however, can be difficult because of the 2-step nature of the estimation procedure. In this essay, I note that IPW is a sequential method of moments (SQMM) estimator which, in cases in which a parametric propensity score model is assumed, has a simple expression of the asymptotic variance. This variance estimator can be used to test not only hypotheses about treatment effects for a given outcome but also hypotheses involving multiple outcomes and/or different estimands. Using Monte Carlo simulations I find that tests based on the proposed SQMM variance estimator have good finite sample size and power compared to competing inference strategies. Tests that ignore the fact that the weights are estimated tend to severely overreject. Tests based on the percentile-t bootstrap method using a bootstrap SQMM variance have very similar size and power properties as the ones obtained using the asymptotic SQMM variance. I interpret this as evidence that the bootstrap percentile-t method is not providing any refinement to the asymptotic variance, which indicates that the SQMM variance estimator is a good enough approximation to the true variance of the treatment effect estimator.

I first propose a simple sequential method of moments variance estimator of inverse probability weighting estimators of average treatment effects. The proposed estimator has several advantages over competing inference strategies. First, it is easy to compute (i.e. faster than the bootstrap). Second, it allows to simply test hypotheses involving different outcomes and estimands because it directly provides an expression for the covariance of the treatment effects across estimands and outcomes. Third, Monte Carlo simulations suggest that the size and power of tests based on the SQMM variance perform well against other competing inference strategy. If anything inference based on the SQMM variance estimator tend to be slightly conservative.

I assess the finite samples size and power of t-tests of no treatment effect using the SQMM variance estimator and find that it performs well even in samples of size 100. The empirical size is close to the nominal size. This is a finding that is robust in settings with homoscedastic and with heteroscedastic errors in the outcome equation. It is also robust to different samples sizes. In contexts in which the ratio of treated to control units is too high t-tests that use the SQMM variance tend to slightly overreject. The same pattern was observed in situations in which the observed maximum estimated propensity score was close to being one and in settings in which the strict overlap assumption, necessary for identification, was violated. In all these DGPs, using the SQMM variance yields better size than alternative estimators that ignore the fact that the weights were estimated.

Finally, I compare inference using SQMM to inference procedures using the bootstrap. The percentile bootstrap tends to reject too often, specially in very small samples. The percentile-t bootstrap implemented using the SQMM variance estimator has similar size and power to the tests using the asymptotic SQMM variance. This is an indication that the bootstrap percentile-t does not provide any refinement to the asymptotic variance, which indicates that the SQMM variance estimator is a good enough approximation of the true variance of the IPW treatment effect estimators.

Bibliography

Abadie, Alberto and Guido W. Imbens, “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, January 2006, *74* (1), 235–267.

Blinder, Alan S., “Wage Discrimination: Reduced Form and Structural Estimates,” *Journal of Human Resources*, Fall 1973, *8*, 436–455.

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

Khan, Shakeeb and Elie Tamer, “Irregular Identification, Support Conditions, and Inverse Weight Estimation,” Unpublished manuscript, Northwestern University 2007.

Oaxaca, Ronald, “Male–Female Wage Differentials in Urban Labor Markets,” *International Economic Review*, 1973, *14*, 693–709.