

New Perspectives on Regression Adjustment in Causal Inference, with Applications to Educational Program Evaluation

by

Adam Chaim Sales

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2013

Doctoral Committee:

Associate Professor Ben B Hansen, Chair
Professor Susan Marie Dynarski
Professor Walter R Mebane Jr
Professor Kerby A Shedden

to Roní, Amitál and Liliana

Acknowledgments

Many many people have helped me make it to this point, a proper subset of whom I will acknowledge here by name. First academic: it is merely analytic to say that without my advisor, Ben Hansen, I would be unable to complete this dissertation; however, to say that Ben added tremendously to its quality, that my scholarship—to the extent that it exists—is due largely to his guidance, or that the past three years would have been much harder and less happy without his support, encouragement and flexibility, is synthetic, and true. The members of my committee, Sue Dynarski, Walter Mebane and Kerby Shedden have each taught me in classes, seminars and especially in personal conversations, and have given me invaluable guidance in life both professional and personal; some of the best ideas in this thesis were responses to their suggestions. Brian Rowan provided me, over the past year, with a job, two datasets, useful comments on my research and, along with Ben, summer tuition funding. Jeff Smith helped me place my results in the context of our sister discipline, econometrics, and in the process helped me better understand my own work. Countless other professors and colleagues deserve my gratitude. The U of M Statistics Department Admins—in particular, Lu Ann, Judy, and Mary Ann—have been really helpful over the past six years.

Next, personal: the person to whom I owe the most gratitude, of course, is my lovely wife, Liliana Martinez, for loving and tolerating me, even when I'm stressed and grumpy, for encouraging me, for making me think, and for raising our two beautiful daughters. She, and, by extension I, have been helped tremendously in the latter by my sister-in-law Angie Martinez and especially by my mother-in-law Nora Martinez, without whom this crazy long-distance arrangement would never have been possible. I owe just about everything in my life to the ones who gave me life, my mom and dad, who raised me with the tools and love of learning which are jointly necessary for a PhD, and have continued to encourage, support and advise me through the process—not to mention, help with thesis proofreading. Say what you will about the tenets of National Socialism, sometimes there's a man, and I'm talking about my brother Ben here, who has helped one hone one's argumentative skills, taught one when to say "this aggression will not stand, man," who, as so many young men of his generation, has some far-out TBL quoting skills (though I don't know why

he has to use so many cuss words) and, in general, makes one's life more fun and thoughtful and thorough. My Ann Arbor roommate/husband/mother/best-friend, Zev "Buffalo Bill" Berger has taken care of me, forgiven my absent-minded neglect, improved my fashion and vocabulary, and made grad school not only tolerable but downright enjoyable.

Finally, I want to thank Sister Ann of the Maria College library and the folks at Tierra Coffee Roasters, both in Albany, NY, for providing me with a table, an outlet, wifi, and, in the latter case, delicious coffee.

Table of Contents

Dedication	ii
Acknowledgments	iii
List of Tables	viii
List of Figures	xi
Abstract	xiii
Chapter 1 Introduction	1
1.1 The Rubin Causal Model and Randomization Inference	4
1.1.1 Notation for Experiments and Observational Studies	4
1.1.2 Analyzing Experiments and Observational Studies	5
1.1.3 When are Observational Studies Believable?	6
1.1.4 Why Not Just Run Regressions?	7
1.2 Applications in Educational Program Evaluation	9
Chapter 2 Analyzing Regression-Discontinuity Designs as Randomized Experiments .	11
2.1 Introduction	11
2.2 Analyzing an RDD as a Randomized Experiment when b is Given	15
2.2.1 Testing Transformed Ignorability: Balance Tests and Diagnostics	18
2.3 Randomization-Based RDD Analysis in the LSO Data	20
2.3.1 Testing Balance of Higher-Order Means	23
2.4 Heterogeneous Treatment Effects: Two Approaches	23
2.4.1 Modeling Treatment Effect Heterogeneity	24
2.4.2 Estimating Randomization-Based Average Treatment Effects	26
2.5 A Data-Driven Approach to Choosing b	28
2.5.1 The Relationship between b^* and the IK Bandwidth	29
2.6 Discussion	31

Chapter 3 Propensity-Score Matching with Very High Dimensional Data: A School-Level Evaluation of a Mathematics Enrichment Program	33
3.1 Introduction	33
3.1.1 Application: an Evaluation of the Agile Mind Algebra 1 Program	35
3.2 Propensity Scores, Prognostic Scores, and Principal Components: A Toolbox for Observational Causal Inference	36
3.2.1 Overview: Propensity-Score Matching	37
3.2.2 Overview: Principal Components Analysis	40
3.2.3 Overview: Prognostic Scores and Principal Components Regression	42
3.3 Combining Principal Components with Propensity Scores: Matching, Evaluating and Estimating	44
3.3.1 Estimating Propensity Scores	45
3.3.2 Evaluating the Match	47
3.3.3 Estimating Effects with Propensity and Prognostic Scores: A Peters-Belson Approach	49
3.4 Evaluating Agile Mind: Background	53
3.4.1 The Agile Mind Algebra 1 Program	53
3.4.2 Defining The Agile Mind Algebra 1 Treatment	55
3.4.3 Selection into the Agile Mind Treatment	56
3.5 Evaluating Agile Mind: Data	57
3.5.1 Texas AEIS	57
3.5.2 Indiana Data	59
3.5.3 Outcomes	60
3.6 Evaluating Agile Mind	61
3.6.1 The Matching Process	62
3.6.2 Outcome Analysis	67
3.6.3 Does AM Affect the Number of 8th-Grade Students Who Pass?	70
3.6.4 Sensitivity Analysis	71
3.7 Discussion	72
Chapter 4 Multivariate Statistics and Machine Learning for Causal Inference: Modern Regression Prognostic Scores	74
4.1 Introduction	74
4.2 The Multivariate Structure of X	76
4.3 Short Overviews of Models to be Considered	78
4.4 Which Models Win? A Cross-Validation Exercise	80
4.5 Test-Sets for Choosing Prediction Models with Less Extrapolation	83
4.6 Conclusion	86
Chapter 5 Multilevel Propensity and Prognostic Score Analysis: An Evaluation of A School-Wide Curricular Program with School- and Student-Level Data	88
5.1 Introduction	88

5.2	Modeling a Multilevel Experiment from Observational Data	89
5.3	Designing a Powerful Test Statistic Using Microdata	90
5.3.1	Multilevel Modeling of Group-Level Treatments with Individual-Level Outcomes	90
5.3.2	Adapting HLM Reasoning to Propensity-Score Designs: Aggregating Outcomes	91
5.3.3	Using Group- and Individual-Level Covariates	94
5.4	Example Dataset: Preliminary Investigation of A School-Level Educational Program	95
5.4.1	Multidimensional Analysis: Principal Components Analysis and Fisher-Transformed Bivariate Correlations	96
5.4.2	The Propensity Model	98
5.4.3	Evaluating the Match: Balance Tests	101
5.4.4	Prognostic Models	102
5.4.5	Outcome Analysis	105
5.4.6	Did Individual-Level Data Help?	107
5.5	Summary	107
	Bibliography	109

List of Tables

Table

3.1	A breakdown of Indiana School-level covariates into types	59
3.2	Covariate balance before and after matching for Indiana middle schools. X1 and X2 are the first two principal components, prog denotes prognostic scores, and p-value denotes omnibus xBalance() χ^2 p-values. There are 505 schools in this stratum, with four treated schools matched to a total of 109 untreated schools.	66
3.3	Covariate balance before and after matching for Indiana high schools. X1 and X2 are the first two principal components, prog denotes prognostic scores, and p-value denotes omnibus xBalance() χ^2 p-values. There are 400 schools in this stratum, with five treated schools matched to a total of 132 untreated schools.	66
3.4	Covariate balance before and after matching for Texas middle schools. X1 is the first principal component, prog denotes prognostic scores, and p-value denotes omnibus xBalance() χ^2 p-values. The third column assesses the match after removing schools for which there was no outcome data available. There are 539 schools in this stratum, with nine treated schools matched to a total of 44 untreated schools	68
3.5	Covariate balance before and after matching for Texas high schools. X1,...,X4 are the first four principal components, prog denotes prognostic scores, and p-value denotes omnibus xBalance() χ^2 p-values. The third column assesses the match after removing schools for which there was no outcome data available. There are 1371 schools in this stratum, with 16 treated schools matched to a total of 94 untreated schools.	69
3.6	Test statistics (weighted difference-in-means averaged over matched pairs) and p-values for the effect of AM in each of the datasets and overall.	69
3.7	Estimates and p-values for the effect of AM on the number of students who took Algebra 1 and the percentage of the total cohort who passed, in each stratum.	70
3.8	Estimates and with 95% margins of error for the effect of AM on the percent of students who passed EOC exams. These margins of error account for possible hidden bias, the hypothetical extent of which is quantified in ρ and T_Z	70
3.9	Benchmarks to choose appropriate values for T_Z and ρ in the sensitivity analysis.	71

4.1	Results from the CV experiment: for each model, LASSO, PCR, Ridge regression or PLS, and for each stratum of the AM dataset, we used 5-fold cross-validation to estimate the optimal tuning parameter λ , and its associated MSE (with standard error SE) and R^2 . We also estimated these quantities (except for the standard errors) for the Random Forest algorithm, using out-of-box error estimates. The table also displays the effective degrees of freedom for each optimal model. For PCR and PLS, this is equal to the number of components in the model; for LASSO this is equal to the number of non-zero coefficients (Zou et al., 2007); for ridge regression, it is the trace of the penalized design matrix and for RF, it is the number of variables considered in each tree.	82
4.2	A potential-match validation of LASSO, PCR, Ridge regression and PLS models in the two Texas strata. The validation uses TAKS scores as outcomes, since EOC scores are not easily accessible for every school.	86
5.1	Sizes of each of the optimal matches for the following matching schemes: matching on only the propensity scores and matching on combinations of the propensity scores and mean ACT composite scores and ACT trends. The last row displays the matches' effective sample sizes.	101
5.2	Covariate balance (standardized differences) for matching, in four scenarios: the unmatched sample, the sample matched using only propensity scores (PS), the sample matched using both propensity scores and ACT prior achievement (PS+ACT) and the match using propensity scores, ACT prior achievement and schools' estimated ACT trends (PS+ACT+Slopes).	102
5.3	Results of a school-level cross-validation study of four modeling strategies, LASSO, PCR, ridge regression and PLS, modeling school-level outcomes. The MSE estimates estimate the MSE of school mean test scores, not individual test scores; similarly, the R^2 's represent the proportions of school-level variation that the models explain.	103
5.4	A comparison of modeling strategies LASSO, PCR, ridge regression and PLS, trained on school- or student-level data or both, using matching validation, as proposed in Section 4.2	104
5.5	Estimates, effect sizes, permutation p-values and 95% confidence intervals for the three different outcomes: prognostic scores (which are not, technically, an outcome, so a null result upholds the model) and unadjusted and adjusted composite average test scores. The test statistic is (5.5). The effect size is the estimate, in points on the test, divided by the overall standard deviation of student scores among the matched group.	106

5.6 The estimate’s sensitivity to hidden bias: sensitivity intervals, accounting for the possible omission of a variable similar to prior ACT scores, % White, % Free or Reduced-Price Lunch, % Special Education, in terms of two parameters: T_Z , which measures a variable’s relationship with treatment assignment, and ρ , which measures a variable’s conditional relationship with the outcome Y . We also considered a “worst case” scenario, with the highest values for T_Z and ρ observed among important covariates. 107

5.7 With only school level data, Estimates, effect sizes, permutation p-values and 95% confidence intervals for the three different outcomes: prognostic scores (which are not, technically, an outcome, so a null result upholds the model) and unadjusted and adjusted composite average ACT PLAN test scores. The test statistic is (5.5). The effect size is the estimate, in points on the test, divided by the overall standard deviation of student scores among the matched group. 107

List of Figures

Figure

2.1	The RDD from LSO. The first-year GPAs were shifted so that the cutoff is at zero—that is, each campus’ cutoff was subtracted from its students first-year GPAs. Subsequent GPA was “cell-mean smoothed”: averaged according to first-year GPA (this causes residual variances to appear inflated towards the edges of the graph; however, the appearance of heteroskedasticity is only because there are few subjects with extreme first-semester-GPAs). The sizes of the plotted circles are proportional to the number of subjects at each value of R . The lines are least-squares linear fits, the red line to the treatment group and the blue line to the control group. The distance between the lines at the cutoff, along the dotted line, is the classic RDD estimate of the “Local Average Treatment Effect.”	13
2.2	A cell-means smoothed plot of $\text{logit}(hsgrade_pct)$ as a function of $dist_from_cut$	20
2.3	A plot of residuals vs fitted-values for the regression of $\widetilde{nextGPA} = nextGPA - 0.23 * Z$ on $dist_from_cut$	21
2.4	Cell-means smoothed plots of (a) $nextGPA$ and (b) $\widetilde{nextGPA}$ versus first-year college GPAs (centered at the cutoff). At each first year college GPA R , this figure plots the average of the $nextGPA$ s of students with first-year college GPA R . The bandwidth for $b = 1.12$; $b = 0.3$ does not include enough unique values of first-year GPA for an easily-interpretable figure.	25
2.5	Confidence regions for ν and τ from model (2.9). The green region is a 90% confidence region, the yellow is 95% and the red is 99%.	26
2.6	P-values from successive balance tests of pre-treatment covariates, one for each possible bandwidth b	30

3.1	A possible pitfall of prognostic-score matching: If prognostic model (3.10) is fit to the control sample but not to the treated sample, overfitting can cause spurious prognostic-score differences between treated and untreated subjects. Here, 100 “treated” and 100 “untreated” observations Y were drawn from the same standard normal distribution, as were 50 “covariates,” which were, in fact, independent of both Z and Y . An OLS model $Y \sim X$ was fit to the control set. The fitted values in the control set were plotted against the model’s predictions in the treated set: in some regions, significant differences in prognostic scores between treated and untreated observations appear, even though all observations were independent and identically distributed. Overfitting is to blame.	43
3.2	Distributions of propensity-score linear predictors for each stratum of the dataset, by AM usage	62
3.3	Covariate balance before and after matching (and with and without schools with missing outcomes—without denoted “CCA”—for Texas schools)	67
4.1	The cumulative proportion of total variance explained by proportion of principal components, for each stratum of the AM dataset	77
4.2	Normal quantile-quantile plots of Fisher-transformed correlation coefficients, multiplied by \sqrt{n} , between individual covariates and outcomes, for each AM dataset stratum, along with a line of slope 1. Corresponding raw correlations are available to the right of the plots.	78
4.3	CV curves for four models, LASSO, PCR, PLS and ridge regression, in each of the four AM strata, clockwise from top left, Texas middle schools, Texas high schools, Indiana high schools and Indiana middle schools.	84
5.1	The cumulative proportion of the data’s total variance accounted for by the first k principal components	97
5.2	Kentucky High Schools, arranged according to the first two PCs. The treated schools are labeled, and the top three positive and negative variables that define the each PC are plotted as arrows. The top positive and negative variables for the first component are in red, and those for the second component are in blue. For clarity, some outlying schools were excluded.	97
5.3	Kentucky High Schools, arranged according to the first and third PCs. The treated schools are labeled, and the top three positive and negative variables that define the each PC are plotted as arrows. The top positive and negative variables for the first component are in red, and those for the third component are in blue. For clarity, some outlying schools were excluded.	98
5.4	Fisher-transformed bivariate correlations between each school-level covariate and the school-mean tests scores.	99
5.5	The distributions of linear predictors from model (5.12)	100

Abstract

Causal inference from observational data—that is, data that did not come from an experiment—is notoriously difficult: because the probability distribution of the treatment variable Z is unknown, measured or unmeasured variables that correlate with both Z and the outcome Y may confound causal estimates. This thesis will present methods for designing and modeling causal observational studies that combine design-based techniques with regression to account for measured covariates X .

Regression discontinuity designs occur when treatment assignment is a function of a variable T : when T exceeds a threshold c , treatment is assigned. Conventionally, researchers analyze RDDs by regressing Y on both T and Z . This thesis argues for modeling RDDs as naturally-randomized experiments in two steps: modeling the relationship between Y and T , and using that design to infer and estimate effects of Z on Y . We illustrate this approach by reanalyzing a dataset used to estimate the effects of academic probation on students' grade point averages.

The rest of the thesis focuses on propensity-score stratification with high-dimensional data ($p \gg n$). If treatment assignment is a random unknown function of X , researchers can adjust causal estimates for X by estimating propensity scores: subjects' respective probabilities of treatment assignment conditional on X . Researchers then stratify subjects based on their propensity scores and model the data as if treatment were randomized within strata. However, when the dimension of X is large, propensity-score estimation is impossible. We propose a method in which a subset of X is used to estimate propensity scores. Next, the entire matrix X can be used to model Y , using a high-dimensional regression technique; the model is trained on subjects excluded from the stratification. The model's predictions of Y can then be used to test balance on, and adjust for, the entire set of covariates in X . We illustrate this method by evaluating two high-school educational programs.

Chapter 1

Introduction

Methods for causal inference from data have, roughly speaking, evolved along two somewhat separate parallel tracks.¹ One focuses on modeling the phenomenon of interest. Along these lines, Pearl (1998, 2000, e.g.), following Haavelmo (1943), emphasizes detection of the “data generating model.” This approach sees a “treatment” (a putative cause), an “outcome” (possible effect) and related variables as part of a system, and models them as such. Of course, one of the most venerable, and almost certainly the most common tools in statistical modeling is the linear regression model, which models an outcome of interest, Y , as a function of a list of explanatory variables, often arranged in a matrix X . Regression is much more than just a causal tool: regression was designed, and has the power, to fully model phenomena: to estimate the function that relates one, or several, variables to each other. Indeed, both Legendre (1805) and Gauss (1809) originally proposed least-squares estimation to model astronomical phenomena. While regression estimates are not generally causal, under certain sets of conditions—in particular, when the regression model is the data generating model—they may be interpreted as such.

Another prominent trend, popularized by Fisher (1935) and Neyman et al. (1935), and epitomized by the randomized controlled experiment, is skeptical of researchers’ ability to estimate a data generating model, and instead, designs experiments specifically to determine whether a causal relationship exists. This approach does not promise a mathematical model of a natural system (how does this system work) as much as one specific, if crucial, aspect of the system: if one were to manipulate a treatment, to what extent would that affect a specific outcome? (see Pearl, 1994; Gelman, 2011) While the goals of experimental inquiry are narrow, its conclusions are dependable. Regression methods depend on sets of assumptions to derive their conclusions; in some scenarios, these assumptions are well-founded, but in others they are not. In contrast, a well-designed experiment, by virtue of its careful design, can be analyzed with virtually no assumptions beyond

¹This is not intended to be a thorough overview of varying philosophies of causal inference, as much as a motivator and (perhaps artificial) structure in which to view this report’s contributions.

the design itself. That is, by randomly assigning a treatment to some subjects but not to others, an experimenter knows enough to infer the existence of a causal effect, as well as a precise measure of uncertainty.

Observational studies that seek to infer causation pose a dilemma: their narrow goals are the same as those of an experiment: the existence and extent of a causal relationship between two variables. The data setting, however, more closely resembles the setting for which regression—especially multiple regression—was developed: a natural system with many interacting and uncontrolled factors. Which paradigm is appropriate? Which methods of analysis are more likely to give rise to believable and useful results?

Some statisticians (Rubin, 2008; Rosenbaum, 2010; Holland, 1986), in some circumstances, argue for expanding the experimental paradigm to include observational studies as well. That is, while some of the advantages of experiments do not extend to observational studies—the mechanism to assign subjects to particular treatment conditions is generally unknown—other advantages, perhaps more subtle, persist into the observational realm. This approach can be referred to as “randomization inference” or “experimental modeling.” However, other statisticians (Yule, 1899; Spirtes et al., 2000; Gelman and Hill, 2007; Bang and Robins, 2005) and allied social scientists (Rindfuss et al., 1980; Tinbergen, 1940; Dielman, 2001, e.g.) believe that the power and flexibility of regression makes it the optimal choice for causal modeling of observational data (see Freedman, 1997, for a broader discussion).

This report argues that, in effect, the dilemma is a false one: at least in some situations, researchers can harness the benefits of regression modeling, while firmly adhering to the design-based logic of experimental modeling. Chapter two addresses the regression-discontinuity design (Thistlethwaite and Campbell, 1960), the common scenario in which treatment assignment is determined as a known function of a continuous “running” variable, which may be correlated with the outcome of interest. When a subject’s running variable is above (or below) a known cutoff, the subject is assigned to the treatment condition. As its name implies, regression—modeling the relationship between the running variable, the outcome, and the treatment—is central to the conventional method of analyzing regression discontinuities. However, following Lee (2005), which argued that in a neighborhood around the cutoff, treatment assignment is random as in an experiment, Cattaneo et al. (2012) has suggested shifting to a strict randomization-inference analysis, essentially excluding regression modeling from the picture. This report, in agreement with Cattaneo et al. (2012), will argue for the benefit of using local randomization to analyze regression discontinuities; it will argue, though, that doing so does not necessitate the abandonment of the conventional regression modeling.

The following three chapters suggest an approach to causal inference in a scenario that will likely prove increasingly common in the “big data” era: more variables than units. Several effective and useful regression methods exist for modeling such data, and chapters three, four and five of this report will suggest a suite of methods, in different scenarios, for harnessing this power in a randomization-inference propensity score design (Rosenbaum and Rubin, 1983a). Chapter three will, in the course of evaluating a high school mathematics enrichment program, propose a theoretical framework for combining high-dimensional regression adjustment and propensity score matching, focusing on principal components regression. Chapters four and five will expand those methods: chapter four will consider more recently-developed high dimensional techniques, along with some methods for choosing which regression strategy to use. Chapter five will, in the course of a preliminary evaluation of another high-school educational program, discuss the use of multi-level data in propensity score and regression studies.

Of course, the distinction between experimental reasoning and regression is not nearly as stark as the discussion here implies. Firstly, across scientific disciplines, randomized experiments are commonly analyzed using regression techniques² (Tang and Tu, 2013; Abdulkadiroğlu et al., 2011; Chattopadhyay and Duflo, 2004; Gerber and Green, 2000, e.g.). Indeed, Fisher himself (Fisher et al., 1970) suggested using analysis of variance modeling, which is essentially a regression technique, to estimate effects in agricultural experiments.³ In regards to observational data, Angrist and Imbens (2002), expressing a common attitude in econometrics, argued for “agnostic regression,” the idea that regression can be effectively used to estimate the narrow causal goals of experiments, provided its results are interpreted carefully. On the other end, randomization-inference proponents commonly use logistic regression or probit regression in the design phase of their studies (Rosenbaum, 2002a).

The approach to be developed here has some precedents in the randomization-inference literature as well. Rosenbaum (2002b), Hansen and Bowers (2009) and Ho and Imai (2006) have suggested methods for regression-based covariance adjustment to increase the precision of randomization-inference estimators. In addition, Rubin (1979, 1973) and Hansen (2008a) have suggested a regression approach to reducing bias in randomization inference based observational studies. This report will build on those methods.

²Freedman (2008) argues against this approach, that multiple linear regression of experimental data can lead to slightly biased estimates and severely biased standard error estimates; see, however, Lin (2013), which argues that these problems are small or easy to fix.

³Pitman (1938), however, derived distribution-free inferential methods for this technique that are explicitly based on the randomization design.

1.1 The Rubin Causal Model and Randomization Inference

1.1.1 Notation for Experiments and Observational Studies

A common and useful approach to formalizing notions of causation and causal inference is called the “Rubin Causal Model” (Holland, 1986; Imbens and Rubin, 2008). Following the framework for randomized experiments in Splawa-Neyman et al. (1990), Rubin (1974) suggested conceptualizing causal effects as comparisons between counterfactual outcomes: between what is, and what could have been. For simplicity, this report will focus on studies with exactly two treatment conditions, referred to as the “treatment” and “control” conditions. Formally, let $Z_i \in \{0, 1\}$ be a random variable coding subject i ’s treatment status ($Z = 1$ signifies treatment). Let Y represent an outcome of interest. Then each subject has two (possibly random) values: Y_{Ci} is subject i ’s outcome *if subject i is not treated* and Y_{Ti} is subject i ’s outcome *if subject i is treated*. For each i , only one of these two values is observed, dependent on Z_i ; subject i ’s observed outcome is $Y_i = Z_i Y_{Ti} + (1 - Z_i) Y_{Ci}$. According to a term coined in Holland (1986), the unobservability of both Y_{Ci} and Y_{Ti} , for a particular subject i , is the “fundamental problem of causal inference”: it is impossible to know, for an individual, what would have been without treatment. The values Y_{Ti} and Y_{Ci} are called “counterfactuals” or “potential outcomes.”

This notation implicitly assumes *non-interference* (Cox, 1958) (also referred to as the “stable unit treatment value assumption (SUTVA)” in Rubin 1978):

Assumption (SUTVA). If $i \neq j$, then subject i ’s treatment does not affect subject j ’s response, or

$$Y_j \perp\!\!\!\perp Z_i \forall j \neq i. \tag{1.1}$$

Assumption (1.1) is necessary for unbiased estimation of treatment effects, but is not necessary for certain inferences.

Subject i ’s “treatment effect” τ_i is defined as $Y_{Ti} - Y_{Ci}$, the difference between subject i ’s treatment and control potential outcomes.⁴ The values τ_i are not observed, or, in general, identifiable; instead, analysts will often estimate aggregate quantities such as the average treatment effect (ATE) $\bar{\tau}$. Alternatively, analysts may assume that $\tau_i = \tau \forall i$, that the treatment effect is identical for all subjects; then, τ may be estimated. If the subjects are modeled as a random sample from a population, the ATE is $\mathbf{E} \tau$, a population mean.

⁴Different parameterizations are also possible, such as the ratio Y_T/Y_C ; however, the difference $Y_{Ti} - Y_{Ci}$ is the most common, and will be the focus of this report.

1.1.2 Analyzing Experiments and Observational Studies

Fisher’s “strict null hypothesis” (Fisher, 1935) is that $\tau_i = 0 \forall i$, that is, that $Y_{C_i} = Y_{T_i} = Y$ for all subjects. Since the value of the vector Y for every possible random draw of Z is hypothetically known, if the distribution of Z is also known, researchers can compute an exact p-value. That is, under the strict null, realized outcomes are invariant to Z , and counterfactual values of T , for various drawings of Z , are also known exactly. Formally, the p-value testing Fisher’s strict null hypothesis is

$$Pr(T(Z, Y) > t) = \sum_{z \in \Omega} \mathbb{1}_{[T(z, Y) > t]} Pr(Z = z) \quad (1.2)$$

where $\mathbb{1}_{[T(z, Y) > t]}$ is equal to one when the hypothetical test statistic T , calculated using the observed values for Y and a hypothetical random draw of Z , z , is greater than the realized value of T , t , and Ω is the set of all possible draws of Z . This analysis ignores variation from randomness in $\{Y_C, Y_T\}$; a researcher may model potential outcomes as fixed, or, equivalently, condition her analysis on their order-statistics, which are sufficient for their distribution (Romano, 2005). In some situations, it is advantageous to estimate this conditional p-value by simulating values from the distribution of Z .

In a simple randomized experiment, the distribution of Z does not depend on the potential outcomes, or on any other (relevant) variables. In that case, the values Y for control subjects—observed Y_{C_i} —are representative of the Y_C values in the entire sample, and the Y values for treated subject, observed Y_{T_i} , are representative of Y_T values for the entire sample. Formally, in a simple randomized experiment, we have that

$$\{Y_C, Y_T\} \perp\!\!\!\perp Z. \quad (1.3)$$

Scenarios in which the distribution of Z is unknown, and must be hypothesized or modeled, but which seek to estimate parameters contrasting treatment and control potential outcomes, have been and will be referred to as “observational studies.” In such settings, (1.3) cannot, in general, be assumed to be true. In particular, there may be a variable W , measured or unmeasured, that correlates with both Z and either Y_C or Y_T ; these relationships induce a “spurious” correlation between Z and the measured outcomes, and violate (1.3). However, in some scenarios, it is reasonable to substitute the following assumption for (1.3):

Assumption (Strong Ignorability).

$$\{Y_C, Y_T\} \perp\!\!\!\perp Z | X \quad (1.4)$$

where X is a vector or matrix of variables to which the researcher has access—covariates—measured

before (or known to be unaffected by) treatment assignment. Under (1.4), within level sets of observed covariates X , there is no confounding, and the data may be modeled as if from a simple randomized experiment. This is referred to as “strong ignorability” (Rosenbaum and Rubin, 1983b), “no unmeasured confounding” (Greenland and Robins, 2009) or “selection on observables” (Heckman and Robb, 1985).

Some more complex randomized experiments use observed covariates X to assign treatment; in this case, strong ignorability (1.4) is known to hold. Either way, in a randomized experiment, the fact that the probability distribution of Z is known exactly ensures not just researchers’ ability to compute exact p-values, but their ability to compute an unbiased estimate of a treatment effect. If (1.3) and (1.1) hold, a simple comparison between treated and untreated units is an unbiased estimate of the sample-average treatment effect $\bar{\tau}$, since the observed Y_C values are a simple random sample of all of the sample’s Y_C values, and the same holds for the Y_T values. Therefore, $\bar{Y}_{Z=0}$, the average of observed control outcomes, is unbiased for \bar{Y}_C , the sample average of all Y_C values, with an analogous result for Y_T values. When (1.3) does not hold, but strong ignorability (1.4) is known to hold, researchers can suitably adjust their analysis methods to produce unbiased estimates.

1.1.3 When are Observational Studies Believable?

The believability of strong ignorability (1.4) is, perhaps, the most vexing problem of causal inference from observational studies. In some settings, it is eminently plausible: for instance, in a regression discontinuity design, as the next chapter will discuss, the treatment assignment mechanism is known. In a broad class of studies known loosely as “natural experiments,” researchers argue—with varying levels of success—that treatment assignment is haphazard, or based (partially) on a variable that is uncorrelated with the potential outcomes (an “instrument”; Angrist et al. 1996). In some other studies, not much is known about the process of treatment assignment; in these scenarios, strong ignorability can be a strong assumption indeed. However, there may be some mitigating factors. The composition of X is surely an important factor: if X contains measures of the outcome variables prior to treatment (pretests), for instance, the bias resulting from assuming strong ignorability may be relatively small (e.g. Cook et al., 2008). Along similar lines, Heckman et al. (1997) emphasize the importance of high quality data for both groups, that all subjects, treated and untreated, exist in the same economic environment and that outcomes in both groups are measured in the same way. A more specific and applied discussion of these and other criteria will appear in chapter three, in the context of an education evaluation.

Even when criteria for a high-quality observational study obtain, and especially when they do

not, causal estimates may be confounded and biased. Of course, one of the central aims of statistics is to quantify uncertainty, and this situation is no exception. Several methods (e.g. Greenland, 2005; Robins et al., 2000; Rosenbaum, 1988; Altonji et al., 2000) are available to assess a studies' sensitivity to departures from strong ignorability, which can help researchers gauge and calibrate their uncertainty regarding this assumption. In chapters three and five, which rely most heavily on strong ignorability, this report will employ the method suggested in Hosman et al. (2010).

1.1.4 Why Not Just Run Regressions?

With the notation here in place, a somewhat more formal account of the difference between the “regression modeling” and the “experimental modeling” paradigms is possible. The distinction comes about in the presence of covariates X ; X , of course, participates in two sets of relationships: relationships between X and the potential outcomes Y_C and Y_T , and between X and Z . If either of these relationships is severed, then X cannot confound a causal estimate of Z on Y ; that is, (1.4) would imply (1.3).

In a simple ordinary least squares (OLS) regression observational study, a researcher will fit the following model:

$$Y = \alpha + \beta X + \gamma Z + \varepsilon \quad (1.5)$$

where α and γ are scalar parameters to be estimated, β is a vector parameter to be estimated, and ε denotes regression errors independent of the regressors X and Z . The parameter γ , here, would be interpreted as a treatment effect: the relationship between Z and Y given X . This regression strategy models both relevant X relationships simultaneously. In fact, the regression estimate of γ is equivalent to the result from the following three-step process:

1. Regress Y on X
2. Regress Z on X
3. Regress the residuals from regression 1 on those from regression 2.

So OLS simultaneously and implicitly models both X relationships in its effect estimate; more complex regression models do the same, if not as neatly.

In contrast, proponents of the experimental modeling approach to observational studies focus on modeling the relationship between Z and X (this is the goal of a propensity score design, as will be discussed in greater detail in Chapter 3). In an experiment, as in an observational study, the

distributions of the potential outcomes are unknown. However, in an experiment, the distribution of Z is known exactly. Therefore, to design observational studies to mimic experiments, researchers must first and foremost attempt to understand the observed distribution of Z . Furthermore, since conditional p-values and hypothesis tests do not depend explicitly on the distribution of Y , the relationship between Y and X is of secondary importance.

This outlook brings with it several distinct advantages (see Rubin, 2007, 2008). First, analogously to randomized experiments, it allows observational studies two distinct phases: design and analysis. The design stage, in which researchers arrange the data so as to mimic data from an experiment, makes use only of Z and X ; Y is unnecessary. This is the stage in which all model building and checking occurs. Researchers can build and test several models without generating several separate effect estimates, and researchers choose a model specification without knowledge of the resulting effect estimate or inference. In this setup, researchers are less able to choose the model whose estimate best confirms their predictions. The second stage is the analysis, in which researchers estimate the effect of Z on Y , based on the design chosen in the first stage. These studies only generate one central⁵ estimate, so some issues of multiple comparisons are avoided.

Another advantage is that a research design that mimics a randomized experiment automatically leads to hypothesis tests that do not refer to Y 's distribution. Hypotheses based on a randomization scheme for Z can be tested with conditional p-values, which can, in turn, be computed exactly. In these scenarios, large sample approximations are unnecessary (though they may, at times, be useful), so small samples pose no inferential problems. For this reason, this style of statistical inference is sometimes referred to as “distribution-free” (eg. Olson, 2013).

Finally, a design mimicking a randomized experiment is bound to be both simple and transparent. For instance, in a propensity-score matching design, under this paradigm, researchers know, and can explain, precisely which treated units are being compared to which untreated units. Then substantive researchers or stakeholders can easily evaluate the design without having to refer to an obscure statistical model (even if an obscure statistical model gave rise to the design in the first place).

One ostensible objection to randomization-based observational study design is that, in some scenarios, there is much to be gained by modeling the relationship between Y and X , as in the regression-based approach. Y - X models can serve two aims: variation in X may help explain some variation in Y , and if this portion of Y 's variance is modeled, effect estimates will be more precise and hypothesis tests will be more powerful. In addition, researchers' X - Z models may be

⁵Nothing stops researchers from checking their estimates' sensitivity to other model specifications after generating their main result; however, these are *post-hoc*, and their estimates are not of interest in themselves, but only as checks to the main estimate. In addition, researchers can compute follow-up estimates, if their first result raises new questions or hypotheses; this process, however, is transparent.

incorrect or incomplete, and modeling Y in these scenarios may correct some of these deficiencies. That is, modeling the relationship between Y and X can improve both the variance and bias of treatment effect estimates. As mentioned, Rosenbaum (2002b), Hansen and Bowers (2009) and Ho and Imai (2006) all suggest similar methods for using regression modeling in the analysis stage of observational studies to reduce estimates' variance. This report will suggest some methods for modeling the X - Y relationship in the design stage to reduce both bias and variance. In the traditional randomization-based observational study, the researcher has no access to Y values in the design stage; the methods proposed here will slightly relax this rule, while preserving its spirit and its advantages.

Randomization-based inference is not the only empirical causal inference strategy, nor is it the only reasonable strategy. However, there are good reasons for researchers to conduct observational causal inference by referencing randomized experiments, and this report will show that, in doing so, they need not sacrifice the power of regression-style modeling.

1.2 Applications in Educational Program Evaluation

Recent decades have seen two important advances in educational program evaluation: first, following a larger trend in the social sciences, education researchers have intensified their skepticism of some statistical methods and, concurrently, use and development of more sophisticated strategies. This trend has occurred at both the grassroots, among individual education researchers, and at higher organizational levels, such as the United States federal government-funded What Works Clearinghouse.⁶ One of the most newly-popular approaches is the regression discontinuity design, which will be applied to an educational question in chapter two. Secondly, as a result of technological and cultural shifts as well as the No Child Left Behind act, there has been a proliferation of data available for analysis. This embarrassment of riches leads to new statistical challenges, which will be addressed in the context of education research in chapters three, four and five. These factors—the need for rigor and the availability of data—make education research an interesting and conducive field for the development of new statistical methods of causal inference, as well as the refinement of existing methods. Academic factors aside, education is a valuable object of study since a good education is both a powerful means towards personal and societal prosperity, equality and justice, as well as an end in and of itself. While the overall focus of this report is methodological, the focus of its applications is educational program evaluation. The methods discussed and developed here

⁶<http://ies.ed.gov/ncee/wwc/> Accessed: 8/26/13.

are designed to be well suited for for improving, by gauging effectiveness, the quality of education.

Chapter two, the next chapter, focuses on a question, and dataset, in higher-education: what is the effect, if any, of academic probation? Academic probation, at its best, is a tool that colleges use to help struggling students, with a combination of carrots and sticks. Lindo et al. (2010) observed that, at one large Canadian university, academic probation was (almost) strictly a function of students' cumulative GPAs, and hence was a regression discontinuity design. This report will re-analyze the data from Lindo et al. (2010), in the service of demonstrating its approach to similar data scenarios in education research.

Chapters three through five use datasets from secondary school evaluations. Agile Mind is a private corporation that offers middle schools and high schools educational enrichment programs that are designed to boost students' achievement—especially the achievement of students who are struggling. Chapter two will attempt to estimate the effect of a school's implementation of the Agile Mind Algebra 1 program on its aggregated scores on standardized-tests . Other than usage information, the evaluation will use exclusively publicly-available school-wide data, of which there is a surfeit: hundreds of variables are available, while only tens of schools use the program. This data scenario, referred to in the statistics world as $p \gg n$, poses a statistical problem, since few $p \gg n$ methods exist for causal inference. (This report, however, is not the only current attempt at this problem; see Belloni et al. 2012 for a more-strictly regression-based approach, and Belloni et al. 2011 for a high-dimensional instrumental-variables model.) In contrast, several $p \gg n$ regression-based methods exist for the purpose of machine learning or prediction; adapting these methods for use in causal inference will be the aim of chapters three, four and five.

Chapter five will present a preliminary analysis of a school-level intervention in six Kentucky high schools, using hundreds of publicly available school-level variables, as well as over 100 student-level variables that have been provided by the Kentucky Department of Education. Although, from a substantive perspective, this analysis is yet incomplete, the dataset will serve as a useful demonstration of how to build on the methods of chapters three and four when student-level variables are available.

Chapter 2

Analyzing Regression-Discontinuity Designs as Randomized Experiments

2.1 Introduction

Randomization of treatment assignment is the gold standard of causal inference in statistics: it is the only way to ensure that there are no confounders biasing a causal estimate. Of the range of options available when experiments are impossible, the “regression discontinuity design” (RDD) (Thistlethwaite and Campbell, 1960; Cook, 2008) approach is among the most credible. In an RDD, the treatment assignment mechanism is known: each subject i is characterized by a “running variable” R_i , and those subjects for whom R_i is greater (or less) than a pre-determined constant c are assigned to treatment. Under seemingly weak assumptions, Lee (2008) has shown that the RDD features “local randomization” of treatment assignment, and is therefore “a highly credible and transparent way of estimating program effects” (Lee and Lemieux, 2010). However, regression discontinuity designs, like all causal analyses, require careful and explicit thought to analyze.

The conventional approach to RDDs attempts to estimate the average functional relationship between the outcome Y , the running variable R and treatment assignment Z , and does not exploit local randomization. Indeed, if local randomization is central to the identification strategy of an RDD, then regression modeling can produce severely biased standard errors (Freedman, 2008).¹ An alternative set of assumptions models treatment assignment itself as random, as in an experiment. A version of this approach appeared in Cattaneo, Frandsen, and Titiunik (2012) (hereafter CFT). This “experimental” modeling approach allows for more explicit and intuitive assumptions, allows analysts to test a wider variety of hypotheses, more-easily allows for small-sample inference, and provides easier ways to test and detect violations of the assumptions and, in some cases, remedies to those violations.

¹See, however Lin (2013)

This chapter will focus on an example of the regression discontinuity design found in Lindo, Sanders, and Oreopoulos (2010) (hereafter LSO). In many colleges and universities, struggling students are put on “academic probation” (AP); the school administration monitors and helps these students more than those who are not struggling. But does AP actually help these students? What is the effect of AP on students’ subsequent GPAs? LSO realized that at a certain large Canadian university, AP was determined almost solely² as function of students’ grade point averages (GPAs): students with GPAs below 1.5 or 1.6, depending on the campus, were put on AP.

According to the formulation in Lee (2008), the RDD insight is that, presumably, there is a random component to each student’s GPA each semester; therefore, for a subset of students whose first-semester GPAs fell close to the cutoff, AP was effectively assigned randomly. Based on this heuristic, CFT argued that analysts may use randomization inference techniques to estimate and infer average treatment effects for this small group of students.³

Alternatively, the conventional approach, which Lee (2008) recommends, argues that one would expect one semester’s GPA to rise, on average, with the previous (first) year’s GPA; if AP had an effect, one would expect a (downward) jump in this relationship at the point at which the first year’s GPA equals the AP cutoff. One could regress next semester’s GPA on the first year’s, and expect the estimated slope to be positive; but if AP had a positive effect, it would provide an additional bump to next semester’s GPA for students below the cutoff. In other words, if the treatment had an effect, the regression line would jump down at the cutoff; by measuring the size of the discontinuity, researchers can quantify the effect of the treatment on the outcome of interest. Figure 2.1 illustrates this analysis.

This chapter proposes a dual approach, combining randomization inference and regression adjustment, along with a procedure, analogous to one suggested in CFT, using pre-treatment covariates to choose the region around the cutoff to study, or to test the validity of a region chosen for substantive reasons. In addition, this chapter will provide a set of randomization-based assumptions under which the conventional method unbiasedly estimates average treatment effects.

In general terminology, the variable upon which treatment assignment depends—in LSO, students’ first-year GPAs—is referred to as the “running variable” (McCrary, 2008) (or “assignment variable,” Berk and Rauma (1983) or “forcing variable” Imbens and Lemieux (2008); in this chapter,

²There are 35 cases, out of a total of 44,362, where students were not put on AP despite GPAs below the cutoff, and three cases in which students were put on AP despite having GPAs above the cutoff. All told, there were 6330 students on AP in this dataset. In this chapter, as in LSO, we will follow the “intent to treat” principle and estimate the effect of treatment *assignment*—that is, having a GPA below the cutoff—instead of actual treatment (AP). Since the number of cases in which these two disagree is such a small proportion of the total, we anticipate that this will have a minimal impact on our estimates

³Some further developments in this direction, from a Bayesian perspective, can be found in Li et al. (2013).

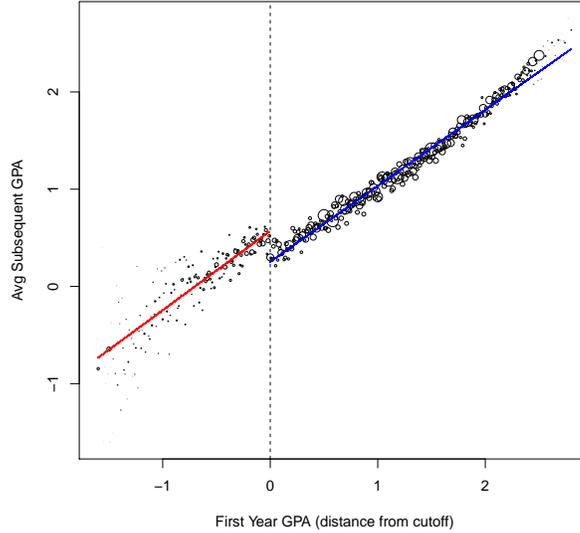


Figure 2.1 The RDD from LSO. The first-year GPAs were shifted so that the cutoff is at zero—that is, each campus’ cutoff was subtracted from its students first-year GPAs. Subsequent GPA was “cell-mean smoothed”: averaged according to first-year GPA (this causes residual variances to appear inflated towards the edges of the graph; however, the appearance of heteroskedasticity is only because there are few subjects with extreme first-semester-GPAs). The sizes of the plotted circles are proportional to the number of subjects at each value of R . The lines are least-squares linear fits, the red line to the treatment group and the blue line to the control group. The distance between the lines at the cutoff, along the dotted line, is the classic RDD estimate of the “Local Average Treatment Effect.”

we will denote it as R), so the conventional, linear regression analysis of an RDD is:

$$Y = \beta_0 + \beta_1 R + \beta_2 \mathbb{1}_{[R > c]} + \beta_3 R : \mathbb{1}_{[R > c]} + \varepsilon \quad (2.1)$$

where Y is the outcome of interest, c is the cutoff, and ε is a mean-0 error (Imbens and Lemieux, 2008).

Imbens and Lemieux (2008) condition their analysis on R , which is typical (Van der Klaauw, 2008; Imbens and Zajonc, 2011; Mealli and Rampichini, 2012, e.g.). In modeling the outcomes, but not treatment status, as random, the conventional approach to RDD does not make use of local randomization. To do so would require precisely selecting a region in which randomization is assumed to take place, and justifying this assumption. Doing so would allow researchers to base their conclusions on the randomization of treatment assignment, which Ronald Fisher famously declared the “reasoned basis for inference,” (Fisher, 1935) and strengthen their causal conclusions.

The question of selecting and verifying a region in which local randomization takes place is related to a prominent discussion in the econometric literature surrounding RDDs. One flaw of the conventional RDD methodology is that a linear model might not accurately account for the relationship between R and Y —when either $R < c$ or $R > c$. The RDD literature has discussed this issue at length; see, in particular, Lee and Card (2008) and Imbens and Lemieux (2008). One common solution is model the relationship with a polynomial (e.g. Oreopoulos 2006). This solution has the disadvantage of allowing observations with large R values, which are generally far from the cutoff, to disproportionately influence the model fit. Hahn et al. (2001) suggests a different approach: fitting a kernel-based “local linear regression” to the data, that weights points closer to the cutoff higher than points far away. Imbens and Lemieux (2008) points out that a simpler version of this approach will work almost as well: simply discard all data outside a given bandwidth—everything of distance h or more from the cutoff c —and fit lines to the data on either side of the cutoff. Making a principled choice for h is still an open question, though recent efforts, particularly in Imbens and Kalyanaraman (2009) (IK) and DesJardins and McCall (2008), have brought us closer to a solution.

This chapter will discuss an approach for selecting, and perhaps widening, a region in which local randomization may take place. Essentially, randomization of treatment assignment, local or otherwise, implies balance of pre-treatment covariates: the theoretical means (and higher moments) of pre-treatment covariates of the subjects selected for treatment and subjects selected for control are equal. Balance of pre-treatment covariates is testable, and will form the basis of our approach to verifying randomization within a small interval around the cutoff. Unlike in CFT, the width of the region of analysis will not depend on at what point random variation in R dominates any relationship between R and Y ; it will, instead, depend on where the relationship between R and Y can be effectively modeled, and suitably accounted for. The chapter will also provide a novel conceptual framework for conducting a randomization-based analysis of RDDs within such a window.

The following section will explain this approach in a situation in which substantive theory or prior research suggests a window. Section three will demonstrate the approach on the LSO dataset, and section four will suggest two methods of accounting for treatment-effect heterogeneity. Section five will discuss a data-driven method for choosing a window of analysis and section six will conclude.

2.2 Analyzing an RDD as a Randomized Experiment when b is Given

The most straightforward randomization-based approach to RDDs is to assume that in a small window around the cutoff $c \pm b$, subjects are randomized to treatment and control conditions as in an experiment. This is roughly the approach discussed in CFT.

Formally, for some $b > 0$,

Assumption (Simple Randomization). $(Y_{Ci}, Y_{Ti}, X_i) \perp\!\!\!\perp Z_i \forall i$ with $R_i \in c \pm b$.

Of course, a method is required of determining the region $c \pm b$ for which this is true; the conventional, OLS-based approach does not address this question. When pre-treatment covariates are available, they may be useful in verifying the simple randomization assumption for an interval-width b . Simple randomization posits that any particular pre-treatment covariate X is independent of treatment assignment Z ; therefore, the sample-means of the covariates in the control and treatment groups,

$$\bar{X}_C = \frac{\sum_i (1 - Z_i) X_i}{\sum_i 1 - Z_i}; \quad \bar{X}_T = \frac{\sum_i Z_i X_i}{\sum_i Z_i} \quad (2.2)$$

are equal in expectation—balanced. One can test this implication for a particular b or for several possible choices of b , and choose a b in which simple randomization seems plausible. This can also be done simultaneously with several pre-treatment covariates X_k $k = 1, \dots, p$.

However, if the potential outcomes Y_C and Y_T and covariates X are correlated with the running variable R , only data in very small windows will exhibit covariate balance. Depending on the criterion for balance, in some cases balance might not be achievable at all. If balance is achievable, it may only be at the expense of the vast majority of the data.

The OLS approach, by modeling the relationship between Y and R , is an attempt to solve this second problem, which remains an issue for a straightforward application of randomization inference ideas. However, as noted above, the OLS approach does not make any reference to or use of randomization of treatment assignment.

A third approach, which combines regression modeling and randomization-inference, could solve both problems. In this approach, we will assume a given $b > 0$ is given that will determine the region of the analysis, and test its validity. The b value will be chosen based on substantive concerns: for which subjects would an estimated treatment effect be interesting, reasonable or coherent? For instance, it makes little sense to ask what the effect of academic probation would be on straight-A students. In following sections, we will propose a data-driven method for choosing b .

The combined approach involves three steps: first, model the relationship between the potential outcomes and R , and use this model to transform potential outcomes Y_C into \check{Y}_C that may be modeled as independent of R . Next, test the validity of the method's assumptions, which will be spelled out shortly, within the given window. Finally, estimate or test the treatment effect using methods from randomized controlled trials.

This approach distinguishes between a design step and an analysis step, as randomized experiments do. However, unlike in a randomized experiment, or in an experimentally-modeled observational study (Rubin, 2008), the design step here involves outcomes Y . However, no estimates or inferences take place in the design stage, which preserves a study's objectivity and avoids some problems of multiple inferences.

The first step is to model the relationship between Y_C and R in order to transform Y_C into \check{Y} : transformed outcome values (hopefully) independent of R . Though other approaches are possible, we will focus in this chapter on a simple-least-squares (SLS) linear regression approach to modeling the relationship between R and Y_C . Under Fisher's strict null hypothesis, $Y_{Ci} = Y_{Ti}$ for all subjects i , so the way to use SLS to model the R - Y_C relationship and construct \check{Y} values is to regress Y on R , and extract the residuals. That is, under the strict null hypothesis, if \mathcal{R} is the matrix formed by joining a column of 1's to the column of R , then

$$\check{Y}_C \equiv Y - \mathcal{R}(\mathcal{R}'\mathcal{R})^{-1}\mathcal{R}'Y \quad (2.3)$$

The new \check{Y} values can be used to estimate and infer causal effects, under the assumption necessary for randomization-based inference:

Assumption (Transformed Ignorability). For all subjects i with $R_i \in c \pm b$, the distribution of R_i is not degenerate, and conditional on the order statistics of R and the potential outcomes \check{Y}_C , R and \check{Y}_C are independent:

$$R \perp \check{Y}_C | (\check{Y}_{C(1)}, \dots, \check{Y}_{C(n)}), (R_{(1)}, \dots, R_{(n)}). \quad (2.4)$$

For simplicity of notation, in the remainder of the chapter we will denote hypothetical \check{Y}_C values simply as \check{Y} . Transformed ignorability (2.4), and conditioning on order statistics, implies that the RDD can be treated as a randomized experiment. Specifically, with R random and $Z = \mathbb{1}_{[R > c]}$ independent of \check{Y} , every combination of R and \check{Y} is equally likely under the null hypothesis—this, in turn, implies the characteristics that recommend randomized experiments: there are no confounders, and the randomization distribution is known.

Transformed ignorability (2.4) depends on picking a b : the width of the region of the analysis. As mentioned above, b can be chosen based on substantive concerns. After choosing b , one can

statistically assess the plausibility of transformed ignorability. If there are pre-treatment covariates available, an analyst can use them to test the assumption. In classical randomized experiments, treatment is assigned independently of pre-treatment covariates, which implies that the means of pre-treatment covariates are balanced, in expectation, between the treatment and control groups. This suggests a crucial test of transformed ignorability, given a procedure for transforming Y into \check{Y} and a region b : transform each covariate X into \check{X} , in the same way the outcomes were transformed, and test the hypothesis that $\mathbf{E}[\check{X}|Z = 1] = \mathbf{E}[\check{X}|Z = 0]$. In other words, analysts can argue for the plausibility of (2.4) by checking the validity of its companion assumption:

Assumption (Transformed Ignorability for Covariates).

$$R \perp\!\!\!\perp \check{X} | (\check{X}_{(1)}, \dots, \check{X}_{(n)}), (R_{(1)}, \dots, R_{(n)}) \quad (2.5)$$

Some classical SLS diagnostic tests can also test transformed ignorability; the following subsection will list some examples.

For transformed ignorability to be plausible, the relationship between Y_C and R must be approximately linear in the region $R \in c \pm b$; if Y_C is modeled as random, and $\mathbf{E}[Y_C|R]$ is differentiable at $R = c$, then Taylor's theorem implies that for some region around c , the relationship is approximately linear.

Transformed ignorability implies a test of the strict null hypothesis of no treatment effect $H_0 : Y_{Ci} = Y_{Ti}$ for all i . Since, conditional on the order statistics, every combination of \check{Y} and Z is equally probable under the null hypothesis, a permutation test of H_0 would achieve the nominal level. In particular, the difference in the mean of \check{Y} between the treatment ($Z = 1$) and control ($Z = 0$) groups could be computed for every possible permutation of Z , holding the vector Y constant: this is the randomization distribution of the difference in means test statistic (Pitman, 1937). For a given α , if the achieved difference in means is greater in magnitude than $1 - \alpha$ proportion of these, then the null hypothesis is rejected at level α . Formally, if $D(Z^*)$ is the difference in means for a given treatment-assignment vector Z^* , Z is the achieved treatment assignment vector, and $n_t = \sum_i Z_i$, the number of treated subjects, then the p-value testing H_0 is

$$p = \#\{Z^* : |D(Z^*)| > |D(Z)|\} / \binom{n}{n_t} \quad (2.6)$$

In practice, p can be computed using monte-carlo methods, or based on a suitable normal approximation, when n is large enough (Good, 2000).

The difference-in-means test statistic is the most straightforward, and emerges most directly

from the comparison with randomized controlled trials, but it is not necessarily the most powerful. Under the alternative hypothesis of a positive additive treatment effect, for instance, so that for all subjects i $Y_{Ci} - Y_{Ti} = \tau > 0$, the slope estimate in equation (2.3) will be biased, so the mean difference between treatment and control values for \check{Y} will most likely be less than τ . A test statistic that may be more powerful against such alternative hypotheses is the F-statistic from the following regression:

$$\check{Y} = \beta_0 + \beta_R R + \beta_Z Z + \varepsilon \quad (2.7)$$

If the model (2.7) is a significantly better fit than a grand mean model $\check{Y} = \beta_0 + \varepsilon$, then it may be concluded that the null hypothesis is implausible. The randomization distribution of this test statistic can be determined through enumeration of permutations of R and \check{Y} , monte-carlo simulations or a normal approximation (Pitman, 1937).

One approach to estimating the magnitude of the treatment effect assumes a constant additive treatment effect, $Y_{Ti} - Y_{Ci} = \tau$ for all i . Under this assumption, for a hypothetical treatment effect τ_0 , Y_{Ci} can be computed for each i as $Y_i - \tau_0 Z_i$. Then, the same permutation-based hypothesis test can be performed for the hypothetical values of Y_{Ci} : regressing Y_C on R , extracting the residuals, and computing the F-statistic (2.7) for the new values \check{Y}_τ (see Rosenbaum, 2002c). One can invert this hypothesis test, testing a range of values for τ . Those values that are not rejected at level α form a $1 - \alpha$ confidence region for τ ; similarly, one can use the Hodges-Lehmann technique (Rosenbaum, 1993) to arrive at a point estimate for τ . Incidentally, the Hodges-Lehmann estimate for τ in this setup is identical to the coefficient of Z from a regression of Y on R and Z (Rosenbaum, 2002c).

2.2.1 Testing Transformed Ignorability: Balance Tests and Diagnostics

The central test of the transformed ignorability assumption (2.4) is the test of covariate balance, described above.⁴ If subjects are essentially randomized to treatment and control conditions, the means of any pre-treatment covariates should be equal between the treatment and control groups: covariates should be balanced. In other words, the analysis should detect no “treatment effect” on pre-treatment covariates. Following this logic, to test transformed ignorability for covariates we may begin by transforming pre-treatment covariates as outcomes were transformed, and testing the strict null hypothesis of no treatment effect.

If several pre-treatment covariates are available, testing transformed ignorability for covariates for each covariate separately could lead to problems of multiple testing, and overly-conservative

⁴CFT also discussed covariate balance as a factor in choosing b ; however, CFT’s balance-assessment procedure differs significantly from the one here.

specification tests. One solution to this problem is to combine information from the covariates into an omnibus test statistic (e.g. Hansen and Bowers, 2007). The appropriate null hypothesis to test is that there is no treatment effect for any of the pre-treatment covariates; indeed, if there were a treatment effect in any one covariate, that would imply a violation of transformed ignorability for covariates. One omnibus test statistic is the maximum of the individual F statistics from testing transformed ignorability for covariates for each of the covariates (see Westfall and Young, 1993). Another possibility for the test statistic is a version of Hotelling's T^2 statistic (Hotelling, 1931): the sum of the individual F statistics. Formally, if $k = 1, \dots, p$ indexes p pre-treatment covariates, and F_k is the F-statistic testing no treatment effect for covariate k , then let

$$F_{tot} = \sum_{i=1}^p F_k \tag{2.8}$$

be the test statistic for overall covariate balance.

As in the case of Y or an individual covariate, the randomization distribution for F_k can be calculated by enumerating all possible permutations of R or approximated with monte carlo methods.

Covariate balance, while a necessary condition of local randomization, is not sufficient. If analysts use regression to generate \check{Y} , as in (2.3), tests of the conventional OLS assumptions (e.g. as in Weisberg, 2005) are in order. If the strict null hypothesis is false, the OLS assumptions in the regression of Y on R will be violated; it is necessary to test transformed ignorability in the absence of a treatment effect. The best way to conduct this test would be to compute the HL estimate of the treatment effect τ_{HL} , subtract τ_{HL} from each of the Y values in the treatment group to calculate Y_{Ci} for each subject i under the null hypothesis that $Y_{Ti}-Y_{Ci} = \tau_{HL}$ for all subjects i . Under this null hypothesis, transformed ignorability is most likely to be confirmed.⁵

In a conference talk, Cattaneo et al. (2010) also proposed subtracting a constant treatment effect from the treatment group and examining the residuals. As in CFT, the paper that emerged from the talk, regression adjustment was not involved. The approach in Cattaneo et al. (2010) was somewhat more formal, involving a dual test of the acceptability of b and τ ; that approach has the advantage of formal hypothesis testing, whereas the approach suggested here is more flexible, allowing researchers to choose their preferred OLS diagnostics, and appropriate interpretations.

The first assessment should be a visual inspection of residuals \check{Y} for any structure. Next, there are several tests in the literature for heteroskedasticity, for instance, such as the Cook-Weisberg test (Cook and Weisberg, 1983) or the Levene test (Levene, 1961). If the variance of \check{Y} is not equal

⁵Since a treatment estimate is necessary for these evaluations, they cannot strictly be considered part of the design stage.

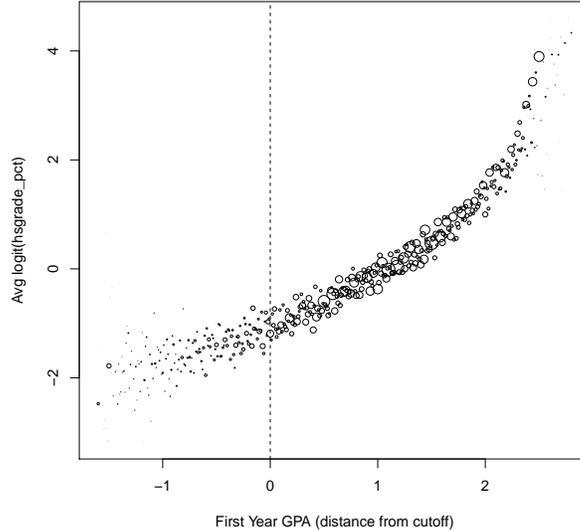


Figure 2.2 A cell-means smoothed plot of $\text{logit}(hsgrade_pct)$ as a function of $dist_from_cut$

between the treatment and the control groups, there may be a violation of transformed ignorability, or the heteroskedasticity may be a consequence of treatment-effect heterogeneity. Therefore, if heteroskedasticity is present, researchers can attempt to model possible treatment-effect heterogeneity to remove the heteroskedasticity. This approach will be treated in section 2.4.1.

2.3 Randomization-Based RDD Analysis in the LSO Data

One of the outcomes that LSO measure is $nextGPA$, students' subsequent GPAs, either for the summer or fall term after students' first years. Their causal question of interest is whether AP causes a change, on average, in $nextGPA$: do students on AP tend to have higher (or lower) subsequent GPAs? Recall that AP is determined almost exclusively by first-year cumulative GPA, which in this case is R : students with GPAs below the cutoff are "treated" with AP, and students above the cutoff are in the control group. One could argue that the most relevant region in which to estimate a treatment effect is within 0.3 grade-points of the cutoff c , since 0.3 conventionally represents the difference in grade points between a C-, say, and a D, or any other grade half-step.

First, we use a covariate, or several covariates, to test transformed ignorability. One suggestive covariate that appears in LSO's data is $hsgrade_pct$, each student's high school grade-point-average.

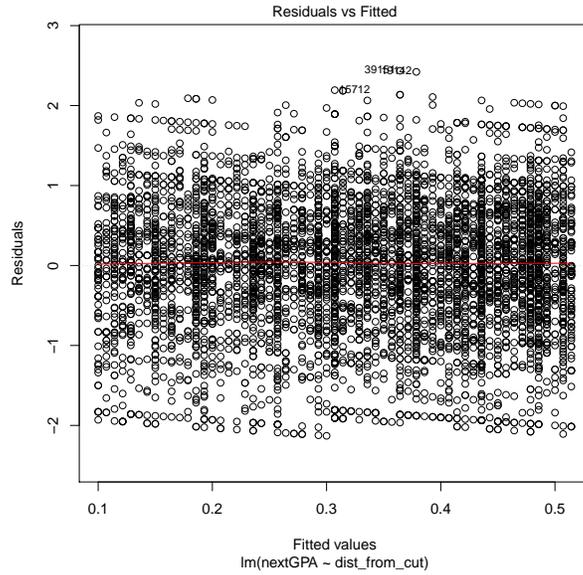


Figure 2.3 A plot of residuals vs fitted-values for the regression of $\widetilde{nextGPA} = nextGPA - 0.23 * Z$ on $dist_from_cut$

Figure 2.2 plots $hsgrade_pct$ as a function of students' first-year GPAs, the running variable, and it seems to have an approximately linear relationship in a neighborhood around c . Linear variation in $hsgrade_pct$ explains about eight percent of the variation in students' first-year GPAs ($\rho = 0.29$) and about four percent of the variation in their subsequent GPAs ($\rho = 0.20$). Since $hsgrade_pct$ is coded with values that fall roughly uniformly between 0 and 100, we use a logit function to transform them $hsgrade_pct \rightarrow \log[0.01 * hsgrade_pct / (1 - 0.01 * hsgrade_pct)]$, to improve the performance of SLS.

We begin by estimating the regression of $hsgrade_pct$ on $dist_from_cut$, and extracting the residuals $\check{hsgrade_pct}$. Next, we can test the null hypothesis that $\check{hsgrade_pct}$ is balanced between treatment and control, using the F-test described above. The permutational p-value for that null hypothesis is 0.22, which means that $\check{hsgrade_pct}$ is more balanced than it would have been in about 22% of hypothetical randomized experiments; at conventional levels, transformed ignorability for covariates is sustained.

For a stronger test, we can test covariate mean balance of several covariates simultaneously. LSO exploits seven pre-treatment covariates to test the RDD assumptions in the AP case: $hsgrade_pct$, the total number of credits attempted in students' first year and students' age at college entry, which are continuous, and dummy variables for campus (the university has three campuses), whether stu-

dents' first language is English, and whether students were born in North America. Using equation (2.8) to simultaneously test balance on all of these covariates yields a permutational p-value of 0.78, sustaining transformed ignorability for covariates.

The next step is to estimate a treatment effect, and to test the strict null-hypothesis of no effect. To test the hypothesis that AP has no effect for $b = 0.3$ means that students whose first-year GPAs were within 0.3 grade-points of the AP cutoff would have the same subsequent GPA regardless of AP assignment. A permutation test of this hypothesis resulted in a p-value of less than $1/500$, so the strict null hypothesis is implausible.

But what is the treatment effect? Using this approach, a 95% confidence interval is the set of τ values whose H_τ is not rejected—this is simply the dual of the hypothesis test if τ is constant. Using this method, a 95% confidence interval for the effect of AP is 0.13–0.33 grade points. Similarly, one would arrive at a point estimate of the effect by determining which hypothetical value of τ_0 produced the highest p-value: this is the Hodges-Lehmann estimate. In the LSO data, the Hodges-Lehmann point estimate was 0.23 grade points; that is, the treatment effect of academic probation is to increase students' subsequent GPA by 0.23 points.

Further tests of transformed ignorability are in order. Using the point estimate of 0.23, we can examine the residuals for evidence against this assumption. Figure 2.3 shows a plot of residuals \check{Y} versus fitted values of the regression in equation (2.3): the regression of $\widetilde{nextGPA} = nextGPA - 0.23 * Z$ on $dist_from_cut$. No structure is readily apparent, so from figure 2.3 there is no reason to suspect a violation of transformed ignorability.

However, it may be useful to statistically test some potential problems. Roughly following Weisberg (2005), one could compare a group-wise model treating $dist_from_cut$ as a factor or grouping variable, to the linear model that treats it as continuous. The F-statistic from this ANOVA comparison 0.92, corresponding to a p-value of 0.65, which also suggests that the residual variance from the linear model is not higher than expected; in other words, the fit is good. The ability to perform this ANOVA is a result of the discrete nature of $dist_from_cut$: in this case, it is an advantage as opposed to a disadvantage.

Similarly, Weisberg (2005) recommends using the Cook-Weisberg test to test for non-constant variance of $next\check{GPA}$. In essence, this regresses transformed residuals on the model fitted values (which, since there is only one regressor in the model, is equivalent to regressing the transformed residuals on $dist_from_cut$) and tests whether the slope on the fitted-values is equal to zero. With a bandwidth of 0.3, the p-value from this test is 0.37, so there is no evidence of non-constant variance.

In the original paper by LSO, the standard errors were clustered by first-semester GPA. Since GPA is essentially discrete, and a function of how many credits a student takes—or even which

classes—it is plausible that students with the same first-semester GPA are similar to each other and therefore have correlated $\check{next}GPA$ values. There is a permutation test that is similarly robust to this clustering: if, instead of allowing all permutations of the treatment variable, we only allow permutations that preserve the grouping structure, the permutation test will be robust to within-group correlation. This is equivalent to assuming that every combination of R and \check{Y} that preserves the grouping structure of R is equally likely; this is somewhat weaker than transformed ignorability (2.4). Weakening transformed ignorability does not significantly change the p-value testing the strict null hypothesis: it is still less than 1/500.

2.3.1 Testing Balance of Higher-Order Means

If transformed ignorability for covariates (2.5) holds and every pre-treatment covariate \check{X} is independent of treatment assignment Z , then \check{X} should be balanced not only in its mean, or first moment, but in all of its moments. Along these lines, some methodologists, such as Caughey and Sekhon (2011), have suggested testing balance of higher-order moments between treatment groups. Incorporating this suggestion into the randomization-based methodology is straightforward: raise the transformed covariate to the specified power p , producing a covariate \check{X}_k^p , and add test balance with \check{X}_k^p as an additional covariate.

In the LSO dataset, a test incorporating 2^{nd} -order natural splines in the balance test for $b = 0.3$ results in a permutational p-value of 0.32.

2.4 Heterogeneous Treatment Effects: Two Approaches

The estimation approach of Section 2, which relies on permutation tests and HL estimates, assumes a constant treatment effect: $Y_{Ci} - Y_{Ti} = \tau \forall i$. However, tests of the strict null hypothesis do not depend on the form or distribution of the treatment effect: under the strict null hypothesis, $Y_{Ti} - Y_{Ci} = 0$ for all i , so the “treatment effect” is a constant 0. Also, under the simple randomization assumption, the difference in means between the treatment and control groups is unbiased for the average-treatment effect. Here we present two approaches to estimating treatment effects under transformed ignorability (or similar assumptions) when the homogeneity assumption is implausible. The first approach continues to rely on permutation tests and HL estimates, but models the treatment-effect heterogeneity. If researchers are confident in their model for the treatment effect, this approach will shed light on how the treatment effect varies. The second approach, in Section 2.4.2, is also

randomization based, but is fairly different from Section 2.2. We show there that an OLS estimate of an average treatment effect will be unbiased given reasonable randomization assumptions.

2.4.1 Modeling Treatment Effect Heterogeneity

The randomization-based approach to RDDs is flexible in that it allows researchers to model the effect of treatment in varying ways, not simply as an additive effect. Given a model for the effect of treatment, an analyst would apply the method suggested above by subtracting the model's hypothetical treatment effect from Y before constructing \check{Y} values. In addition, the analyst must choose the hypothesis test with care, to test the equivalence of the treatment and control groups.

For instance, the methodological RDD literature frequently recommends allowing the slope of the relationship between R and Y to vary from the control group to the treatment group (Lee and Lemieux, 2010). In practice, this means including an interaction term in the linear regression (2.1) interacting Z with R . The randomization-based approach allows researchers to not only estimate but infer the effect of Z on the slope between Y and R . In this more complex model, there are two causal parameters to be estimated: a constant additive effect τ , and an effect that increases (or decreases) linearly with R , that we will denote ν . Therefore, we have that

$$Y_{Ti} = Y_{Ci} + \tau + \nu R_i. \quad (2.9)$$

For a particular null hypothesis $H_{\tau\nu}$, the analyst would subtract $\tau + \nu R_i$ from each member of the treatment group, creating the new variable

$$\tilde{Y}_i = Y_i - Z_i(\tau + \nu R_i) \stackrel{H_{\tau\nu}}{=} Y_C.$$

Under $H_{\tau\nu}$, the slope of the relationship between \tilde{Y} and R should be constant throughout the sample, with no discontinuity, allowing the analyst to use SLS to construct \check{Y} values that satisfy transformed ignorability. Finally, under transformed ignorability, a level- α randomization test designed to test the absence of a treatment of the form (2.9) should reject with probability α .

The challenge, then, is to select a hypothesis tests no treatment effect of the form (2.9). Figure 2.4 illustrates the importance of choosing the right hypothesis test. The difference-in-means test statistic, $\check{\bar{Y}}_T - \check{\bar{Y}}_C$, tests strictly for a difference in means between the control and treatment groups; the parameter ν , however, provides an additional degree of freedom that the difference-in-means does not take into account.

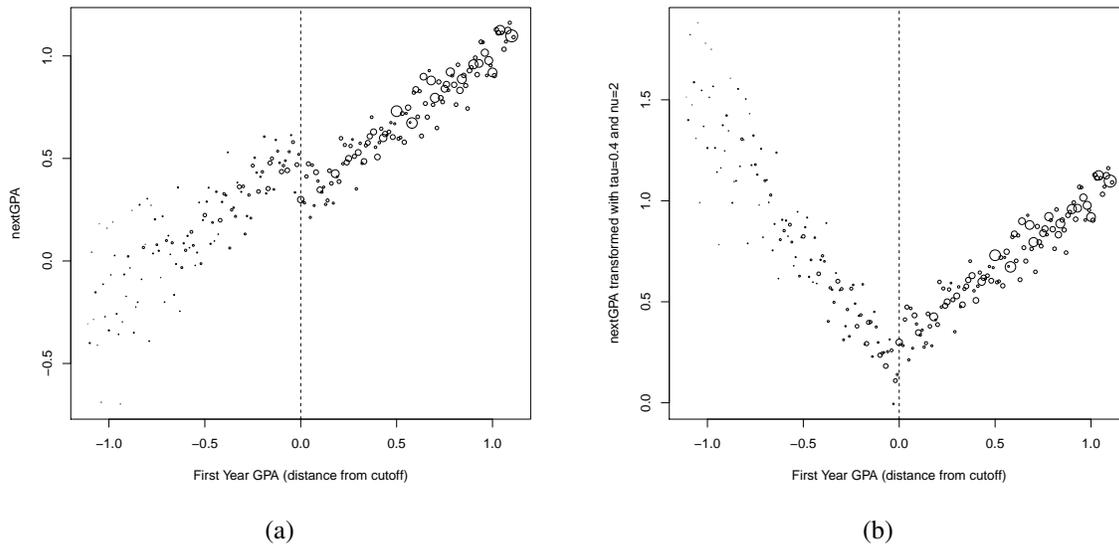


Figure 2.4 Cell-means smoothed plots of (a) $nextGPA$ and (b) $next\check{GPA}$ versus first-year college GPAs (centered at the cutoff). At each first year college GPA R , this figure plots the average of the $nextGPAs$ of students with first-year college GPA R . The bandwidth for $b = 1.12$; $b = 0.3$ does not include enough unique values of first-year GPA for an easily-interpretable figure.

In Figure 2.4, the Y values are transformed with $\nu = -2$ (which seems impossible) and $\tau = 0.4$, resulting in an permutational p -value $p > 0.6$, with an observed difference in means of approximately 0.009. While the difference-in-means test, as used, fails to reject this pair of values, Figure 2.4 clearly shows that $\tau = 0.4$ and $\nu = -2$ do not suitably transform the data to be roughly equivalent on either side; they are not values a researcher would want in her $1 - \alpha$ confidence region.

Fortunately, an alternative hypothesis test performs better. One hypothesis test that directly tests both equal slopes and equal intercepts in the treatment and control groups is a classical analysis-of-variance F-test. After transforming the data with a hypothetical τ and ν pair and computing \check{Y} , an analyst would use a permutational F-test to compare a model regressing \check{Y} on a constant, R , Z and the interaction $R : Z$ with the regression of \check{Y} on just a constant. This test would reject if it found that either the slope or the intercept varied from one group to another. It would not reject, however, if other parameters in the two groups differed, such as the residual variance. Inverting this testing procedure, by sweeping over values of both τ and ν simultaneously, yields a $1 - \alpha$ confidence region for τ and ν .

The results of this procedure, applied to the LSO dataset, appear in figure 2.5; the green region is a 90% confidence region, the yellow is 95%, and the red is 99%.

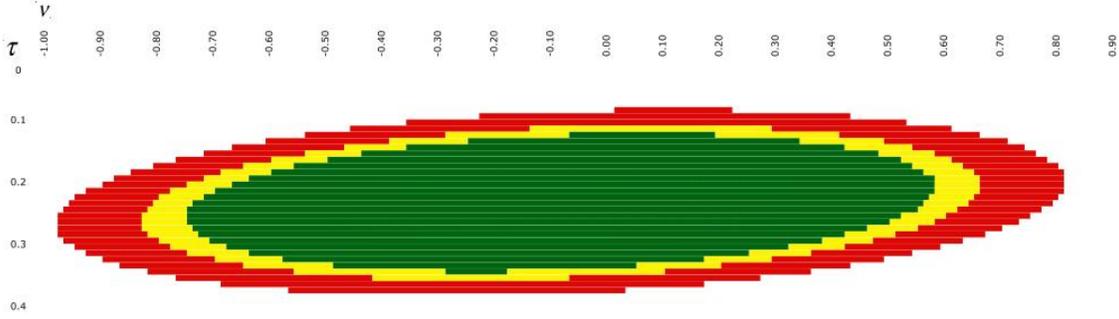


Figure 2.5 Confidence regions for v and τ from model (2.9). The green region is a 90% confidence region, the yellow is 95% and the red is 99%.

Marginally, τ varies, at most, from 0.1 to 0.35, with 95% confidence, and v varies from -0.96 to 0.82, also with 95% confidence. Apparently, allowing the slope to vary from treatment to control does not substantially change our conclusion about τ , but the data are, somewhat surprisingly, fairly uninformative about v .

2.4.2 Estimating Randomization-Based Average Treatment Effects

To discuss the average of heterogeneous treatment effects, let $\vec{\tau}$ represent the vector $Y_T - Y_C$. Let $\bar{\tau}$ be the average of $\vec{\tau}$, the average treatment effect in the sample. If one subtracts $\bar{\tau}$ from the Y value of the treatment group, one will still have failed to recover the values Y_{Ci} , so if one constructs \check{Y} as in Section 2.2, transformed ignorability will not hold for the transformed outcomes. Therefore, it is not clear if the method outlined in Section 2.2 can be used to estimate an average treatment effect in the presence of heterogeneity (though, as mentioned, heterogeneity does not affect the validity of tests of the strict null hypothesis).

A slightly different approach to estimation is necessary in the presence of heterogeneous treatment effects. This approach begins by specifying some alternative, but similar, assumptions to transformed ignorability:⁶

Assumption (ATE-1). $\forall i$ with $R_i \in c \pm b$, with constants α_i and τ_i , $i = 1, \dots, n$:

1. The distribution of R is not degenerate

⁶While this approach is similar in spirit to the approach in Section 2.2, it is technically incompatible: transformed ignorability conditions on the order statistics of \check{Y} , whereas this approach treats Y as fixed; under those conditions there are often not enough degrees of freedom for R and \check{Y} to vary independently, as transformed ignorability demands.

$$2. y_{Ci} = \alpha_i + \beta R_i$$

$$3. Y_i = y_{Ci} + \tau_i * \mathbb{1}_{[R_i > c]}.$$

Following the lead of Freedman (2008), it is possible to re-write Assumptions ATE-1 in a more familiar form. For ease of notation, let $Z_i = \mathbb{1}_{[R_i > c]}$. Represent α_i as $\bar{\alpha} + \tilde{\epsilon}_i$ and τ_i as $\bar{\tau} + \eta_i$, and let $\epsilon_i = Z_i \eta_i + \tilde{\epsilon}_i$. Note that the only random quantity in this set-up is R , and by extension Y , Z , and ϵ . The vectors $\tilde{\epsilon}$ and η , however, are not modeled as random.

Then, we can re-write assumption ATE-1 as

$$Y_i = \bar{\alpha} + \beta R_i + \bar{\tau} Z_i + \epsilon_i. \quad (2.10)$$

Here, $\bar{\tau}$, the average treatment effect (ATE), is the target estimand. Unlike in the classical OLS setup, ϵ_i 's randomness is purely a function of Z_i ; it is not homoskedastic and for $i \neq j$ it cannot be assumed that $\epsilon_i \perp \epsilon_j$.

However, it is still desirable that $\sum_i \mathbf{E}[Z_i \epsilon_i | n_t] = 0$; for that to be true, another assumption is necessary:

Assumption (ATE-2). Let $n_t = \sum_i Z_i$, the number of treated subjects. Then, for all subjects i with $R_i \in c \pm b$

$$\sum_i \mathbf{E}[Z_i | n_t] \eta_i = 0 \quad (2.11)$$

and

$$\sum_i \mathbf{E}[Z_i | n_t] \tilde{\epsilon}_i = 0. \quad (2.12)$$

Equation (2.11) puts a limit on the extent of the treatment-effect heterogeneity: in aggregate, the random assignment of treatment cannot correlate with the effect of treatment. Equation (2.12) is basically a no-confounding assumption: there can be no correlation between Y_C and treatment assignment, other than the component of Y_C explained by R . All told, Assumption ATE-2 ensures that $\sum_i \mathbf{E}[Z_i \epsilon_i | n_t] = 0$.

Under these assumptions, the OLS estimator for $\bar{\tau}$ is unbiased:

Proposition 1. Let \mathcal{R} be the $n \times 3$ matrix formed by joining a $n \times 1$ column of 1s, a column of R and a column of Z . Then, under Assumptions ATE-1 and ATE-2, the OLS estimator is unbiased:

$$\mathbf{E}(\mathcal{R}' \mathcal{R})^{-1} \mathcal{R}' Y_3 = \bar{\tau} \quad (2.13)$$

Proof. Let $\vec{\beta} = \{\bar{\alpha}, \beta, \bar{\tau}\}'$. So Assumption ATE-1 is that $Y = \mathcal{R}\vec{\beta} + \varepsilon$. We have that

$$\begin{aligned} (\mathcal{R}'\mathcal{R})^{-1}\mathcal{R}'Y &= (\mathcal{R}'\mathcal{R})^{-1}\mathcal{R}'(\mathcal{R}\vec{\beta} + \varepsilon) \\ &= \vec{\beta} + (\mathcal{R}'\mathcal{R})^{-1}\mathcal{R}'\varepsilon \end{aligned}$$

To show that the third element of $\mathbf{E}(\mathcal{R}'\mathcal{R})^{-1}\mathcal{R}'\varepsilon$, corresponding to the error in the estimate of $\bar{\tau}$, is 0, condition on $\{R_{(1)}, \dots, R_{(n)}\}$, and note that $\mathcal{R}'\mathcal{R}$ is invariant to permutations of R or Z . This is because Z is a deterministic function of R . Also, note that $\mathbf{E}[Z_i|\{R_{(1)}, \dots, R_{(n)}\}] = \mathbf{E}[Z_i|n_t]$, since Z_i is binary.

$$\begin{aligned} \mathbf{E}(\mathcal{R}'\mathcal{R})^{-1}\mathcal{R}'\varepsilon &= \mathbf{E}\mathbf{E}[(\mathcal{R}'\mathcal{R})^{-1}\mathcal{R}'\varepsilon|\{R_P(1), \dots, R_{(n)}\}] \\ &= \mathbf{E}(\mathcal{R}'\mathcal{R})^{-1}\mathbf{E}[\mathcal{R}'\varepsilon|\{R_P(1), \dots, R_{(n)}\}] \\ &= \mathbf{E}(\mathcal{R}'\mathcal{R})^{-1}\left(\sum_i \mathbf{E}[\varepsilon_i|n_t], \sum_i \mathbf{E}[R_i\varepsilon_i|\{R_P(1), \dots, R_{(n)}\}], \sum_i \mathbf{E}[Z_i\varepsilon_i|n_t]\right)' \end{aligned}$$

Assumption ATE-2 guarantees that the third element above, corresponding to error in the estimate for $\bar{\tau}$, is equal to zero. \square

A stronger assumption would be necessary for unbiasedness in β , but since this is not the target of estimation, strengthening the assumption would be unnecessary.

In principal, it is also possible to calculate conservative randomization-based standard errors in the manner of Freedman (2008) or Gadbury (2001); however, this is beyond the scope of this chapter. Similarly, the same approach as above shows that that the OLS estimate of $\bar{\tau}$ from equation (2.1) is also unbiased, since both R and $R : Z$ are deterministic functions of R .

2.5 A Data-Driven Approach to Choosing b

If theory or past research do not suggest a bandwidth b , one can sweep over the set of possible b s to find one.⁷ As b increases, so does sample size, and hence power and precision. Therefore, the optimal bandwidth b^* is the maximal b such that the RDD estimator, applied to data within the window of analysis, has approximately zero bias. In other words, b^* is the largest b for which transformed ignorability is plausible. To choose an optimal b^* , perform the balance test on ever-

⁷CFT discusses an approach similar in spirit, if not in detail, to the one here.

expanding data sets, letting b vary over its entire range. By examining the p-values from this set of hypothesis tests, one should be able to identify at which b the \check{X} s exhibit enough imbalance to render transformed ignorability implausible. For instance, an analyst might decide, in advance, that she would be willing to tolerate covariates that are more balanced than, say, 10% of randomized experiments—in other words, reject a balance hypothesis with a p-value of 0.1 or smaller. Then, she would incrementally increase b , and at each b test covariate balance as described in Section 2.2.1. She could pick b^* as the largest b corresponding to a p-value above 0.1.

There are two issues with this approach: the first is the arbitrariness involved in picking the level of the test at which to reject. The second is that p-values naturally fluctuate; indeed, under the null hypothesis a p-value is drawn from a uniform $(0, 1)$ distribution, and, of course, will fall below 0.1 10% of the time. This is related to the problem here of multiple-comparisons, one for each possible bandwidth.

A different, though equally valid approach is to choose an data-driven bandwidth on a case-by-case basis, by inspecting the pattern of p-values at different bandwidths. For instance, if the graph of p-value versus bandwidth shows a fluctuating pattern followed by a steeply decreasing pattern, the b at which the pattern changes would be a reasonable choice for b^* ; for values of b greater than b^* , there may be patterns in the data, reflected in the covariates, that increasingly invalidate transformed ignorability.

Regardless of the method analysts use to pick b^* , it is good practice to estimate the treatment effect or ATE at several plausible b values, as a sensitivity analysis. If the ATE estimates vary considerably—that is, show a high sensitivity to b —there may be deeper problems with the model.

Figure 2.6 shows the p-values that form successive balance tests at each possible bandwidth b ranging from 0.01 to 1.0. Based on Figure 2.6, one could choose a b^* as 0.6, since this seems to be approximately where the downward trend in p-values begins. At $b = 0.6$, the HL estimate of a treatment effect is 0.23 grade points, with a p-value of less than 1/500 and a 95% confidence interval of 0.16–0.29, roughly the same as the estimate in Section refLSOsec.

2.5.1 The Relationship between b^* and the IK Bandwidth

As noted above, the problem discussed in this chapter bares a close similarity to the problem in conventional RDD analysis of choosing a bandwidth h for a local-linear regression in an RDD. The approaches in IK and DesJardins and McCall (2008) seek to minimize the mean-squared-error of the RDD estimate. IK bases its approach on the RDD analysis based on local linear regression (Porter, 2003), which uses a kernel to weight points closer to the cutoff more than points farther,

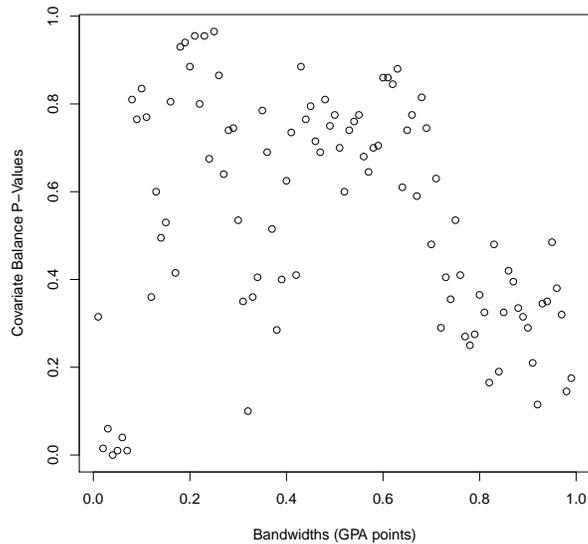


Figure 2.6 P-values from successive balance tests of pre-treatment covariates, one for each possible bandwidth b .

and runs a linear regression on the weighted data. The trade-off in their approach is that as the bandwidth increases, so does the sample size, so the standard error of the estimate decreases. On the other hand, as h increases, so does the potential bias from a non-linear relationship between R and Y , the outcome. The IK bandwidth procedure non-parametrically estimates the curvature of the relationship and balances that against the standard error improvements that accompany a wider bandwidth.

One potential problem with this approach is that mean-squared-error is a sum of the variance (squared standard error) and squared bias of the estimator. While decreasing the standard error is indeed a worthwhile aim, it is questionable whether doing so at the expense of incurring substantial bias is advisable. Standard error is a quantity that is estimable from the data, and can be quantified and accounted for in the presentation of the data's results; this is not the case with bias, which might cause hypothesis tests and confidence intervals to depart from their nominal levels.

The approach outlined here seeks to avoid these problems by constraining bias to be close to zero. Covariate balance, indeed, may be interpreted as an assessment of the validity of the linear approximation to the relationship between R and Y : if we assume that the relationship between at least one of the covariates and Y has a similar form to the relationship between R and Y , then departures from linearity in the former relationship would indicate similar departures in the latter.

Conversely, if at a particular bandwidth b a linear regression correctly estimates the discontinuities of the covariates at the cutoff of R (that is, zero), then it is reasonable to assume that, within b of the cutoff, the relationship between R and Y is also approximately linear. Imbens and Lemieux (2008) suggests that fitting simple linear regressions to only the data within b of the cutoff is a simple way to estimate a local linear regression, so using b^* as this chapter has suggested may be roughly equivalent to estimating a local linear regression RDD analysis with a bandwidth that constrains bias at approximately zero. This argument may serve as an additional, secondary justification for using covariate balance tests to choose a bandwidth for an RDD analysis.

2.6 Discussion

This chapter presents a novel interpretation and modeling approach to regression discontinuity designs. The new approach has several advantages over the conventional approach. Its assumptions are (arguably) more intuitive than the conventional RDD assumptions (see eg. Imbens and Lemieux, 2008). Transformed ignorability speaks directly to the data's similarity to a randomized experiment, which should stand at the heart of causal inference. This approach is robust for to small-sample inference, since it relies on permutation tests, as was demonstrated in CFT. Similarly, the conceptual approach does not demand continuity in the running variable R , since it does not rely on taking limits of continuous functions of R . Finally, the approach presented here suggests two simple and intuitive ways to choose a bandwidth for the RDD analysis: preferably, to use prior substantive knowledge to choose a region in which an estimated treatment effect or ATE is desirable and interpretable (and then to validate this bandwidth with a balance test) or, alternatively, to use sequential balance tests to allow the data to suggest a bandwidth.

The approach presented here may have some weaknesses as well. Firstly, it requires a set of covariates in the data whose relationships to the outcome of interest are similar to the running variable R 's. Not only are such covariates not always available, but even in a rich dataset it may be hard to assess the covariates' usefulness. Secondly, there may be some scenarios in which the conventional RDD assumptions are more plausible than those presented here.

These issues suggest some further opportunities for future research. Researchers should test the new method, both on additional datasets that have already been analyzed in the conventional manner and with monte-carlo simulation methods. Hopefully, these studies could clarify the new method's relationship to the conventional method: is one method more powerful than the other? Under which scenarios do the two methods differ, and under which scenarios do they produce similar estimates?

Which method tends to perform better?

Though starting from a similar point, this chapter's approach differs in some important ways from the approach in CFT. CFT suggests assuming that subjects in a window close to the cutoff are randomized, with equal probabilities to various treatment conditions. In particular, within the window of analysis, the potential outcomes are not related to treatment assignment—and, therefore, the running variable R . In contrast, this chapter allows a relationship between potential outcomes and R in the window of analysis. The question of which set of assumptions is more plausible will depend on the substantive question of interest. CFT's approach allows more flexibility in the choice of estimands, whereas the approach here is limited to estimating mean-based estimands. This chapter's approach will hopefully be attractive to researchers who are drawn to randomization-based arguments, but are reluctant to abandon the familiar trappings of the conventional RDD approach.

Finally, since the methods presented here come from the family of randomization-inference statistical methods, the Rosenbaum-bound sensitivity analysis techniques (Rosenbaum, 2002a) may apply. A sensitivity analysis tailored to RDDs, along with an appropriate interpretation, would bolster the believability of the resulting statistical estimates.

Chapter 3

Propensity-Score Matching with Very High Dimensional Data: A School-Level Evaluation of a Mathematics Enrichment Program

3.1 Introduction

The No Child Left Behind Act (NCLB) of 2001 mandated the collection and recording of large amounts of data, and has left educational researchers with a dilemma: are these newly (or, in some cases, not-so-newly) available data useful? If so, new statistical techniques are necessary to optimally incorporate all available data into causal estimates. This is a concern, in particular, in observational causal inference, when there may be a large number of observed, potential confounders. Inferences from observational studies are susceptible to bias from confounders; hence, the question of how to best adjust causal estimates for observed, pre-treatment covariates is a central question in educational studies. Of course, almost any dataset can become high dimensional, as researchers can examine the raw covariates themselves, along with functions of the covariates, such as splines, polynomials, or interactions. In this sense, high-dimensional data has always been available, but, perhaps, out of reach.

Indeed, in some publicly-available, school-level education datasets—including those that are analyzed in this chapter as well as in Chapter 5—much of the high-dimensionality is due to disaggregation. That is, where a smaller school-level dataset would include, for instance, school average test scores and simple demographic breakdowns (percent Black, percent economically disadvantaged and the like), larger datasets, because of statutory and administrative factors, include interactions between these variables (such as average test scores for particular subgroups of students). In that sense, these new variables do not provide any new information that would not be available in a

student-level dataset. However, as mentioned, researchers may be interested in these interactions, and want to include them in their models, as the approaches that will be presented here allow. In addition, in some cases even student-level data, without interactions between variables, contain enough variables to stymie a regression when the sample size is small or moderate. This is all to say that the question of high-dimensional data is not only a result of researchers' reliance on school-level data, but also must be addressed when student-level data are available. Chapter 5 of this report addresses an example in which this is the case.

This paper follows a long tradition in statistics of modeling observational data as emerging from a hypothetical and unobserved—but hopefully reconstructible—block-randomized experiment (see Rubin, 1977, 2008; Rosenbaum, 2010). That is, similar subjects were grouped into strata, and a randomly-selected subset of each stratum was assigned to treatment. Of course, this model is not literally true, in the sense that such a deliberate randomization did not take place; that being said, it may be quite useful (à la Box and Draper 1987). In particular, if treatment status is a random function of observed variables ("selection on observables," Heckman and Robb 1985) then this experimental modeling approach is likely to yield nearly unbiased estimates of treatment effects under a lucid and transparent framework. Further, this approach comes with natural extensions that relax the selection-on-observables assumption (eg Rosenbaum, 1987; Hosman et al., 2010).

Under this model, the central challenge in an observational study is to reconstruct the design of the block-randomization; that is, to reconstruct the strata. When treatment assignment depends on one observed covariate x , reconstructed strata will also depend on that covariate (Rubin, 1977). In other words, an analyst would assume that subjects with similar values of x were placed in the same stratum, and that treatment was randomized within these strata.

When treatment assignment may have been a function of several "assignment" covariates, the curse of dimensionality makes direct stratification very difficult, or impossible. In such situations, analysts can attempt to reconstruct the hypothetical strata by estimating subjects' propensity scores: their respective probabilities of treatment, conditional on the assignment covariates (Rosenbaum and Rubin, 1983b, 1984). Then, analysts would group subjects with similar estimated propensity scores, and posit a block randomization within those strata.

Actual propensity scores, though, are estimates, and estimating propensity scores can be a challenge. Conventionally, researchers will use logistic regression or a variant thereof (Rosenbaum, 2002d), modeling treatment assignment as a function of the covariates. A surfeit of covariates can doom this modeling step, as well: the propensity score estimates lose stability as the number of covariates grows relative to the number of observations, and becomes impossible when the number of covariates is greater than the number of observations. This paper suggests an approach

to overcoming this difficulty, based on one common multivariate statistical analysis technique: principal components analysis (PCA).

Previously, the standard approach to causal inference with many covariates was theory-based covariate selection: choosing a small number of covariates that substantive theory and prior knowledge suggest to be important predictors of both treatment assignment and outcomes. In many applications, this approach may be sufficient. In particular, if the researcher has, either because of a robust, thorough and (sufficiently) correct theory or because of luck, chosen the correct subset of covariates, a treatment-assignment or outcome model based on this subset will sufficiently eliminate confounding bias. However, when a researcher is not entirely confident in his model, a method that attempts to account for all available covariates may be attractive.

This chapter suggests beginning an analysis in the traditional way, with theory-based covariate selection, but then augmenting the analysis with PCA-based multivariate techniques to account for all of the available covariates.

3.1.1 Application: an Evaluation of the Agile Mind Algebra 1 Program

Agile Mind, a company founded in 2001, produces tools to help secondary science and mathematics teachers achieve better outcomes, particularly for “at-risk” students. Among other products, AM offers a suite of online tools to supplement middle and high school Algebra 1 curricula. These include lesson planning guides for teachers, computer applets and animations for students and tests and other assessments of the material. Schools individually choose to buy the AM program, and within those schools math teachers are free to use AM as little or as much as they choose. Importantly, the AM staff is able to observe the amount of time teachers and students at particular schools use the online AM resources.

The Agile Mind company contracted with the Educational and Well Being (EWB) division of the University of Michigan Institute for Social Research, to evaluate their Algebra 1 program; EWB, in turn, hired this report’s author to perform the statistical analysis. This evaluation acts as a motivation and an illustration of the methods proposed in this chapter. This chapter will model AM Algebra 1 usage as a school-level intervention, and attempt to estimate the effect of high AM Algebra 1 implementation on schools’ aggregated Algebra 1 end-of-course standardized test scores. To define “high implementation,” we will exploit the AM staff’s data on online AM usage and focus on estimating the effect of the online components of AM. A previous AM report (Correnti et al., 2008) was unable to detect any effects; it did, however, discover that many schools which had bought the AM curriculum did not contain any teachers who used it. In contradistinction,

this chapter, pursuant to the AM staff’s request, will consider “treated” AM schools those schools containing teachers who used the online AM tools for at least 70 hours over the course of a school year. We will also consider only schools in Indiana and Texas. Section 3.4 will provide a more complete description of the AM Algebra 1 program.

To conduct this causal estimation, we will use propensity score matching, using publicly-available school-level covariates, which will be described in Section 3.5. However, there are merely 38 high and middle schools in Indiana and Texas which use AM to this extent, and thousands of covariates, such as student demographics, several years of standardized test achievement measures, disaggregated into demographic groups, and school finances; hence, high dimensional causal inference techniques will be necessary.

The small number of AM schools poses an additional problem, aside from the difficulty of fitting a high-dimensional model: 38 schools may not provide enough power to detect a small effect. However, rather than dismissing the data offhand, it may be profitable to attempt the best analysis possible: is there information about AM to be gleaned from these data? Indeed, the multivariate techniques that this chapter will advocate may improve standard errors as well as biases, by using covariates to model some of the variance in the outcome. Even if the resulting standard errors are not small enough to detect a small effect, they may still yield informative confidence intervals, which will, in turn, limit the range of possible effects.

The following section will review the three principal techniques that form the basis of our method: propensity score matching, prognostic scores, and principal component analysis. Section 3.3 will introduce the novel statistical methods, focusing on principal components analysis as the main multivariate tool. Sections 3.4–3.6 will comprise the actual evaluation of AM: Section 3.4 will provide some background for the analysis, Section 3.5 will describe the data in depth, and Section 3.6 will carry out the analysis. Section 3.7 will conclude.

3.2 Propensity Scores, Prognostic Scores, and Principal Components: A Toolbox for Observational Causal Inference

This section will review three important methodological tools, each of which is useful in high-dimensional propensity-score causal inference. No novel results are present in this section.

3.2.1 Overview: Propensity-Score Matching

Recall the Rubin Causal Model (Section 1.1): each subject i has two potential outcomes, Y_{Ci} and Y_{Ti} : subject i 's outcomes if he is untreated or treated, respectively; a treatment assignment $Z_i \in \{0, 1\}$; a vector of covariates X_i measured before treatment (or unaffected by treatment). Also recall that when the treatment assignment mechanism is unknown, the following assumption, labeled in section 1.1 as (1.4) and called strong ignorability, aides causal inference:

$$\{Y_{Ci}, Y_{Ti}\} \perp\!\!\!\perp Z_i | X_i. \quad (3.1)$$

For subjects with identical values of X_i , potential outcomes are independent of treatment assignment, and effect estimates comparing treated and untreated subjects are unconfounded.

There are experiments in which strong ignorability (1.4) is known to hold: experiments in which, for reasons of feasibility or to improve estimates' precision, researchers stratify random assignment on a set of covariates. One approach to an observational study which assumes (1.4) is to model the data as if they emerged from such an experiment, which researchers must then reconstruct (see, eg. Rubin, 2008). That is, assume that within certain unknown groups of subjects, determined by measured covariates, treatment was assigned randomly. The challenge, then, is to use the observed relationships between X and Z to reconstruct those groups. This approach follows the reasoning in Section 1.1.4: that the best way to analyze an observational study is by mimicking a randomized experiment. The attempt to reconstruct this hypothetical experiment is the “design” stage of the study, which is where most of the interesting work takes place: establishing and justifying an identification strategy, and modeling the crucial relationship between X and Z . Outcome data and causal estimates first appear in the analysis stage, after the modeling is done.

Given (1.4), the most straightforward approach to estimating average treatment effects $\hat{\tau}$ is to match subjects exactly on all of the measured covariates X : that is, identify the level sets of X within which (1.3)—independence of potential outcomes and treatment assignment—is assumed to hold. However, with many covariates, this becomes difficult or impossible (Cochran, 1953).

Rosenbaum and Rubin (1983b) proposed a solution to the multivariate matching problem: the propensity score. Define the subject i 's propensity score, π_i as

$$\pi_i(X) = Pr(Z_i = 1 | X). \quad (3.2)$$

The propensity score is a dimension-reduction technique with an attractive property; conditioning on π —one dimension—is equivalent to conditioning on multi-dimensional X . That is, if strong

ignorability (1.4) holds, then

$$\{Y_C, Y_T\} \perp\!\!\!\perp Z \mid \pi(X). \quad (3.3)$$

In the presence of heterogeneous treatment effects—when $Y_T - Y_C$ is not constant—one may distinguish between the “average treatment effect,” $\mathbf{E}[Y_T - Y_C]$, and the average “effect of the treatment on the treated,” (ETT) or $\mathbf{E}[Y_T - Y_C \mid Z = 1]$. For unbiased estimation of the latter quantity, a somewhat weaker form of (3.3) will suffice: $Y_C \perp\!\!\!\perp Z \mid \pi(X)$; for the remainder of the paper we only assume this weaker form, and attempt to estimate the effect of the treatment on the treated.

There are two practical steps that researchers must take in order to reconstruct a hypothetical stratified randomized experiment using (3.3). Firstly, in observational studies, $\pi(X)$ is unknown, and must be estimated; secondly, in a finite sample exact propensity-score matches are generally impossible, so approximate matching must suffice. Despite these practical issues, approximate propensity-score-matching can often remove a large portion of the bias that would result from an unadjusted causal comparison. To check the plausibility of the experimental model, after estimating propensity scores and choosing a match, researchers can check the match for covariate balance. In particular, the definition of propensity scores implies, analogously to (3.3), that

$$X \perp\!\!\!\perp Z \mid \pi. \quad (3.4)$$

This suggests that testing covariate balance, stratified by propensity score matches, serves as a diagnostic test. When covariates are balanced within matched sets, the data resemble data from a stratified randomized experiment, which would also have this property.

Of course, to effectively match treated subjects to control subjects on the propensity score, treated and control groups must have overlapping propensity-score distributions, leading to a second assumption necessary for propensity-score analysis:

Assumption (Overlap).

$$\forall i \ 0 < \pi_i < 1. \quad (3.5)$$

There can be no subjects in the sample, with propensity scores of zero or one. If some such subjects exist, they may be removed from the analysis; matching with “calipers” will be discussed below.

While propensity scores are a powerful dimension reduction technique, very high dimension can still overwhelm them. Conventionally, propensity scores are estimated with a general linear model, such as logistic regression or probit regression (Rosenbaum, 2002a). For this reason, estimating propensity scores in a high-dimensional setting is itself a challenge; in particular, when there are more covariates than treated or untreated subjects, it may be impossible to incorporate every

covariate into a propensity score model.

Other Propensity Score Techniques

The approach to propensity score matching presented here—which is roughly equivalent to the approach favored by its discoverers, Rubin (2008) and Rosenbaum (2002c)—is not the only way propensity scores are used in the interdisciplinary causal inference literature. In particular, different academic disciplines tend to approach propensity score estimation differently. A thorough discussion of propensity score methods, including an assessment of their respective strengths and weaknesses, would be out of place here—some overviews are Heinze and Jüni (2011), Austin and Mamdani (2006) and Heinrich et al. (2010)—but a brief mention of two popular alternatives is in order.

The most straightforward use of propensity-scores would be to use them as weights in a weighted average. This approach builds on ideas from survey sampling: if some subjects are sampled with a higher probability than others, an unbiased estimate of population averages weights each subject by the inverse of his probability of selection (Horvitz and Thompson, 1952). Similarly, the average Y_T and Y_C values in a sample—and hence, the sample-average treatment effect—can be estimated unbiasedly, in theory, by computing weighted averages, with the reciprocals of the propensity scores serving as weights. This inverse-probability weighting approach is relatively popular in epidemiology, and has been developed extensively by Robins (e.g. Hernán and Robins, 2006) who has suggested combining propensity score weighting and linear regression in “doubly-robust” estimators (Bang and Robins, 2005). These strategies suffer from two weaknesses: first, they depend on correct estimation of propensity scores; in contrast, it is not hard to imagine how, in the method preferred here, propensity scores could be estimated incorrectly, yet produce a reasonable matched design that would lead to approximately unbiased estimates. Such a case would arise, for instance, if the estimated propensity scores were, in expectation, a monotonic transformation f of the true propensity scores, $\mathbf{E} \hat{\pi} = f(\pi)$. Secondly, these methods lack the advantages of models that seek to mimic randomized experiments: the objectivity that comes with a clear division between design and analysis, small-sample robustness and transparency.

Another set of methods, more common in economics, bears more similarity to the methods discussed here. Smith and Todd (2005) and Heckman et al. (1998) suggest kernel-weighted matching estimators, in which researchers match treated subjects to (possibly overlapping) groups of untreated subjects, weighted by their distances to the treated subjects in propensity scores. This can be expressed as a kernel-weighted local linear regression. This has the advantage, over the

method here, of accounting for differing levels of similarity between treated and untreated subjects within matched sets. Similar to inverse-probability weighting approaches, but unlike approaches that attempt to mimic randomized experiments, weighted-matching estimators may not be robust for small-samples, and lack the degree of transparency that accompanies the ability to enumerate the members of matched sets. See Section 1.1.4, above, for more details.

3.2.2 Overview: Principal Components Analysis

The problem of very high dimensional datasets calls for an additional dimension reduction technique. One of the most popular and oldest of these is principal components analysis. PCA can be thought of as the solution to a linear algebraic optimization problem. Let $X_{n \times p}$ be a (covariate) matrix, in which each column (covariate) has been studentized: transformed so that its sample mean is zero and its sample variance is one. Then the first principal component is Xu^1 , with u^1 defined as

$$u^1 = \arg \max_{u: u'u=1} u'X'Xu. \quad (3.6)$$

If X is a matrix of n samples of p covariates, each with mean 0, then $X'X$ is the sample covariance matrix of X . Similarly, $u'X'Xu$ is the variance of a linear combination of the covariates in X , with coefficients u . The first principal component Xu^1 represents the linear combination with the maximum variance. If variance is considered as a proxy for information, then Xu_i^1 is the linear combination of covariates providing the maximal information about subject i .

Similarly, higher principal component coefficients u^k are defined as

$$u^k = \arg \max_{u: u'u=1, u'u^i=0, i < k} u'X'Xu \quad (3.7)$$

This is the linear combination of variables in X with maximal variance with coefficients orthogonal to the coefficients in all the previous principal components. For the right choice of K , information from a high-dimensional dataset can often be effectively summarized in the first K principal components.

Expressions (3.6) and (3.7) turn out to be eigen-value problems. That is, $u^1, \dots, u^{\max\{n,p\}}$ are the eigenvectors of $X'X$ ordered according to the magnitude of their corresponding eigenvalues. Then the objective equation $u'X'Xu = \lambda^2 u'u = \lambda^2$.

In high-dimensional datasets, and, in particular, when $p \gg n$, a computationally-efficient method of solving equations (3.6) and (3.7) is the singular-value decomposition (SVD) (Mandel,

1982): any matrix $X_{n \times p}$, with rank r , can be decomposed as

$$X' = USV' \tag{3.8}$$

where $U_{n \times r}$ and $V_{p \times r}$ are orthogonal, and $S_{r \times r}$ is diagonal. Following (3.8),

$$\begin{aligned} X'X &= USV'VSU' \\ &= US^2U'. \end{aligned}$$

Then (3.6) becomes

$$u'X'Xu = u'US^2U'u$$

with U orthogonal.

As mentioned, the vectors u^k are eigenvectors of $X'X = US^2U'$. It turns out that the eigenvectors u^k are the columns of U , U_1, \dots, U_n . To see this, consider,

$$\begin{aligned} US^2U'U_k &= Us_k \text{ (where } s_k \text{ is a vector with } S_{k,k}^2 \text{ in the } k^{\text{th}} \text{ place and 0s elsewhere)} \\ &= U_k S_{k,k}^2. \end{aligned}$$

Therefore, the principal components u^k are actually the columns of U , U_k , and the corresponding eigenvalues are the squared ‘‘singular values,’’ the diagonal entries of S .

The singular values themselves contain useful information about the structure of the dataset. If $X'X$ is the empirical covariance matrix of the set of covariates in X , then the diagonal entries of $X'X$ are the sample variances of its component covariates. The sum of these variances is the trace of $X'X$, which is also equal to the sum of its eigenvalues, the squared singular values S^2 . Each squared singular value, divided by the sum of S^2 , can be thought of as the proportion of X 's total variance that can be attributed to that singular value's eigenvector. Often, a large proportion of a matrix's variance can be attributed to a relatively small number of principal components: this is the case, for instance, if the variables in X are highly correlated with each other. In such a case, one compound variable can capture much of the information in several separate highly collinear raw variables. By plotting the singular values in order, an analyst can check to see if this is the case: if it is, the singular values will decrease rapidly before leveling out near zero.

In such situations, PCA can be a useful dimension-reduction technique. By focusing on the first K principal components and discarding the rest, an analyst can capture a large proportion of the information in a dataset in few dimensions. This has implications for causal inference in

high-dimensional data: if the covariate matrix X could be effectively summarized in a much smaller matrix of principal components, the problem of high-dimensional data would be mostly solved. How to choose which principal components to use, and how to incorporate them into a matching scheme, are not trivial questions; prognostic scoring, an analogue to propensity scoring, may be an important part of the solution.

3.2.3 Overview: Prognostic Scores and Principal Components Regression

Strong ignorability asserts that, conditional on X , treatment assignment Z and potential outcomes $\{Y_C, Y_T\}$ are independent. In order for a potential confounder W to cause a violation of strong ignorability, and to induce a dependence between Z and $\{Y_C, Y_T\}$, W must share a dependence with both Z and $\{Y_C, Y_T\}$.

Propensity scores are attempts to model the first of these possible dependences, between covariates X and Z , and, indeed, under strong ignorability, at level sets of propensity scores, $\{Y_C, Y_T\}$ and Z are independent, as in equation (3.3).

Alternatively, analysts can model the possible relationship between X and Y_C , and estimate “prognostic scores” (Hansen, 2008b).¹ In the same sense that propensity scores capture the relationship between covariates X and treatment assignment Z , prognostic scores $\psi(X)$ capture the relationship between X and control potential outcomes Y_C . More precisely, $\psi(X)$ is a prognostic score if

$$Y_C \perp\!\!\!\perp X \mid \psi(X) \tag{3.9}$$

that is, $\psi(X)$ captures all of the information X holds about Y_C ; it is, in this sense, a sufficient statistic. Conceptually, $\psi(X)$ can be thought of as a prediction of Y_C given X .

In practice, analysts can estimate $\psi(X)$ by modeling Y_C as a function of X , often using linear regression. In a case in which the number of covariates p is much greater than the sample size n , ordinary least squares (OLS) is impossible; fortunately, there is a range of options for predicting a continuous response as a function of covariates when $p \gg n$. One of these methods is principal components regression.

After computing the principal components of X , an analyst can construct a matrix of k principal components, U_k , and treat them as regressors, in the regression model

$$Y_C = \alpha + U_k \beta_{PCR} + \epsilon_C \tag{3.10}$$

¹As mentioned above, we will focus on confounding with Y_C , and not Y_T , and attempt to estimate the ETT.

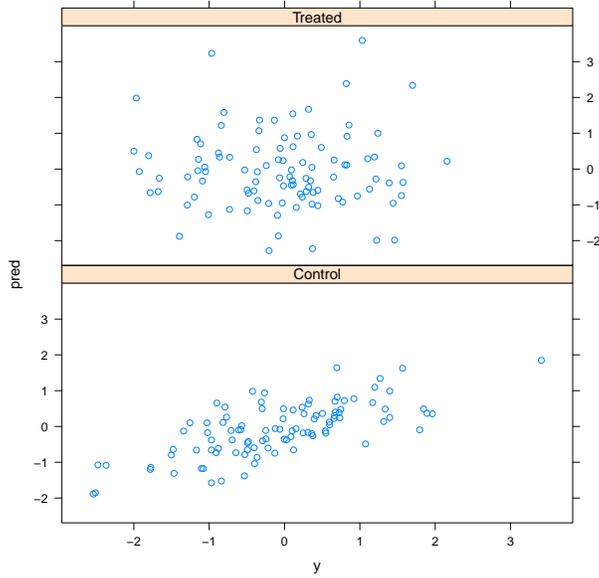


Figure 3.1 A possible pitfall of prognostic-score matching: If prognostic model (3.10) is fit to the control sample but not to the treated sample, overfitting can cause spurious prognostic-score differences between treated and untreated subjects. Here, 100 “treated” and 100 “untreated” observations Y were drawn from the same standard normal distribution, as were 50 “covariates,” which were, in fact, independent of both Z and Y . An OLS model $Y \sim X$ was fit to the control set. The fitted values in the control set were plotted against the model’s predictions in the treated set: in some regions, significant differences in prognostic scores between treated and untreated observations appear, even though all observations were independent and identically distributed. Overfitting is to blame.

where α is a fixed intercept term and ε_C is a vector of mean-0 errors, observed under the control condition. Predictions from this regression, \hat{Y}_C , can serve as estimates of prognostic scores $\psi(X)$.² If \hat{Y}_C is balanced, within matched sets, then it is unlikely that variables in X are confounding the causal estimate: the values of Y_C one would have predicted, based on X , are approximately equal in treated and untreated subjects.

One potential pitfall of prognostic score-matching—analyzed in Hansen (2008b)—is avoidable in this context. If model (3.10) is fit to the entire dataset, including both treated and untreated subjects, then the fitted values are not necessarily predictive of Y_C . When treated subjects are included in the model fit, the fitted values can depend on treatment assignment; therefore, conditioning on them can bias the causal estimate.

Alternatively, a statistician can fit model (3.10) to only the untreated sample. However, this

²Unlike propensity scores, prognostic scores may be multidimensional (Hosman, 2011); here, however, we will consider one-dimensional estimates of $\psi(X)$.

leads to another, more subtle problem: due to inevitable overfitting, the model will fit the data in the untreated sample more closely than in the treated sample. This is illustrated in a toy example in Figure 3.1. In the example, there are 100 treated and 100 untreated subjects. Each subject has a vector x_i of $p = 50$ covariates, which are mutually independent and generated from a standard normal distribution; together, they form matrix X . The outcome Y is also generated from a standard normal distribution, and independent of both X and treatment assignment. The upper plot in Figure 3.1 plots Y against prognostic scores \hat{Y} : fitted values from regressing Y on X . Even though Y is independent of X , the least squares algorithm produced a fit with an impressive within-sample prediction: the coefficient of determination $R^2 = 0.51$. However, this is not due to any inherent pattern in the data, but is the result of overfitting. This can be seen by inspecting the bottom plot of Figure 3.1 which plots prognostic scores—predictions from the same model—in the treated group against actual values of Y —there is no apparent relationship, and $R^2 = 0.0002$. While overall the fitted values have the same population mean in the treated and control groups—zero—if one were to compare the fitted values between treated and control subjects in some subsets of the data spurious treatment-control differences would appear. Even though the data from the treatment and control groups were drawn from the same distribution, prognostic score overfitting here would lead an analyst to reject a comparison of the groups.

One way to avoid this potential pitfall is to set aside a portion of the dataset to fit the prognostic model, and use the remainder of the dataset to estimate treatment effects. In that case, rather than a comparison of within-sample fitted values in the control group to out-of-sample predictions in the treated group, the comparison would be between out-of-sample predictions in both groups. In general, this approach is unappealing because discarding data may reduce the power of a hypothesis test and widen confidence intervals. However, the following section will show that when combining prognostic score modeling with propensity score matching, a subset of the control group that can be used to fit prognostic models emerges naturally.

3.3 Combining Principal Components with Propensity Scores: Matching, Evaluating and Estimating

A propensity-score analysis involves three stages: estimating and matching on the propensity scores, evaluating the match and analyzing the outcome. When $p \gg n$, PCA has a role to play in each of those steps: principal components can be included as variables in a propensity-score model, a prognostic model based on principal components regression can be used to test the success of

the propensity-score match, and principal components regression can be used to adjust final effect estimates.

3.3.1 Estimating Propensity Scores

We will focus on propensity-score estimation based on logistic regression. Let $\text{logit}(Z)$ be the log-odds a subject is assigned to treatment, and let W_1, \dots, W_k be potential pre-treatment predictors of treatment assignment. Then we fit the model

$$\text{logit}(Z) = \alpha_P + \beta_{P1}W_1 + \dots + \beta_{Pk}W_k = W\beta_P \quad (3.11)$$

where α_P and β_P are parameters of the propensity model; without subscripts, W should be interpreted as an $n \times (k + 1)$, with the first column a vector of ones, and β_P a vector of coefficients $\{\alpha_P, \beta_{P1}, \dots, \beta_{Pk}\}$. The vector β_P is estimated as $\hat{\beta}_P$, yielding fitted values, or linear predictors

$$\hat{p}_{lin} = W\hat{\beta}_P. \quad (3.12)$$

When the number of predictors k meets or exceeds the number of treated or untreated subjects, a classical logistic model will find a separating hyperplane—i.e. a vector of estimated coefficients $\hat{\beta}_P$, such that there is no overlap between the linear predictors \hat{p}_{lin} for the treatment and control groups (Hastie et al., 2005). Such a model predicts treatment status perfectly, but only within the sample used to fit the model. This only seemingly violates the overlap assumption (3.5) and is really a symptom of overfitting, not (necessarily) a property of the underlying covariate distribution: even a set of n randomly-generated covariates, independent of Z , would yield a perfect model.

A bayesian approach to logistic regression relaxes this property somewhat (Gelman et al., 2008); if weakly-informative priors on β encourage $\hat{\beta}$ to fall close to zero, it is harder to generate a separating hyperplane. With very small treatment or control groups, using bayesian logistic regression allows some extra leeway (Kleyman, 2009).

To estimate propensity scores when the total covariate matrix X has very high dimension, one could use substantive knowledge to choose a set of variables W —whose size depends on the sample size—which have the largest potential to confound a causal analysis. In the AM example, as following sections will discuss in detail, one could use student demographics, pretreatment exam passing rates, and, perhaps, school characteristics such as the size of the school. These variables, socioeconomic status, ethnicity, and prior performance, are the most likely confounders in an

educational causal analysis.

But what of the rest of the variables? PCA constructs small-dimensional summaries of the matrix of covariates: including one or a few of these summaries in the propensity score model would allow some information from the covariates that were not already included in the match to play a role. The strategy, then, would be to include some principal components as columns in the matrix W in the propensity-score model (3.11). The number of principal components may be chosen based on the constraints of the data: more components are preferable, so long as model (3.11) converges and does not produce degenerate fitted values. Further, analysts can choose the number of components by inspecting the properties of various fitted models and the matches that result from them, the same criteria by which other propensity model specification choices are judged.

But which principal components should be included? The goal, of course, is to satisfy (3.3): to assure that, within level sets of $\hat{\pi}$, treatment assignment is independent of potential outcomes. Equation (3.3) will hold if π predicts either Z or the potential outcomes Y_C and Y_T ; it follows that the most important components to include are those that most highly correlate with the observed outcomes (see Steiner et al., 2010). This, however, raises a dilemma: conditioning a causal estimate on covariates that may, themselves, be effects of treatment assignment can induce “included variable bias” (Rosenbaum, 1984; Ayres, 2005); in other words, all adjustment covariates must be pre-treatment. So how does one determine which principal components most highly correlate with potential outcomes? Estimating the correlation between each of the principal components and all of the measured outcomes would induce included variable bias. On the other hand, if an analyst estimates correlations between the principal components and Y_C scores—that is, using only data from the control group—the resulting bias is likely to be small or non-existent.³ An analyst can include information from the entire covariate matrix X in the propensity-score model by including the few principal components that correlate, in the control group, most highly with Y .

³The bias that results from this maneuver may not be exactly zero; although the treatment itself did not affect the outcomes in this calculation, treatment *assignment* determined which outcomes to include in the calculation. However, since the choice of which components to use is a discrete choice, it is likely that, under alternative treatment assignments, the choice would remain the same. Even if that is not the case in a particular analysis, and a component is included in the logit model (3.11) that would not have been included under alternative treatment assignments, that component would have to substantially influence the propensity-score estimation in order to substantially bias the outcome. This procedure, therefore, seems more likely to reduce bias—by including information from more, and more important, variables in the propensity-score estimation—than to increase it.

3.3.2 Evaluating the Match

The ignorability assumption mathematically implies equation (3.3); however, a practical propensity score analysis is less straightforward. In virtually every propensity score study, the propensity scores π themselves must be estimated as $\hat{\pi}$. Furthermore, in a finite sample, no treated or untreated subjects will have precisely the same estimated propensity score, so conditioning on $\hat{\pi}$ must be approximate. Specifically, in a propensity-score matching design, a statistician assumes that

Assumption (Matched Ignorability). Subjects i and j are in the same matched set implies that $\pi_i = \pi_j$, so that

$$Y_C \perp\!\!\!\perp Z | m(\hat{\pi}) \quad (3.13)$$

where $m(\hat{\pi})$ represents the matched sets based on estimated propensity scores $\hat{\pi}$.

One compelling approach to assessing the accuracy of these two approximations is to check that equation (3.4) holds. In particular, an analyst can test whether, within matched sets, the means of pre-treatment covariates are equal between the untreated and treated subjects (Rosenbaum and Rubin, 1984). The logic behind this approach is that if there is evidence for imbalance—the means of a covariate are unequal between treated and untreated observations in a matched set—then matched ignorability is unlikely to hold, especially if the imbalanced covariates correlate with Y_C or Y_T . Conversely, if covariates are balanced, then the most straightforward confounding, in which monotonic relationships between Z and X , and X and Y_C or Y_T bias the causal estimate, is impossible (however, more subtle types of confounding, involving non-monotonic relationships or interactions, are still possible).

When the observed covariate matrix X is high dimensional, there is an additional layer of separation between (3.3) and an applied analysis: only a subset of available pre-treatment covariates contributed to the match. A strong analysis, however, would attempt to gauge the plausibility of bias resulting from this omission.

After matching on a propensity score $\hat{\pi}$, an analyst can use estimated prognostic scores \hat{Y}_C to check if the matching scheme effectively balanced not just the subset of variables used in the propensity score model (3.11), but the entire covariate matrix X . In other words, he can fit a prognostic model to the entire set of pre-treatment covariates X and the outcome Y , and check if there is imbalance between the predictions of the prognostic model in the control and treatment groups. Recall that prognostic modeling can suffer from a pitfall if the model is fit using data from either the whole sample or just the control sample. In particular, if the prognostic model is fit using data from the control group, then overfitting can cause spurious differences to appear between the groups, even if no actual difference exists—this is illustrated in Figure 3.1.

In this case, however, a solution to the overfitting problem may emerge naturally from the data setup. Often, a propensity match will involve the use of calipers: only control subjects with propensity scores within a pre-specified distance from treated subjects are included as candidates for a match. Calipers prevent a matching routine from matching two subjects who are wildly different, in an over-zealous attempt to use the entire dataset in the estimating procedure; it sacrifices the power that would come from using the entire dataset to mitigate the bias that would result from unrealistic matches. The use of calipers in constructing propensity-score matches causes a portion of the data—often a sizable number of control subjects—to be discarded from the analysis, and hence solves the overfitting problem that potentially plagues prognostic scores. Furthermore, marginal returns to power and efficiency diminish as match sizes grow; in particular, the gains in efficiency from adding more untreated subjects to a matched set that already includes five or so untreated subjects are modest. For the sake of simplicity, as well as for other reasons, researchers may want to limit the sizes of their matched sets; this adds more subjects to the pool of discarded controls. If a prognostic score model is fit to the discarded portion of the dataset, then its predictions suffer from overfitting in neither the treatment group nor the matched-controls.

In other words, the propensity score matching routine with calipers divides the dataset into three groups: the treatment group, the control subjects that are matched to treatment subjects, and the control subjects which are not included in the match due to their dissimilarity with the treated subjects or their superfluousness. The prognostic model would be fit in this third group. A balance test could then compare the model’s predictions in the treated group to its predictions in the matched-control group; both of these sets of predictions would be “out of box”—that is, using data that was not involved in the fitting procedure—so overfitting would not be present.

If a prediction algorithm based on pre-treatment covariates predicts no difference in Y_C values between the control and treatment groups, the potential bias that may result from these covariates must be severely limited in scope. Of course, this depends on the quality of the prognostic fit, which may be measured in mean-squared-error (MSE), $\mathbf{E}(Y_C - \hat{Y}_C)^2$. This can be estimated with cross-validation within the unmatched control subjects. However, it may be the case that MSE estimates from cross-validation are biased downwards, because the matched controls are not representative of the entire control population. Nevertheless, balanced prognostic scores present valuable evidence against confounding from any of the measured covariates. This issue is discussed more fully in the following chapter.

When $p \gg n$, prognostic scores become particularly valuable in the evaluation of a propensity-score match. Since the propensity score model (3.11) can only include a small subset of measured covariates, there is a potential for substantial confounding from the rest of the covariates not included

in W . However, prognostic score estimation is not as severely limited: prognostic scores may be estimated using high-dimensional prediction techniques such as principal components regression.⁴ Then, principal-components regression and prognostic scores strengthen the believability of a propensity-score design in a high-dimensional dataset.

3.3.3 Estimating Effects with Propensity and Prognostic Scores: A Peters-Belson Approach

Even if imbalance in prognostic scores between treatment and control subjects is statistically undetectable, it will generally be non-zero. In this case, an additional bias-correcting step may be useful. Following Peters (1941) and Belson (1956), decompose each observed outcome Y_i into the portion predicted by the prognostic model and prediction error:

$$Y_C = \hat{Y}_C + e_C \quad (3.14)$$

and

$$Y_T = \hat{Y}_C + e_T = \hat{Y}_C + e_C + \tau \quad (3.15)$$

where e_C and e_T represent prediction errors and τ is a vector of possible treatment effects. Prognostic predictions \hat{Y}_C are pre-treatment, in the sense that they are functions only of pre-treatment covariates and outcomes from un-treated subjects not involved in the assumed pseudo-experiment. That being the case, treatment assignment could not affect \hat{Y}_C ; any effect of treatment must be entirely present in the difference between prediction errors e_C and e_T . Hence, we may write, as in (3.15), $\tau = e_T - e_C$. Therefore, causal estimation may be based only on the values e_C ; in other words, instead of treating Y as the dependent variable, researchers may treat $e = Ze_T + (1 - Z)e_C$ as the dependent variable.

Replacing Y with e in the outcome analysis may reduce both the bias and the variance of a resulting estimate. Indeed, doing so replaces the ignorability-after-matching assumption (3.13) with the following:

Assumption (Adjusted Ignorability).

$$e_C = Y_C - \hat{Y}_C \perp\!\!\!\perp Z | m(\hat{\pi}) \quad (3.16)$$

⁴More modern techniques, such as the LASSO, may perform as well or better than PCR in some cases; this possibility will be discussed in Chapter 4.

In other words, the potential prediction residuals, not the actual potential outcomes, must be independent of Z . This assumption is somewhat weaker than (3.13), in a sense described below.

The Peters-Belson-propensity ETT estimate is

$$\hat{\tau}_{PBP} = \sum_m w_m \left(\frac{e'_m Z_m}{1' Z_m} - \frac{e'_m (1 - Z_m)}{1' (1 - Z_m)} \right) \quad (3.17)$$

where $m = 1, \dots, M$ indexes the matched sets, e_m and Z_m are vectors of prognostic residuals and treatment status for subjects in set m , and 1 is a vector of 1s. Strata weights w_m may be chosen flexibly, in such a way that they sum to one; when $w_m = n_m/n$, with n_m the size of matched-set m , then $\hat{\tau}_{PBP}$ reduces to a simple difference in means between treated and untreated subjects (in such cases, the impact of the propensity score matching on the estimate would come from selecting a subset of controls from a larger control pool that are similar to the treated subjects). Generally, $\hat{\tau}_{PBP}$ is an average of the average differences between treatment and control subjects within each matched set.

For the following, it will be convenient to replace (3.10) with a more general form,

$$Y_C = f(X) + \varepsilon_C \quad (3.18)$$

for some function f .

Under strong ignorability, in addition to certain known, though perhaps idealized, conditions, conducting analysis on prediction residuals rather than on outcomes—that is, replacing Y with e —leads to an attractive property, which we will call “double unbiasedness.” This condition states that if either the prognostic model or the propensity-score match is correct, then the resulting estimate of the average effect of the treatment on the treated will be unbiased. Double-unbiasedness is analogous to “double-robustness” (Bang and Robins, 2005; Neugebauer and van der Laan, 2005), the property of some estimators that combine an outcome regression with propensity scores: if either an outcome model or a propensity-score model is valid, the resulting estimate will be consistent. As its name suggests, double-unbiasedness suggests conditions under which $\hat{\tau}_{PBP}$ will be unbiased; it does not refer to consistency.

Double unbiasedness holds for Peters-Belson propensity modeling if the method for estimating $f(X)$ in (3.18) is unbiased. It is also necessary to choose w_m in such a way so that $w_m = Pr(M = m)$, the probability of an observation falling in matched-set m . This is dependent on the sampling and conditioning model chosen by the researcher; in general, a reasonable choice is $w_m = n_m/n$, so $w_m = Pr(M = m)$ is the proportion of all the subjects in set m .

Proposition 2. *If strong ignorability (1.4) holds, and if, in addition, either*

1. $Y_C = f(X) + \varepsilon_C$, with $\{Y_C, X\} \perp\!\!\!\perp \varepsilon_C$ (the prognostic model is correct) or
2. $i, j \in m \Rightarrow \pi_i = \pi_j$ (the propensity-score match is correct)

and $f(X)$ is estimated as $\hat{Y} = \hat{f}(X)$, from a separate, independent sample, with $\mathbf{E}[\hat{f}(X)|X] = f(X)$, then

$$\mathbf{E} \hat{\tau}_{PBP} = \tau_{ETT} \quad (3.19)$$

where τ_{ETT} is the ETT, $\mathbf{E}[Y_T - Y_C|Z = 1]$.

Proof. Assume the prognostic model is correct. Then we will show first that $\mathbf{E}[e_C|Z] = \mathbf{E}e_C$, i.e. that e_C is mean-independent of Z , and then that $\hat{\tau}_{PBP}$ is unbiased for the ETT. Since \hat{f} was estimated unbiasedly from a separate, independent sample, $\mathbf{E}[e_C|X] = \mathbf{E}e_C$:

$$\begin{aligned} \mathbf{E}[e_C|X] &= \mathbf{E}[\hat{f}(X) - f(X)|X] + \mathbf{E}[\varepsilon|X] \\ &= \mathbf{E}[\hat{f}(X) - f(X)|X] + \mathbf{E}e_C = \mathbf{E}e_C \end{aligned}$$

Due to ignorability (1.4), $e_C \perp\!\!\!\perp Z|X$ (since $X \perp\!\!\!\perp Z|X$ trivially and e_C is a function of Y_C and X). Then

$$\mathbf{E}[e_C|Z] = \mathbf{E}\mathbf{E}[e_C|Z, X] = \mathbf{E}[e_C|X] = \mathbf{E}[e_C]$$

So e_C is mean-independent of Z . The prediction error e_T is observed for the treatment group, for which $Z = 1$. For subject i let $e_{Ti} = e_{Ci} + \tau_i$. The treatment effect τ_i can be negative, positive or zero, and can vary between subjects. Let constants $n_{Tm} = 1'Z$ and $n_{Cm} = 1'(1 - Z)$ be the numbers of subjects in the treatment and control groups of each matched set m . Finally, denote by e_{Cm} and

τ_m the vector of e_{Ci} and τ_{Ci} in each matched set m . Then

$$\begin{aligned}
\mathbf{E} \hat{\tau}_{PBP} &= \mathbf{E} \sum w_m \left(\frac{(e_{Cm} + \tau_m)'Z}{n_{Tm}} - \frac{e'_{Cm}(1-Z)}{n_{Cm}} \right) \\
&= \sum w_m \left(\frac{\mathbf{E}(e_{Cm} + \tau_m)'Z}{n_{Tm}} - \frac{\mathbf{E}e'_{Cm}(1-Z)}{n_{Cm}} \right) \\
&= \sum w_m (\mathbf{E}[e_C|Z=1, m] - \mathbf{E}[e_C|Z=0, m] + \mathbf{E}[\tau|Z=1, m]) \\
&= \mathbf{E}_m (\mathbf{E}[e_C|Z=1, m] - \mathbf{E}[e_C|Z=0, m] + \mathbf{E}[\tau|Z=1, m]) \text{ for correctly chosen } w_m \\
&= \mathbf{E}[e_C|Z=1] - \mathbf{E}[e_C|Z=0] + \mathbf{E}[\tau|Z=1] \text{ by smoothing} \\
&= \mathbf{E}e_C - \mathbf{E}e_C + \mathbf{E}[\tau|Z=1] \text{ since } e_C \text{ is mean-independent of } Z \\
&= \mathbf{E}[\tau|Z=1]
\end{aligned}$$

Which is the ETT. The other direction is simpler: if $i, j \in m \Rightarrow \pi_i = \pi_j$, then $e_C \perp\!\!\!\perp Z|m(\hat{\pi})$ as well, since $e_C = Y_C - \hat{Y}_C$ and $\hat{Y}_C \perp\!\!\!\perp Z|m(\hat{\pi})$ since \hat{Y}_C is a function of independent Y_C values and X , which is independent of Z given π by (3.4). \square

Proposition 2 motivates the combination of Peters-Belson adjustment and propensity-score matching; however, generally, neither the propensity-score matches nor the prognostic model will be entirely correct. Indeed, when $p \gg n$, unbiased estimation of $f(X)$ in (3.18) may be impossible. However, Proposition 2 may indicate that if either model is somewhat close to the truth, then $\hat{\tau}_{PBP}$ will be approximately unbiased.

Indeed, this intuition may be made more precise. As a first step, it is possible to construct a bound on the squared bias of a Peters-Belson estimate, given the prognostic model's prediction mean-squared-error (MSE) $\mathbf{E}(Y_C - \hat{Y}_C)^2$. The intuition behind this approach is that the more accurately the prognostic model predicts Y_C values, the more accurate, and hence unbiased, the resulting ETT estimate will be. Without propensity-score stratification, a Peters-Belson ETT estimate is

$$\tau_{PB} = \bar{e}_T - \bar{e}_C = \bar{\tau}_{Z=1} + \bar{e}_{CZ=1} - \bar{e}_{CZ=0}$$

where $\bar{e}_{CZ=0}$ is the sample mean of the prognostic model's error in predicting Y_C among subjects for whom $Z=0$; other sample means above are defined analogously. Since $\bar{\tau}_{Z=1}$ is unbiased for the ETT, the bias of τ_{PB} is the expected difference between average prognostic estimation error in the

treatment group and in the control group. If we assume that $\mathbf{E}[\bar{e}_{CZ=1}\bar{e}_{CZ=0}] \geq 0$, then

$$\begin{aligned}
bias(\tau_{PB})^2 &= (\mathbf{E}(\bar{e}_{CZ=1} - \bar{e}_{CZ=0}))^2 \\
&\leq \mathbf{E}(\bar{e}_{CZ=1} - \bar{e}_{CZ=0})^2 \text{ by Jensen's inequality} \\
&= \mathbf{E}[\bar{e}_{CZ=1}^2 + \bar{e}_{CZ=0}^2 - 2\bar{e}_{CZ=1}\bar{e}_{CZ=0}] \\
&\leq \mathbf{E}[\bar{e}_{CZ=1}^2] + \mathbf{E}[\bar{e}_{CZ=0}^2] \\
&\leq \mathbf{E}[e_C^2|Z=1] + \mathbf{E}[e_C^2|Z=0] \\
&= \mathbf{E}[(Y_C - \hat{Y}_C)^2|Z=1] + \mathbf{E}[(Y_C - \hat{Y}_C)^2|Z=0] \tag{3.20}
\end{aligned}$$

which is the sum of the MSE for the treated subjects and the MSE for the control group. As noted, this does not account for propensity-score stratification, and may be considerably tightened. However, it suggests that for prognostic models with low prediction MSE, the bias of $\hat{\tau}_{PBP}$ may be small, even if neither the prognostic nor propensity models is entirely correct.

Peters-Belson adjustment may also improve the estimated ETT's standard error, as well as its bias. In many applications, a large proportion of the variance in Y_C can be explained by covariates X . In such cases, the residuals e from the prognostic models may be more stable than the outcomes themselves, so treating e as the outcome will lead to higher power. This is, at least informally, testable: using cross-validation, researchers can estimate the prediction R^2 of a prognostic model in the set of unmatched schools. A high cross-validated R^2 indicates substantial standard-error improvements from Peters-Belson covariance adjustment. This parallels the covariance adjustment suggested in Rosenbaum (2002c) and in Hansen and Bowers (2009).

3.4 Evaluating Agile Mind: Background

3.4.1 The Agile Mind Algebra 1 Program

Agile Mind, Inc. (AM) is a company based in Austin, TX, whose stated goal “is to provide the programs, the tools, and the instructional improvement systems you need to transform student engagement and achievement through exemplary, sustainable teaching.”⁵ In collaboration with education researchers at the Charles A. Dana center at the University of Texas, they have created sets of tools for nine courses: Biology, Middle School Math, Algebra 1, Intensified Algebra, Algebra

⁵ Agile Mind website: <http://www.agilemind.com>. Accessed: 1 July, 2013.

2, Geometry, Pre-calculus, Advanced Placement Calculus and Advance Placement Statistics. In addition, they provide a program called “Academic Youth Development,” which is not tied to any particular school course. These services are meant to supplement, not replace, the standard curriculum in Texas and Indiana schools. That said, they are designed to cover the state standards in these states; however, since the program is identical in the two states, the services cannot be precisely aligned with state standards.

The details of the AM Algebra 1 program have been thoroughly described in a report by Correnti et al. (2008), and on the AM website;⁵ in addition, there are various other sources of information currently available on the internet, including a demonstration video,⁶ an Institute of Education Sciences “Small Business Innovation Research Success Stories” press release⁷ and a qualitative evaluation from the Indiana Department of Education “Indiana Education Roundtable” along with the Charles A. Dana Center at the University of Texas, Austin.⁸ Like Correnti et al. (2008), this report will focus on the AM Algebra 1 program.

The AM programs have roughly two parts: a set of online “high-tech” tools, and face-to-face “high-touch” services. The “high-touch” tools include a two-day seminar for teachers beginning to use AM, seminars for administrators and district leaders, customized individual or school-based mentoring, and an AM guidebook.

The online tools attempt to teach students to “use algebraic tools to represent problem situations,”⁵ understand functions and rate of change, and model problems with linear, quadratic and exponential functions. The set of tools includes course content and assessments for students as well as planning tools and professional development for teachers. The course content is composed of “animations, simulations, in-depth exploration and practice.”⁵ The assessments include multiple-choice options, to prepare students for standardized tests, and are graded in real time, providing students with immediate feedback and teachers with guidance on student understanding. The online planning tools for teachers include “customizable lesson plans, teaching tips, and research-based strategies for improving student performance”⁵ which are aligned with state standards.

In particular, teachers have access to topic overviews and animations or applets to use in group instruction, topic summaries, guided assessments, self tests and multiple choice tests for assessments and planning tools (Correnti et al., 2008). The planning tools include help topics called “planning the course,” “scope and sequence,” “goals and objectives,” “topics at a glance,” “prerequisite skills” and “language support,” along with activity sheets.⁶ According to Correnti et al. (2008), teachers

⁶<http://http://www.youtube.com/watch?v=FN8VjMmuOo>. Accessed: 1 July, 2013

⁷<http://ies.ed.gov/sbir/agile.asp>. Accessed: 1 July, 2013

⁸<http://www.doe.in.gov/sites/default/files/curriculum/agilemindalgebra1.pdf>, Accessed: 1 July, 2013

tend to spend almost three quarters of their AM time on the overviews and animations—the direct instructional resources—and most of the remainder on assessment resources, though the proportion of time spent on assessment tends to increase with each additional school year teachers spend using AM. Students, on the other hand, spent most of their time on assessment functions, and the remainder on the animations and applets (students, of course, do not have access to teacher planning functions). Correnti et al. (2008) does not present any evidence that within an AM classroom, certain types of students (for instance, low- or high-performing students) used AM more than others.

3.4.2 Defining The Agile Mind Algebra 1 Treatment

The definition of AM implementation—what it means to be “treated”—was determined by the AM staff, who requested the analysis, and this was the treatment information that they provided. Correnti et al. (2008) found a high level of variability in teachers’ use of AM Algebra 1 services; in particular, that report divided Texas schools into three categories, high, medium and low, based on a teacher-level score composed of four factors: “the count of number of curricular topics teachers used, the average number of minutes teachers spent per curriculum topic, the logarithmic transformation of the total number of minutes teachers used AM services of all kinds, and the maximum percent of time teachers spent on a single curricular topic” (Correnti et al., 2008). Each school’s implementation index was aggregated from its teachers’ scores. First year AM teachers in low-implementing schools used online AM functions for about three hours over the course of a school year on average; first year AM teachers in high-implementing schools used these functions for about twenty-five hours. Correnti et al. (2008) attempted to measure the effect of AM Algebra 1 usage using “high implementation” classification as its treatment measure, and failed to detect an effect.

Perhaps in reaction to this result, the AM staff decided to commission an evaluation using a somewhat more stringent measure of usage: treated schools were those that employed a teacher who used AM online tools for at least 70 hours over the course of an academic year. AM provided the research team with a list of schools in Texas and Indiana which met this criterion, and requested that this list be considered the list of “treated” schools. To be sure, there are some statistical and substantive issues with this definition, mostly because the motivation for the choice of 70 hours as a cutoff is unclear. Nevertheless, these were the data that the research team had at its disposal, reflecting the wishes of the AM corporation which commissioned the study; it is with this definition that this report will proceed.

3.4.3 Selection into the Agile Mind Treatment

There are two levels of selection involved in our AM analysis: some schools buy the AM curriculum, and some teachers within those schools tend to use it. For the purposes of this study, we conflate those two into a binary variable that divides schools into those which use AM extensively and those which do not. Extensive use here means that a school contains a teacher who uses AM for at least 70 hours over the course of a school year. This selection mechanism, then, seems to hinge more on individual teachers' teaching decisions than on administrative decisions, since a small percentage of schools which have bought AM are classified as extensive users.

Unfortunately, we do not have direct information as to how those decisions are made; such data would be difficult to collect and even more difficult to interpret.⁹ However, certain aspects of the design of the AM program suggest some of the factors in its uptake. One of the principal aims of the AM program is educational “equity”;⁵ this suggests an emphasis on under-served student populations. Indeed, the AM website trumpets its emergence from research on under-served students:

The design of the Agile Mind AP support system builds on the work of Philip Uri Treisman, professor of mathematics and of public affairs at The University of Texas at Austin and executive director of the Charles A. Dana Center. In particular, it draws on his early research at the University of California at Berkeley investigating the factors that support minority student high achievement in calculus.

The website further states “Our work builds on Dana Center studies of high-performing, high-poverty schools.”⁵ These stated aims and design factors suggest that school level variables, such as ethnic makeup and socioeconomic status, could be important predictors of school-level selection into treatment.

To be sure, it is likely that there are unmeasured, or unmeasurable predictors of treatment assignment that also predict potential outcomes, that is, confounders; this, of course, is the principal disadvantage of observational studies over randomized experiments. We will attempt to account for this possibility, as part of our uncertainty assessment, in a sensitivity analysis below.

⁹In addition, I am told, interviewing teachers would require an amendment to the Institutional Review Board agreement, which I am unable to obtain.

3.5 Evaluating Agile Mind: Data

Broadly speaking, our method requires three categories of variables: variables recording information about “treatment,” i.e. AM implementation, “outcomes,” i.e. school-average test-scores, and “covariates,” such as student demographics and scores for tests taken prior to the introduction of AM (“pre-tests”). Both Texas and Indiana provide school-level data for the latter two categories. Texas’ “Academic Excellence Indicator System” (AEIS)¹⁰ provides information both on school-level passing rates for tests—which can serve as outcomes of interest—and thousands of covariates. The Texas Education Agency also provides school average scores on the Texas Assessment of Knowledge and Skills (TAKS) standardized tests. The Indiana Department of Education provides similar data for download from its website¹¹ in the form of Microsoft Excel “Data Reports.”

Information on school-level treatment status was provided by the Agile Mind staff. Because AM course-content is delivered online, AM staff are able to monitor each teacher’s usage; therefore, data are available on how many hours each Algebra 1 teacher has used AM. Agile Mind provided us with a spreadsheet listing schools employing teachers who used the AM Algebra 1 course for more than 70 hours of instruction. The spreadsheet also contains information on what proportion of these school’s teachers are under an AM contract,¹² what proportion of AM-contracted teachers used AM for more than 70 hours, what proportion of all school teachers used AM for more than 70 hours and at what date the school started using the AM curriculum. Finally, the spreadsheet includes a list of teachers who used AM Algebra 1 for more than 70 hours, along with the actual number of hours they were used.

We model treatment status as binary: every school with a teacher who used AM for more than 70 hours is considered a “treatment” school, and the rest are considered “control.”

3.5.1 Texas AEIS

The Texas Education Agency makes publicly available thousands of school-level variables for each of its public and charter schools. Data are available for each academic year from 1992–1993 through 2010–2011; while the form of the data (i.e. which variables are available) changes from year to year, it is mostly consistent. For pre-treatment covariates, we relied mostly on the 2009–2010 data, which included 3890 variables for each of 8322 schools of all levels and types.

¹⁰<http://ritter.tea.state.tx.us/perfreport/aeis/>, accessed 11/08/2012

¹¹<http://www.doe.in.gov/improvement/accountability/find-school-and-corporation-data-reports>, Accessed 11/8/12

¹²Schools or districts chose whether to adopt AM treatment, but in some cases the contracts were limited to particular teachers.

AEIS includes data on college-readiness,¹³ standardized tests such as TAKS, SAT and ACT, completion and dropout rates, school finance, student demographics—including data on English language learners (ELLs), special education students, student ethnicities and students’ economic backgrounds—advanced courses and instruction, and staff and class-size. Where applicable, each variable was disaggregated both by school and by 10 school subgroups: Black, Hispanic, Native America, Asian or Pacific Islander, White, Male, Female, ELL, economically disadvantaged and special education students. Each dataset contains both current data and data from the previous year; therefore, there is some overlap between AEIS datasets of different years.

Many variables are not applicable to every school: for instance, data on 9th-grade TAKS tests only apply to schools that contain a 9th grade; for all other schools, these variables are coded as NA. Similarly, when there are too few students of a particular subgroup at a school to ensure confidentiality and “statistically reliable information,” test-score data for that subgroup are masked in order to preserve confidentiality.¹⁴ Also, schools that were recently built do not have historical data—these schools will have NAs for certain historical variables.

Some AM schools began treatment before 2009, so some variables in the 2009–2010 AEIS dataset are not pre-treatment, and including them in an analysis may bias the results. We therefore deleted certain suspect data values from the dataset for schools that began treatment early enough. We assumed that the AM treatment would only affect mathematics test scores and those variables directly related to test scores—that is, we assumed data on language or social studies tests, student and staff demographics and school finances were *causally prior* to AM instruction, even if they were measured after AM instruction began.

In summary, there are four well-defined reasons for missing data: inapplicability of the variable to the school, small student subgroups, newly-built or founded schools, and the requirement for variables to be pre-treatment. Importantly, there is no missing data for the main student demographic variables: a school’s percent economically disadvantaged, percent English-language learners, percent special-education and ethnic/racial composition.

We removed all variables that were more than two-thirds missing, as well as, for each outcome of interest, removing schools with no data on that outcome. The latter step served to limit each analysis to schools that catered to students of the same age group: schools that do not have a 9th grade, for instance, do not have data for the 9th-grade TAKS test. In addition, we appended school-average TAKS results from 2006 and 2010 as a pre-tests (schools that started using AM curriculum before 2010 could not use 2010 TAKS results as pre-test variables, whereas schools that

¹³for precise definitions, see the glossary on the TEA website, <http://ritter.tea.state.tx.us/perfreport/aeis/2010/glossary.html>

¹⁴<http://www.tea.state.tx.us/index4.aspx?id=4638>. Accessed: August 26, 2013.

did not exist in 2006 had no 2006 TAKS data—hence both pretests are necessary), with indicators for missingness for both variables. The rest of the missing values were imputed with their variables’ grand means. After processing, 2330 variables remained in the dataset.

There are three AM schools that are not represented in the 2009–2010 AEIS dataset. The Estrada Achievement Center in San Antonio is specialized for students with serious disciplinary issues, and students attend Estrada only temporarily, for varying periods of time. The Pickett Center in San Antonio has fewer than 40 students, and covers grades 7–12; therefore, no grade is large enough to report average test scores. These two schools not only do not have sufficient data, but are likely to be exceptional, and are left out of the analysis. The third missing high-school, the College Transitional Academy of La Joya Independent School District (now renamed Jimmy Carter Early College High School) was established in 2010, for the 2010-2011 school year, and did not exist in the 2009-2010 school year.

3.5.2 Indiana Data

The Indiana Department of Education provides “data reports” for each of its public schools.¹⁵ Most variables relate to tests: the number of students who took the test, the number who passed and the passing rate for Advanced Placement tests, the ISTEP+ test, the IREAD-3 test, the International Baccalaureate test, End-of-Course Assessments (ECA), college placement exams and alternative assessments. Results from these tests were disaggregated (separately) into grade level, ethnicity (American Indian, Asian, Black, Hispanic, Multiracial and White), free or reduced-price lunch (FL) status, English-language-learner (ELL) status and special education status. Historical data are also available, starting as early as 2006 for some variables.

There were also variables with information on school demographics, enrollment, and graduation rates; Table 3.1 shows what proportion of the dataset is devoted to each type of variable.

	Proportion
demographics	0.25
enrollment	0.10
grad rate	0.02
test score	0.64

Table 3.1 A breakdown of Indiana School-level covariates into types

¹⁵Publicly available for download at <http://www.doe.in.gov/improvement/accountability/find-school-and-corporation-data-reports>. Accessed 1/1/2013

Results from math tests cannot be considered as pre-treatment if they record data from after the adoption of the Agile Mind curriculum; these data were deleted for treated schools.

We took similar steps to process the Indiana covariates as we did the Texas covariates: first, we removed many redundant or uninformative covariates, such as school names and IDs. We also removed variables that could not be considered pre-treatment, variables with no variance and variables which were more than 90% missing. We imputed the remaining missing values with their variables' grand means. The raw data contained 2126 variables; after processing, 1210 variables remained for 864 schools, 10 of which were treated. One of the treated schools, New Augusta Public Academy-North, administered the ECA in the summer, and therefore does not have publicly available outcome data and was excluded from the analysis.

3.5.3 Outcomes

The outcomes of interest are all standardized test results. In Texas, there are two tests of interest: the Texas Assessment of Knowledge and Skills (TAKS) and the End-of-Course STAAR (State of Texas Assessments of Academic Readiness) test (EOC). The TAKS is divided into mathematics and reading sections, one for each grade-level. For younger grades, there is a Spanish version, but for high school only an English version is available. The content of the TAKS mathematics test is based not on course content but on grade-level; therefore, it covers more mathematical topics than just Algebra, which was the target of AM curricula. In particular, the test for middle-schools that use AM was the 8th-grade TAKS, which, though it contained some Algebra content, is mostly pre-algebra and arithmetic. The last TAKS test for High-School students is the 2011 TAKS test. For two reasons, we do not expect to detect an effect of the AM curriculum on this variable: first, as noted, the TAKS test is not specifically tailored to Algebra 1, especially for eighth-grade; second, eight out of the 28 Texas schools which used AM began using it in the spring or summer of 2011—too late for it to affect the 2011 TAKS scores.

The EOC, by contrast, is based on course content: there is an EOC Algebra 1 test, which is our outcome of interest. The EOC focuses on understanding and representing functions, and solving and graphing linear and quadratic equations and inequalities.¹⁶ Graduation depends on students' cumulative scores over all of their EOCs. The STAAR tests are administered by Pearson Testing,

¹⁶A more complete list of tested concepts is available at <http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147488337&libID=2147488336>. Accessed 7/24/13. A brochure explaining the STAAR is available (also accessed 7/24/13) at <http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147502688&libID=2147502682>.

which hires graders to grade the tests.¹⁷ To date, there is no easily accessible school-level EOC data online. There are, however, school-by-school reports that can be downloaded one at a time. For each test, there are three outcomes of interest: the number of students who took the test, the percent/number of students who passed and the average score.

In Indiana, the main outcomes of interest in this study are the 2012 percents of eighth and ninth grade students who passed ECA Algebra 1. The numbers of students who took the exams and who passed are also available as outcomes. However, average scale scores are not available. The ECA Algebra 1 test is a standardized end-of-course test which mostly focuses on solving and graphing linear equations and inequalities, and solving quadratic inequalities.¹⁸ It is graded by “trained evaluators,” not the students’ teachers.

One Indiana school was missing outcome data: New Augusta Public Academy North administered its ECA over the summer, instead of in the spring, so the scores were not available online.

3.6 Evaluating Agile Mind

Now, we apply the methods outlined above to evaluate the AM Algebra 1 online program, estimating the ETT of AM usage. Technically, as mentioned above, and as will be described in some more detail below, we consider “treated” schools to be those schools in which at least one teacher used the AM Algebra 1 online resources for at least 70 hours over the course of a school year. The causal question here is whether such usage causes a change in those schools’ Algebra 1 end-of-course exam passing rates.

There are three stages to the analysis: designing the hypothetical stratified experiment by identifying a group of non-AM-using schools that highly resembles the sample of “treatment” schools, checking that the identified comparison group is indeed similar to the treatment group, and analyzing the difference in test-scores between the two groups of schools.

¹⁷Apparently it recruits on craigslist.com, among other venues, and pays \$12.00 an hour. <http://www.myfoxdfw.com/story/20662399/craigslist-ad-solicits-staar-test-graders>. Accessed 7/24/2013.

¹⁸An example, with scoring rubrics, is available at <http://www.doe.in.gov/sites/default/files/assessment/2012-algebra.pdf>. Accessed 7/24/2013.

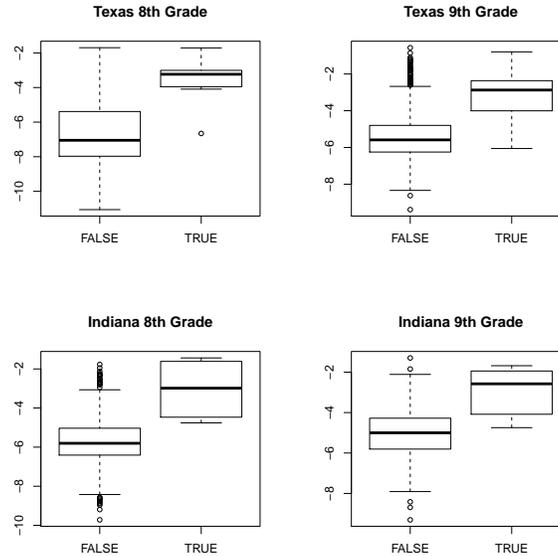


Figure 3.2 Distributions of propensity-score linear predictors for each stratum of the dataset, by AM usage

3.6.1 The Matching Process

Following Section 3.4.3, we selected a small set of school-level covariates that would pose the most concern as potential confounders, including student demographics: percent African-American, percent Hispanic, percent special-education and percent economically disadvantaged, and pre-test information: for Texas schools, school-average TAKS score from 2006 and 2010 and for Indiana schools, prior ECA passing-rates. To balance the rest of the covariates, we included principal components in the propensity-score model.

Next, for each treatment school we tried to find a set of control schools with approximately matching propensity scores. We only allowed matched schools to differ by at most 1–10% of the overall standard deviation of the total set of linear predictors; this is referred to as a “caliper” (Rosenbaum and Rubin, 1985). Within that restriction, we used the `fullmatch()` command from the R package `optmatch` (Hansen, 2007) to construct a match.

Texas Schools

We matched schools and evaluated school-level matches separately for each Algebra 1 grade, 8th or 9th, in each state. To perform the matching for the Texas high-schools, we chose to include the four principal components that most highly correlated with the average 2011 TAKS score.

These were the first, the fourteenth, the eighteenth, and the sixth. The first principal component primarily distinguished schools with high historical TAKS scores against those with low scores and high at-risk student populations. The fourteenth primarily distinguished schools with high African American student populations and high school-leadership funding but low dropout rates against those with high dropout rates and high GED rates. The eighteenth primarily distinguished schools with high dropout rates and low historical TAKS scores and passing rates against schools with low dropout rates and high TAKS scores. Finally, the sixth component primarily contained information on TAKS participation rates (that is, what proportion of students took the TAKS exam). Together, these four PCs account for 10% of the variation in the dataset, but about 64% of the variation in 2011 TAKS passing rates for unmatched control schools. To find a comparison group for the middle schools, we only included one principal component, since there are fewer AM middle schools than high schools. The first principal component—which is also the principal component that most highly correlates with 8th grade 2011 TAKS scores—is similar to the first principal component in the high-school sample. It contributes 7% of the total variation in the sample, and about 54% of the variation in 2011 8th-grade TAKS passing rates among unmatched control schools.

For practical reasons—we needed to manually record EOC data for each of the matched schools in Texas—we did not want the total number of matched schools in Texas to be too high. To limit the total number of matched schools without sacrificing too much effective sample size and the quality of the match, we varied the maximum number of controls matched to each treatment and the maximum allowable difference in propensity scores between the treatment and control schools in the same matched set (the caliper).

For Texas middle schools, the optimal configuration was a caliper of 5% of the overall standard deviation of the propensity scores between the treatment and control groups (so AM and non-AM schools in the same matched set could differ by at most 0.0875 in their estimated log-odds of being treated). Also, the maximum number of control schools for each AM-school was set to five. With this configuration, the analysis will include nine matched sets, each of which contains one AM school and varying numbers of non-AM schools. The total number of matched schools, including AM and non-AM schools, was 43, with an effective sample size of 13.3.

In the high-school sample, we chose a caliper of 10% of the standard deviation of propensity scores (allowing schools in the same matched set to differ by 0.12 in their estimated log-odds of being treated) and a maximum of seven non-AM schools matched to each AM school. This resulted in 16 matched sets (one for each AM school), with a total of 110 matched schools, and an effective sample size of 26.7.

Before matching, the covariate profiles of the treated and control schools were significantly

different; the differences are displayed in Tables 3.2 through 3.5. In particular, in the Texas high-school sample, the AM schools had, on average, lower pre-test scores and higher Hispanic and economically-disadvantaged student populations. The samples also differed significantly on the first and fourteenth principal components of the covariates. Also, the prognostic scores (see below) were imbalanced between the AM-using and non-AM-using schools.¹⁹ The probability of a randomized experiment resulting in this (or higher) magnitude of covariate imbalance is less than 10^{-8} . The imbalance pattern in the middle-school sample is similar, although in the prognostic scores do not appear significantly imbalanced.

Propensity-score matching significantly improved the balance of the pre-treatment covariates between the treatment and control groups, as Figure 3.3 and Tables 3.2 through 3.5 show. None of the covariates used in the matching were significantly different between the AM schools and the matched non-AM schools. In fact, the covariate balance achieved between the treatment and matched controls was better than would be expected in over 90% of randomized experiments.

Indiana Schools

The AM schools were matched to sets of similar non-AM schools within the same grade-level: that is, schools that administered the Algebra 1 ECA in eighth grade were matched to other schools with eighth-grade Algebra 1 ECA data, and schools that administered the test in ninth grade were matched to other similar schools. Operationally, we divided the data into two sets: “middle schools” with 8th-grade ECA data and “high schools”²⁰ with 9th grade ECA data. There were four AM schools in the middle school set, and five in the high school set.

Within each dataset, propensity scores were estimated based on the following covariates: percent FL, percent ELL, percent Black, percent Hispanic, percent special education, total enrollment, percent of eighth- or ninth-grade students who passed the 2010 Algebra 1 ECA, and the first two principal components, which had the highest correlation with the outcomes. In 8th grade schools, 26% of the data’s variance was attributable to these two components, and 8% of the variance in ECA passing rates, in the unmatched control schools, was attributable to these components. In 9th grade schools, 29% of the data’s variance was attributable to the first two components, and about 27% of the variance in ECA passing rates, in the unmatched control schools, was attributable to these components. The distributions of propensity score linear predictors are shown in figure 3.2.

¹⁹Comparing prognostic score balance between the matched set and the entire sample is somewhat problematic, for reasons discussed in Section 3.2.3.

²⁰“middle schools” and “high schools,” here, are in quotes since not all “middle schools” are necessarily actually middle schools—they are just schools with 8th grade data.

Next, we used the R function `fullmatch()` from the `optmatch` package to find propensity score-based matches. We varied the caliper to optimize effective sample size and covariate balance. In the high school sample, we settled on a caliper of 0.05 of the pooled standard deviation of the propensity score linear predictors, which yielded a match with an effective sample size of 8.15 and an `xBalance` χ^2 p-value of 0.90. In the high school sample, we settled on a caliper of 0.1 of the pooled standard deviation of the propensity score linear predictors, which yielded a match with an effective sample size of 6.25 and an `xBalance` χ^2 p-value of 0.76.

Testing the Matches

After the matches were specified, we used PCR to model the outcomes of the non-matched schools.²¹ Next, we computed the predicted values of each of those models for the matched schools, and computed balance statistics. The results, before and after matching, are in Tables 3.2, 3.3, 3.4 and 3.5; these tables also contain overall χ^2 p-values, which indicate that matching drastically improved covariate balance. A graphical representation of covariate balance before and after matching appears in Figure 3.3.

Some of the matched schools in Texas did not have EOC data available, and were dropped from the dataset; as can be seen from the omnibus p-values, excluding these schools degrades balance slightly, but not to a degree that may be worrisome. This missing data, however, may be worrisome for another reason: the fact that the outcomes were missing could, conceivably, be influenced by AM usage. If that were the case, outcome missingness would not be causally prior to treatment, and conditioning on it, which is the effect of conducting analysis without these schools, would bias the result. For instance, if AM caused the potentially highest-performing schools to decide not to hold an exam in 2012, the estimate would be biased downwards. On the other hand, if some control schools would have conducted exams had they used AM, and this subset of schools would have had potentially higher test scores, then the estimate would be biased upwards. One approach to solving this problem would be to apply partial identification methods (Horowitz and Manski, 2000; Tamer, 2010), which attempt to bound the size of a treatment effect given the observed data and minimal assumptions about the data's distribution (such as the distribution's support). This is beyond the scope of the present study, but would be a fertile area for future research. We will assume that missingness in EOC passing rates is not affected by AM, and that conditioning on it would not bias causal estimates.

²¹Following the results of Chapter 4, we also used ridge regression for similar calculations; the results were similar to those with PCR.

	Full Sample			Matched Sets	
	std.diff	z		std.diff	z
black	1.64	3.23	**	-0.11	-0.19
hispanic	0.33	0.66		-0.64	-0.77
freeLunch	0.73	1.44		-0.39	-0.62
ell	1.27	2.51	*	-1.08	-1.13
specialEd	0.76	1.52		0.23	0.29
totalEnrollment	0.32	0.63		-0.10	-0.18
pretest	0.08	0.17		0.02	0.06
X1	0.08	0.16		0.45	0.76
X2	1.35	2.66	**	-0.18	-0.30
prog	-1.16	-2.30	*	0.12	0.17
p-value	0.02			0.72	

Table 3.2 Covariate balance before and after matching for Indiana middle schools. X1 and X2 are the first two principal components, prog denotes prognostic scores, and p-value denotes omnibus xBalance() χ^2 p-values. There are 505 schools in this stratum, with four treated schools matched to a total of 109 untreated schools.

	Full Sample			Matched Sets	
	std.diff	z		std.diff	z
black	0.46	1.02		-0.19	-0.37
hispanic	0.16	0.37		-0.08	-0.22
freeLunch	0.08	0.18		-0.21	-0.63
ell	0.45	1.00		-0.05	-0.10
specialEd	0.69	1.54		-0.45	-0.57
totalEnrollment	0.64	1.43		0.02	0.03
pretest	-0.89	-1.96	*	-0.03	-0.07
X1	0.73	1.61		0.21	0.38
X2	-0.64	-1.42		0.09	0.21
prog	-0.28	-0.61		0.65	1.26
p-value	0.10			0.85	

Table 3.3 Covariate balance before and after matching for Indiana high schools. X1 and X2 are the first two principal components, prog denotes prognostic scores, and p-value denotes omnibus xBalance() χ^2 p-values. There are 400 schools in this stratum, with five treated schools matched to a total of 132 untreated schools.

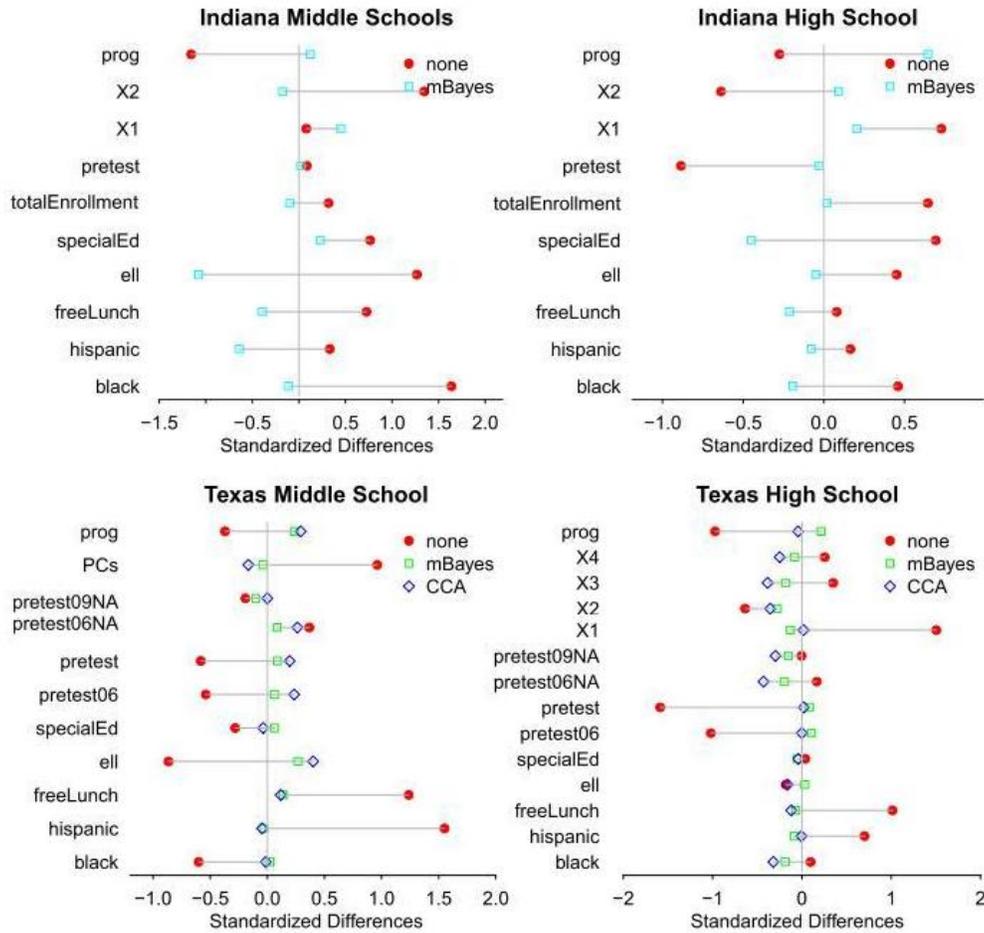


Figure 3.3 Covariate balance before and after matching (and with and without schools with missing outcomes—without denoted “CCA”—for Texas schools)

Since End-of-Course (EOC) data is only available for the matched schools in the Texas samples, the cross-validation used to specify the PCR for the Texas samples was done using TAKS data as an outcome.

3.6.2 Outcome Analysis

To estimate the effect of AM on learning, we focused on the percent of a school’s 8th or 9th grade students who passed the Algebra 1 ECA or EOC, for Indiana and Texas, respectively. Unfortunately, average scale scores, which may contain more aggregate information, were not available in Indiana, so when we aggregated the data across state and grade-level, we could only use percent passing as

	Full Sample			Matched Sets		Matched Schools with Outcomes	
	std.diff	z		std.diff	z	std.diff	z
black	-0.60	-1.80	.	0.02	0.26	-0.01	-0.15
hispanic	1.55	4.61	***	-0.03	-0.49	-0.05	-0.75
freeLunch	1.24	3.69	***	0.14	0.94	0.12	0.80
ell	-0.87	-2.58	**	0.27	0.72	0.40	1.13
specialEd	-0.28	-0.84		0.06	0.23	-0.04	-0.15
pretest06	-0.54	-1.61		0.06	0.19	0.23	0.75
pretest	-0.58	-1.74	.	0.09	0.23	0.20	0.60
pretest06NA	0.37	1.10		0.09	0.18	0.26	0.57
pretest09NA	-0.19	-0.57		-0.10	-0.38	0.00	0.00
X1	0.96	2.87	**	-0.04	-0.17	-0.17	-0.75
prog	-0.37	-1.11		0.24	0.78	0.29	1.20
p-value	0.00			0.96		0.66	

Table 3.4 Covariate balance before and after matching for Texas middle schools. X1 is the first principal component, prog denotes prognostic scores, and p-value denotes omnibus χ^2 p-values. The third column assesses the match after removing schools for which there was no outcome data available. There are 539 schools in this stratum, with nine treated schools matched to a total of 44 untreated schools

an outcome.

The following results assume that, after accounting for all measured covariates—the matrix X —AM and non-AM schools are directly comparable: there is no statistical relationship between AM usage and test scores other than whatever effect AM may have had on test scores. As in any study based only on observational data, there is a possibility of hidden bias: if an unmeasured variable is correlated with both test scores and AM usage, our results may be skewed. Below, in subsection 3.6.4, we will attempt to directly assess this possibility, and adjust our estimates accordingly.

Table 3.6 shows the results of the outcome analysis. Averaging across Texas and Indiana, and across 8th and 9th grades, we estimate the effect of a teacher using AM for at least 70 hours as increasing the number of students who pass the Algebra 1 End-of-Course exam by about one percentage point; however, this estimate was statistically indistinguishable from zero ($p=0.63$). Indeed, due to limitations in the available data, we were unable to reach any firm conclusions about the nature and extent of AM’s effectiveness in improving end-of-course test scores. In Indiana, in Texas high schools and overall, we were unable to detect any statistically-significant effects. In Texas middle schools, there is a significant negative effect on the percent of students who passed.

	Full Sample			Matched Sets		Matched Schools with Outcomes	
	std.diff	z		std.diff	z	std.diff	z
black	0.10	0.39		-0.19	-0.56	-0.32	-0.88
hispanic	0.70	2.78	**	-0.08	-0.31	-0.01	-0.02
freeLunch	1.01	4.01	***	-0.07	-0.37	-0.12	-0.59
ell	-0.18	-0.72		0.03	0.11	-0.16	-0.65
specialEd	0.04	0.15		-0.05	-0.17	-0.04	-0.12
pretest06	-1.02	-4.03	***	0.11	0.41	-0.00	-0.00
pretest	-1.59	-6.22	***	0.08	0.39	0.02	0.10
pretest06NA	0.17	0.66		-0.20	-0.56	-0.43	-1.07
pretest09NA	-0.00	-0.02		-0.15	-0.46	-0.30	-0.99
X1	1.50	5.91	***	-0.13	-0.61	0.02	0.09
X2	-0.64	-2.53	*	-0.28	-0.84	-0.36	-1.00
X3	0.35	1.39		-0.18	-0.43	-0.38	-0.79
X4	0.26	1.02		-0.08	-0.32	-0.25	-0.92
prog	-0.97	-3.85	***	0.21	1.13	-0.04	-0.22
p-value	0.00			0.91		0.76	

Table 3.5 Covariate balance before and after matching for Texas high schools. X1,...,X4 are the first four principal components, prog denotes prognostic scores, and p-value denotes omnibus xBalance() χ^2 p-values. The third column assesses the match after removing schools for which there was no outcome data available. There are 1371 schools in this stratum, with 16 treated schools matched to a total of 94 untreated schools.

		Raw Passing Rates		Peters-Belson-Adjusted	
		Effect Estimate	p	Effect Estimate	p
Indiana	8th	2.86	0.78	1.04	0.92
	9th	11.45	0.27	-1.62	0.75
Texas	8th	-9.59	0.03	-8.94	0.02
	9th	4.47	0.41	2.56	0.59
Overall		2.12	0.52	1.28	0.63

Table 3.6 Test statistics (weighted difference-in-means averaged over matched pairs) and p-values for the effect of AM in each of the datasets and overall.

3.6.3 Does AM Affect the Number of 8th-Grade Students Who Pass?

	Effect: % Took Test	p-value.	Effect: % Passed Test	p-value
IN 8	-0.02	0.83	0.00	0.97
IN 9	-0.03	0.71	0.04	0.57
TX 8	0.10	0.55	11.58	0.42
TX 9	0.02	0.81	0.44	0.91
Combined	0.03	0.66	3.19	0.48

Table 3.7 Estimates and p-values for the effect of AM on the number of students who took Algebra 1 and the percentage of the total cohort who passed, in each stratum.

Table 3.6 shows that, for Texas middle schools, AM seems to *decrease* the percentage of students passing the EOC exam, by about 9 percentage points ($p=0.02$). This may be a result of random chance (indeed, Gelman and Weakliem (2009) cautions that statistically-significant effects that are surprisingly large should be interpreted with extra caution, especially when sample sizes are small). An alternative explanation is that AM causes more eighth-grade students to attempt the EOC exam; these “marginal” students have a lower passing rate than the students who would have taken the exam anyway, and pull down the school’s overall passing rate.

	State-Grade	T_Z	ρ	Effect Estimate	Margin of Error
IN	8th	3.00	0.20	1	21
	9th	3.00	0.50	-1.6	23
TX	8th	2.00	0.30	-8.9	9.9
	9th	4.00	0.40	2.6	17.5
Overall		4.00	0.50	-1.3	10.8
Overall		2.00	0.20	-1.3	6.8

Table 3.8 Estimates and with 95% margins of error for the effect of AM on the percent of students who passed EOC exams. These margins of error account for possible hidden bias, the hypothetical extent of which is quantified in ρ and T_Z .

This raises the question: does AM tend to increase the absolute number of students who take or pass the exam? The size of the 8th- or 9th-grade class was not a matching variable, but would be highly relevant to this question. To solve this problem, we divided the number of students passing the EOC (in Texas) and ECA (in Indiana) by the size of the 8th- or 9th-grade class, which we assume is unaffected by AM. Then, we tested the effect of AM on this ratio, using both raw ratios and Peters-Belson-adjusted ratios, as above. We used the same number of principal components in the prognostic models as above. The results of the tests are found in Table 3.7: we estimate

that AM increased the percentage of students taking algebra 1 by 0.03 percentage points, and the percent of the cohort who passed by about 3 percentage points, averaging over states and grades. However, these results also failed to achieve statistical significance: we are unable to reach any firm conclusions about whether AM affected the percentage of students who took Algebra 1 in 8th or 9th grade, or the percentages of the total classes who passed the Algebra 1 end-of-course exams.

	IN 8th		IN 9th		TX 8th		TX 9th		
	ρ	T_z	ρ	T_z	ρ	T_z	ρ	T_z	
%Black	-0.12	2.77	-0.06	0.73	-0.07	-1.18	-0.08	-1.59	
%Hispanic	0.03	-1.43	-0.07	0.01	-0.12	1.33	0.04	-1.61	
% Free Lunch	-0.04	-0.43	-0.04	-1.27	0.30	1.05	-0.13	1.50	
% ELL	-0.06	1.89	0.05	0.44	-0.05	-1.57	-0.13	1.82	
% SpecialEd	0.00	1.04	-0.08	1.56	-0.19	-0.68	-0.09	-1.75	
Total Enrollment	0.15	-0.34	0.04	-1.79					
pretest	0.03	0.46	0.43	-2.38	-0.11	0.34	-0.14	1.25	
pretest09					-0.10	0.95	0.11	-3.39	
pretest06NA					0.12	1.28	0.03	0.84	
pretest09NA					0.00	-0.81	-0.01	0.22	
PCs	1	0.09	0.81	0.06	2.64	-0.26	0.61	-0.37	1.73
	2	-0.18	0.69	0.09	-0.90			-0.27	-3.06
	3							0.25	1.43
	4							0.06	-0.39

Table 3.9 Benchmarks to choose appropriate values for T_Z and ρ in the sensitivity analysis.

3.6.4 Sensitivity Analysis

Since, in this case, the treatment assignment mechanism is unknown, there is a possibility of hidden bias in this analysis; there may be an unmeasured variable U that correlates with both AM usage and the outcome, and biases the causal estimate. To assess this risk, we conducted a sensitivity analysis of the result—an attempt to include uncertainty about hidden bias in the final estimates. The method, following Hosman et al. (2010), parameterizes the extent of possible confounding in terms of two parameters: ρ , which represents the hypothetical confounding variable U 's relationship with the outcome Y , and T_Z , which represents U 's relationship with the Agile Mind usage. Given values for T_Z and ρ , the method produces a “sensitivity interval”: an interval of plausible values for the causal effect, given the presence of a confounder with relationships with Z and Y no greater than T_Z and ρ , respectively.

First, to gauge which values of T_Z and ρ are plausible, we calculate real values of these parameters for some of the most important measured variables in the dataset. Based on the results of this study, shown in Table 3.9, we decided on a set of values for T_Z and ρ , for each state-grade combination, and calculated sensitivity intervals. These choices, and the resulting intervals, are displayed in Table 3.8. The sensitivity intervals strengthen the conclusion that, given the limitations of the data, it is impossible to precisely estimate the effect of AM. Indeed, the effect of a teacher using AM for at least 70 hours, averaged across 8th and 9th grade schools in Texas and Indiana could be as low as -12.1 percentage points (that is, lowering the passing rate by 12.1 points) and as high as 9.5 percentage points, using conservative criteria to account for possible hidden bias. With less conservative criteria, effects are estimated as being between -8.1 and 6.8 percentage points. An additional notable conclusion is that the negative, statistically significant result for Texas 8th grade students ceases to be significant when possible hidden bias is included in the analysis; this indicates that the result may be a result of such bias.

3.7 Discussion

As more data becomes available, social scientists interested in questions of causation will need methods to incorporate large numbers of covariates into their causal models. Propensity scores, themselves powerful dimension-reducers, can be inadequate to the task, especially when the number of covariates is large compared to the number of treated or control units. Some variants of regression adjustment, however, can accommodate arbitrarily large numbers of covariates; these methods, though, are designed not to unbiasedly estimate treatment effects, but to build predictive models.

This chapter bridged some of the gap between a particular high-dimensional multivariate technique—principal components analysis—and causal inference with propensity score matching. The central idea is that researchers seek to estimate treatment effects unbiasedly, but the relationships between covariates and outcomes need not be estimated correctly; only prediction accuracy is necessary. Therefore, principal components regression modeling of the relationship between covariates and outcomes can assist propensity score matching when many covariates are present.

Three separate, if synergistic, uses of principal components appeared in this chapter. The first was to include principal components in a propensity score model, along with a small set of variables thought, based on substantive background, to be important to selection into treatment and predictive of outcomes. The second and third depended on estimating prognostic scores: predictions of control potential outcomes from pre-treatment covariates. Prognostic models were trained on a set of control

units not included in the propensity-score match. This chapter suggests using these predictions to test the propensity score match's validity, as well as adjust the propensity score analysis. If either the propensity score match or the prognostic score estimation were sufficient to eliminate bias, their use together produces unbiased treatment effect estimates. Even if this is not the case, prognostic score potential outcome predictions with low mean-squared-errors can limit the bias of a causal analysis.

Next, this chapter attempted to use these new statistical methods to estimate the effect of the Agile Mind Algebra 1 mathematics enrichment program on school-level standardized test passing rates. One contribution of this paper over Correnti et al. (2008) is that we accounted for the fact that most teachers under AM contract do not seem to take full advantage of AM services. By considering only schools whose teachers use AM for more than 70 hours as "AM schools," we increased the likelihood of finding a significant effect. Another important difference is that this paper uses EOC data in Texas and ECA data in Indiana as the outcomes of interest. These tests directly measure Algebra 1 skills, and are more likely than the TAKS to capture the effects of AM usage.

The causal-inference technique used, propensity score matching, was successful, yielding large matched samples containing schools using AM and extremely similar sets of schools that did not use AM. We were able to implement novel methods to reduce or eliminate bias resulting from thousands of covariates that did not appear directly in the matching algorithm. In particular, any information we were able to extract from these covariates that would have predicted schools' outcomes without AM usage was well-balanced between treatment and control groups. The techniques applied in this paper were designed to maximize the chances of detecting an effect of AM without incurring bias from omitted variables.

Both the methodological and substantive contributions in this chapter may be strengthened by future analysis. A simulation study, or an analysis of a dataset with a known treatment effect, would shed light on the extent of the ability of principal components analysis to help propensity score matched designs accurately estimate treatment effects. Some weaknesses in the analysis of Agile Mind could also be corrected with access to better data: the measure of implementation that we had access to—whether a teacher in each school used AM for at least 70 hours—may be overly simplistic and arbitrary; access to better implementation data may lead to a more informative estimate. Also, school-average test scores, instead of passing rates, would likely be a more informative outcome to measure.

Chapter 4

Multivariate Statistics and Machine Learning for Causal Inference: Modern Regression Prognostic Scores

4.1 Introduction

Machine learning, and, in particular, high-dimensional data analysis, has been one of the most productive areas of statistics in the past two or three decades. However, few of its contributions have been adapted to causal inference. This is not surprising: machine learning, for the most part, is concerned with prediction more than with inference; its goal is not to learn correct relationships or models, but to predict outcomes reliably. That is to say, in machine learning, correlation and causation are of equal value. In causal inference, of course, this is not the case. Furthermore, high-dimensional data—the fodder for machine learning—has, until recently, been fairly uncommon in causal inference settings.

Now, however, high-dimensional data is poised to become more common in causal inference, and, in particular, in educational evaluations, due to the increasing availability of educational data. Some recent attempts at integrating machine learning and multivariate techniques into causal inference are Schneeweiss et al. (2009); Belloni et al. (2011, 2012).

The previous chapter has suggested a causal-inference application in which prediction is the (proximate) goal: prognostic scoring in a propensity-score matching design. Say a researcher has used a small number of covariates to estimate propensity scores and reconstruct a hypothetical stratified experiment, and she has set aside a portion of the untreated group—perhaps a portion that was not sufficiently similar to the treated group—and excluded it from her matched design. However, she has access to a large number of additional covariates, and she is concerned that their omission from the propensity score model will bias her causal estimates. Then the prognostic-score

solution is to use a high-dimensional regression technique, along with the full set of covariates, to train a prediction model on the set of units that were excluded from the match. This model's predictions in the matched sample are prognostic scores: researchers can use them to test overall covariate balance among matched units, or to adjust causal estimates, as described above.

Section 3.3 suggests estimating prognostic scores with principal components regression. This approach is attractive for a number of reasons. These include the fact that principal components can give insight into possible underlying structure in the covariate matrix X , and, in some instances, analysts can interpret principal components as new variables. Further, using principal components regression dovetails nicely with the method in Section 3.3.1, in which principal components are added to the prognostic model.

However, several other prediction methods in high dimensional data are available. In particular, motivated, perhaps, by the rise of genomics and facilitated by the explosion of computing power, the the past two decades have seen rapid development in high-dimensional data analysis, including regression methods such as the LASSO (Tibshirani, 1996), the Elastic Net (Zou and Hastie, 2005) and Supervised Principal Components Regression (Bair et al., 2006). Of course, these methods join some other, older high-dimensional regression methods, including ridge regression (Hoerl and Kennard, 1970) and partial least squares regression (Wold, 1966).

Proposition 2, double unbiasedness,¹ requires that the prognostic model (3.18) unbiasedly estimate β , or unbiasedly predict Y_C .¹ Further, the proposition suggests that the model must be correct—that is, account for the entire relationship between Y_C and X —in order to improve the validity of ETT estimate $\hat{\tau}_{PBP}$. However, when $p \gg n$, these conditions will be extremely difficult, if not impossible, to fulfill. On the other hand, (3.20) suggests that a model that reliably predicts Y_C , even if it is neither an unbiased technique nor does it replicate the data-generating-model, may go a long way towards reducing the bias of $\hat{\tau}_{PBP}$. In particular, (3.20) suggests that the prediction mean-squared-error (MSE) of the prognostic model is an important criterion for its usefulness. An analyst can, on a case-by-case basis, choose the optimal regression technique from among the possibilities in the literature. Given the difficulty of finding the correct model, this choice may be motivated purely by prediction-accuracy.

This chapter will suggest some possible approaches to investigating various high-dimensional modeling techniques and choosing one for causal inference. To do so, we will focus on the data from our study of Agile Mind, described in Section 3.5.

The following section suggests some descriptive multivariate techniques to explore the structure

¹I believe that the proposition would still hold if the bias in the model's prediction of Y_C were independent of treatment status Z , but I have not proven that to be true.

of the covariate matrix X , which may be helpful for high-dimensional modeling (and interesting in its own right). Section 4.3 will briefly describe several high-dimensional modeling techniques. Section 4.4 will discuss model selection by cross-validation (CV), and use CV to determine which modeling technique would perform optimally for the AM data. Section 4.2 will present an alternative validation technique for choosing a model and estimating prediction MSE. Finally, Section 4.6 will conclude.

4.2 The Multivariate Structure of X

Before choosing a multivariate method, it may be reasonable to examine the multivariate structure of the covariate matrix X , and its relationship to the control outcomes, Y_C . One of the first steps in this direction is, of course, the singular value decomposition and principal components analysis already computed. As mentioned in subsection 3.2.2, the squared singular values S^2 —eigenvalues of $X'X$, the observed covariance matrix for studentized X —summarize information about the dataset's collinearity and variance structure. In particular, the sum of S^2 is equal to the sum of the variances of X 's variables, and each squared singular value S_k , divided by this sum, $S_k^2 / \sum S^2$, is a measure of the proportion of X 's variance is attributable to the k th principal component. Figure 4.1 displays the cumulative proportion of variance explained by the first p proportion of principal components in each stratum of the dataset. So, for instance, the first 20% of principal components in the Indiana strata explain about half of the total variance in their respective strata, but the first 20% of Texas PCs explain almost 70% in their own strata. This suggests that the underlying structure in the Texas strata has a lower dimension than in the Indiana strata. In other words, more information is summarized in fewer principal components.²

Next, it may be useful to examine the pattern of correlations between the variables in X and the control outcomes Y_C . The Fisher transformation (Fisher, 1915) of the Pearson correlation coefficient, which is identical to the inverse hyperbolic tangent function, transforms correlation coefficients so that, if the two correlated variables are jointly normal and their observations are independent, their distribution will be approximately normal. By calculating the Fisher transformations of the correlations between each of the covariates and the outcomes, and plotting their quantiles against the quantiles of a normal distribution, researchers can often observe interesting patterns. Figure 4.2 does just that. Since, with jointly normal variables, the standard deviation of the Fisher-transformed

²The same pattern is present when the variable on the x-axis is the number of principal components, not the proportion of the total.

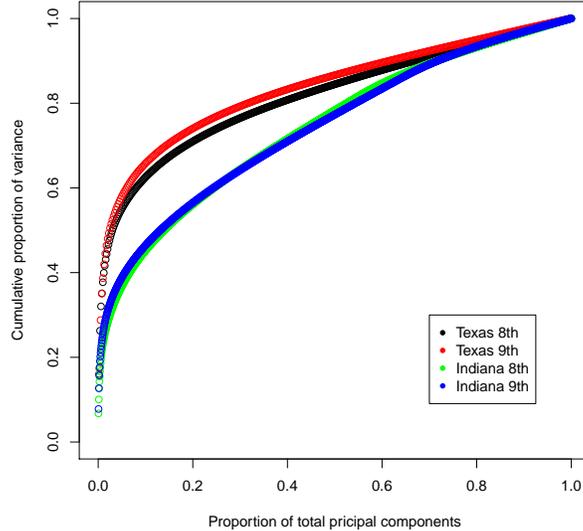


Figure 4.1 The cumulative proportion of total variance explained by proportion of principal components, for each stratum of the AM dataset

correlation is approximately $1/\sqrt{n}$, where n is the sample size, the transformed correlations here are multiplied by \sqrt{n} , giving them an approximate standard deviation of one, so they are, roughly, Z-scores; the corresponding raw correlations are also marked, on the right-hand side of the plot.

In both datasets, but especially in Texas, the correlations of the covariates with the outcomes are almost all highly significant, with Z-scores as high as 40 and as low as -40. In Indiana, the correlations are also significant, but to a lesser extent; in addition, a few variables stand out as being unexpectedly highly correlated with the outcomes. In other words, the Indiana plots suggest that there are a few variables that correlate very strongly with the outcomes, and that a prediction strategy that relies on identifying and focusing on those variables would bear the most fruit. On the other hand, the distribution of correlations in the Texas data suggests that predictive information about Texas outcomes is more smoothly distributed among the covariates.

Does this point in a different direction than the information from Figure 4.1? The distribution of singular values indicates that the underlying structures of the Texas strata have lower dimension than the Indiana strata; the distribution of bivariate correlations with the outcome indicates that, possibly, a small number of covariates are necessary to optimally predict outcomes in Indiana, but not in Texas. Two differences between the plots resolve this apparent contradiction: the singular values refer to variation explained by principal components, which themselves are combinations of many covariates, whereas the bivariate correlations refer to specific, raw covariates. It may be, for instance,

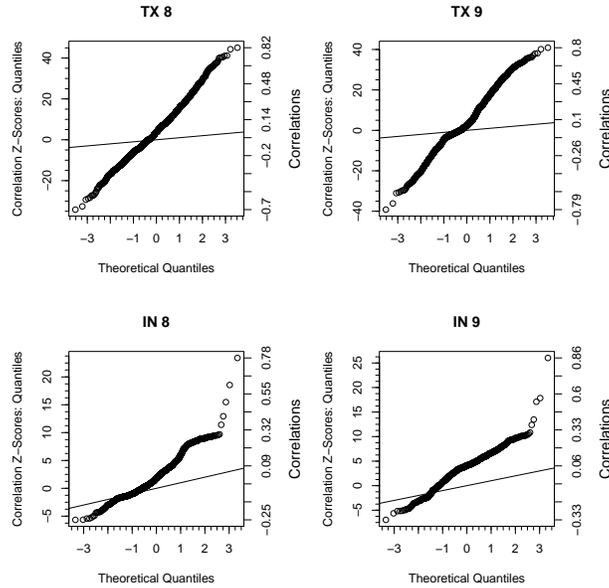


Figure 4.2 Normal quantile-quantile plots of Fisher-transformed correlation coefficients, multiplied by \sqrt{n} , between individual covariates and outcomes, for each AM dataset stratum, along with a line of slope 1. Corresponding raw correlations are available to the right of the plots.

that a large number of highly-correlated covariates (such as measures of pretest scores) are all similarly correlated with outcomes, but can all be roughly summarized in one principal component. In such a case, the underlying structure would have low dimension, but many covariates would be highly correlated with the outcomes. In addition, the correlations refer to the relationships between the covariates and the outcomes, whereas the principal components refer to the relationships among the covariates.

This information from the distribution of singular values and of correlations between covariates and outcomes suggests that, perhaps, the LASSO will perform well for Indiana schools by identifying a few important covariates. On the other hand, principal components regression will require fewer components for accurate predictions in Texas than in Indiana.

4.3 Short Overviews of Models to be Considered

A host of techniques for high-dimensional regression are available; Hastie et al. (2005) provides a nice survey, from which much of this sub-section is drawn. In this paper, we will consider a subset of those techniques, which we feel roughly represents the field. We have attempted to include

the most well-known and popular techniques: ridge regression, the LASSO, partial-least-squares regression and Random Forest prediction.

The LASSO and ridge regression emerge from the same principle, but with a subtle yet important difference. Ordinary least squares (OLS) attempts to find a vector of estimated coefficients $\hat{\beta}$ such that the sum of squared differences between the actual data values and the predicted values, $\|Y - \hat{Y}\|^2$, is minimized. When variables in the dataset are collinear, the OLS estimates $\hat{\beta}$ can become unstable. Penalized least squares techniques, such as ridge regression and the LASSO, attempt to limit this instability, at the price of some bias, by introducing a penalty on large coefficients:

$$\hat{\beta}_{PR} = \arg \min_{\hat{\beta}} \|Y - \hat{Y}_{\hat{\beta}}\|^2 + \lambda \|\hat{\beta}\|_{\downarrow} \quad (4.1)$$

where λ is a tuning parameter which controls the amount of shrinkage, or the strength of the penalty. In ridge regression, the penalty is a multiple of the L_2 norm of $\hat{\beta}$: the sum of squared $\hat{\beta}$ coefficients. In the LASSO, the penalty is a multiple of the L_1 norm: the sum of the absolute values of $\hat{\beta}$ coefficients. This subtle mathematical difference has an important practical effect: the LASSO will shrink some variables' coefficients all the way to zero, effectively removing them from the model. For that reason, it is often conceptualized as a variable selection technique (Meinshausen and Bühlmann, 2006; Tibshirani et al., 1997, e.g.). Ridge regression does not share this property, and every variable contributes to a ridge model. For this reason, it seems likely that when an outcome can be predicted best from a small subset of covariates, the LASSO would outperform ridge regression, whereas when a large proportion of available coefficients covary with the outcome at similar scales, ridge regression may carry the day.

Ridge regression, in fact, can be conceptualized as a smoothed version of PCR. Hastie et al. (2005) rewrite the ridge fitted values, $X(X'X + \lambda I)^{-1}X'Y$ substituting the SVD, USV' , for X , and show that

$$\hat{Y}_{ridge} = \sum_i u_i \frac{d_i^2}{d_i^2 + \lambda} u_i' Y \quad (4.2)$$

meaning that principal components with lower eigenvalues are shrunk more, and vice-versa. This suggests that PCR and ridge regression will produce similar results.

Partial least squares regression (PLS) is also similar to PCR, in that it involves constructing composite variables, and regressing Y on them. In the case of PLS, however, the derived variables are constructed with reference to covariates' relationship with Y , not only on their relationships to each other. While PCR finds mutually orthogonal derived variables with maximal variance, PLS finds mutually orthogonal derived variables that maximize the product of the variable's variance,

and its squared correlation with Y (Frank and Friedman, 1993). Just like in PCR, performing PLS regression requires choosing the number of composite variables to include as regressors.

A final option, Random Forests (RF) (Breiman, 2001) fully replaces the parametric assumptions in the linear regression models with increased computational cost. RF builds on another non-parametric prediction algorithm, the regression-tree (Morgan and Sonquist, 1963; Breiman and Ihaka, 1984). A RF predictor repeatedly (B times) randomly sub-samples from the dataset, each time drawing, with replacement, a sample the same size as the dataset itself. In each sub-sample, the algorithm grows a regression tree: repeatedly, it randomly chooses a subset of size m of the available predictors, and partitions the subjects on those predictors in a way that minimizes the within-partition variance of the sample. This process is repeated, in each tree, until each partition fits the sub-sampled data perfectly: the within-partition variances are all zero. The result is B trees, each of which can predict an outcome for a new case, using that case's predictors to choose an appropriate partition, and predicting the value of the training set in that partition. The RF algorithm improves on the predictions of the individual trees by averaging their respective predictions.

4.4 Which Models Win? A Cross-Validation Exercise

Perhaps the most powerful technique to determine which high-dimensional modeling technique will yield estimated prognostic scores with the lowest mean-squared-error is cross-validation (CV) (Kohavi et al., 1995; Efron and Gong, 1983). To implement k -fold CV, a researcher randomly divides the rows of X , corresponding to subjects, into k roughly equal parts. Then, k times, use $k - 1$ of those sections (the “training set”) to estimate a model, and use the remaining section (the “testing” set) to test the model's predictions. The mean of the squares of the estimation errors is an estimate for the model's MSE.³

The choice of k is, like many choices in statistics, a trade-off between bias and variance. The highest possible value of k is $k = n$, the sample size; this is referred to as leave-one-out cross-validation, and has the lowest possible bias: it almost exactly replicates the process of using the full dataset to predict new values. However, because in this case the training sets resemble each other so strongly, the variance of the CV estimate is high. Conversely, for lower values of k , the variance of the CV estimate is lower, at the expense of some bias. Kohavi et al. (1995) recommends using $k = 10$, and Hastie et al. (2005) similarly recommends $k = 5-10$; for our study, based both on

³More complicated, and perhaps slightly more accurate, methods of CV are available (Efron and Tibshirani, 1997, e.g.), but here we restrict ourselves to the old-fashioned version.

computational concerns and on observed variance of CV estimates we used $k = 5$.

Cross-validation would be extremely expensive for the RF algorithm, but also, fortunately, unnecessary. RF grows a separate regression-tree in each of B random sub-samples of the data. Each subsample will, with very high probability, exclude certain subjects, and the MSE of that subsample's tree can be estimated as its empirical MSE when predicting the excluded cases. The average of these MSEs, over all B trees, referred to as the out-of-box (OOB) error estimate, is an estimate of the model's prediction MSE.

To examine various high-dimensional prognostic-score estimation techniques, we conducted a 5-fold CV in each stratum of the AM dataset. Figures 4.1 and 4.2 suggest that estimators may perform differently on datasets from different states, but similarly within states. Each modeling technique involves a tuning parameter, denoted here as λ . For PCR and PLS, we tried from 1 to 101 components in the CV. To choose the tuning parameters for ridge regression, we first narrowed the options down by randomly splitting each stratum half-way into training and a testing set, and estimating the MSE for a very wide range of parameters: a sequence of λ values ranging from 1 to 5000, with step sizes of 10. We then chose the a range of 101 values of λ that contained the best-performing λ from the initial pass-over. The statistical package we used to estimate LASSO models automatically chose a set of tuning parameters to try. We chose to grow 10,000 trees for RF, since $p \gg n$, and used the function `tuneRF()` from the R package `randomForest` to choose the number of variables to try in each tree.

To ensure that only pre-treatment information informed our modeling choices, and that the models were not tuned more closely to the control sample than to the treated sample, the cross-validation was performed using only unmatched schools. In the Texas samples, only TAKS scores were available for the entire sample of schools, since EOC scores needed to be recorded by hand. Therefore, the CV is based on TAKS passing rates, which may lessen the CV MSE estimates' accuracy when predicting MSEs for EOC scores. Since the prognostic model would need to be fit to TAKS scores, and then used to predict EOC scores, it seems likely that the MSE estimates are biased downward in the Texas sample.

All of the estimators were implemented in R (R Development Core Team, 2011), some using specialized packages. In particular, the LASSO estimate was implemented using the `glmnet` package (Friedman et al., 2010), PCR and PLS regressions were implemented by the `pls` package (Mevik et al., 2011), and RF was implemented by `randomForest` (Liaw and Wiener, 2002).

	TX 8th				TX 9th				IN 8th				IN 9th			
	MSE	SE	R^2	DF	MSE	SE	R^2	DF	MSE	SE	R^2	DF	MSE	SE	R^2	DF
LASSO	42.59	4.21	0.69	101	101.79	8.07	0.75	101	53.30	9.27	0.81	101	80.72	8.12	0.84	101
PCR	42.57	4.46	0.69	101	101.18	7.41	0.75	101	129.18	12.27	0.54	101	244.05	14	0.50	101
Ridge	43.31	3.70	0.69	167	98.58	7.41	0.76	175	112.76	13.20	0.59	307	224.27	30.18	0.54	133
PLS	44.06	4.19	0.68	4	98.10	6.95	0.76	6	101.83	11.96	0.63	58	207.11	16.17	0.58	14
RF*	40.28		0.71	50	87.18		0.79	500	68.27		0.75	500	106.13		0.78	500
5-fold Cross Validation estimates, except (*) OOB error estimates																

Table 4.1 Results from the CV experiment: for each model, LASSO, PCR, Ridge regression or PLS, and for each stratum of the AM dataset, we used 5-fold cross-validation to estimate the optimal tuning parameter λ , and its associated MSE (with standard error SE) and R^2 . We also estimated these quantities (except for the standard errors) for the Random Forest algorithm, using out-of-box error estimates. The table also displays the effective degrees of freedom for each optimal model. For PCR and PLS, this is equal to the number of components in the model; for LASSO this is equal to the number of non-zero coefficients (Zou et al., 2007); for ridge regression, it is the trace of the penalized design matrix and for RF, it is the number of variables considered in each tree.

Table 4.1 displays the best estimated prediction MSE and R^2 that were observed for each model type, for each stratum. Figure 4.3 displays each model’s CV curve in each stratum. In the Indiana strata, LASSO and Random Forests outperformed the rest—this may be related to our observations in Figure 4.2, that a few variables in the Indiana datasets are very highly correlated with outcomes. In contradistinction, all models performed approximately equally well in the Texas strata. We estimated standard errors for all of the cross-validated models, LASSO, PCR, Ridge and PLS, by computing the standard deviation of the estimated MSEs from each of the folds, divided by the square root of 5, the number of folds.⁴ Unlike the predictions in Section 4.2, that fewer principal components would be necessary in Texas than in Indiana, the CV preferred 101 components (the maximum allowed in this CV setup) for PCR in all strata; this indicates that there was enough useful information in the 101st component, even in Texas strata, for that component to be included in an optimal PCR model. That is, apparently, its usefulness outweighed the additional variance it would add, as an additional covariate, to the model’s predictions.

4.5 Test-Sets for Choosing Prediction Models with Less Extrapolation

The theory of cross validation is that the model is validated based on a set of observations which is a simple random sample of the larger training set. If the training set is itself a random sample from a larger population, then the CV validation samples are random samples from the larger population, as well. Cross-validation, then, is well suited to estimate a model’s performance when predicting outcome values for new subjects drawn from this population.

In the Peters-Belson-adjusted propensity score case, however, the training set is the set of untreated subjects which were not included in the propensity-score match, but the model is expected to predict values from the set of matched treated and untreated subjects. It cannot, in general, be assumed that the set of matched observations and the set of unmatched observations are random samples from the same population. Indeed, one of the advantages of propensity-score matching is that it identifies the set of untreated subjects most similar, in relevant ways, to the treated subjects; this advantage is most pronounced when treated subjects, and hence matched controls, are not representative of the entire population. The validation samples from CV are, on the other hand, representative of the rest of the training sample. This all adds up to the fact that conventional CV

⁴This is the default SE estimation method for the CV routine in the R package *lars*, on which we based our CV code (Hastie and Efron, 2011).

estimates of MSE are likely to underestimate a model’s true prediction MSE, when it is applied to matched subjects. On the other hand, if there are outlier schools that were not included in the match, and the matched schools are relatively typical, the CV MSE estimates may be biased upwards.

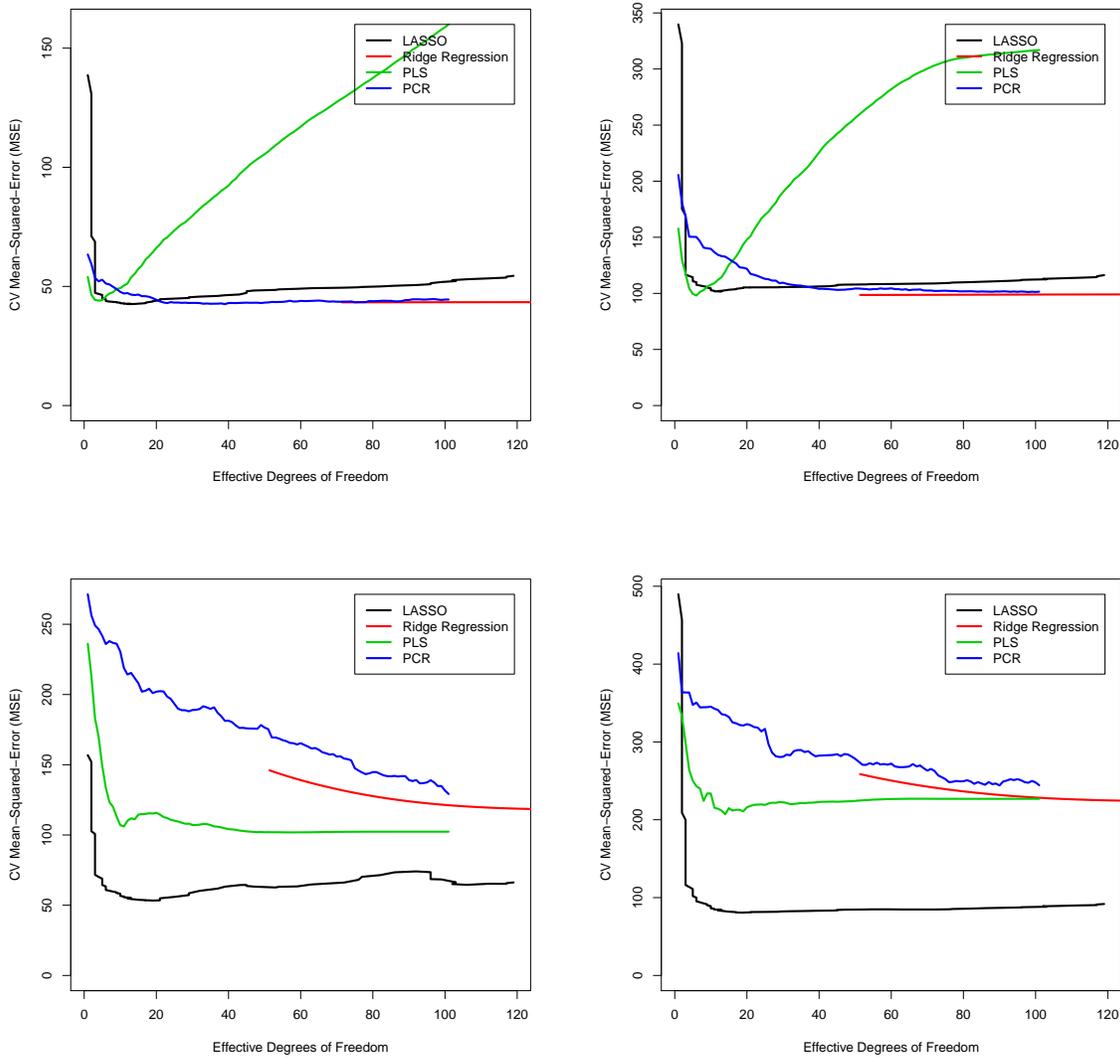


Figure 4.3 CV curves for four models, LASSO, PCR, PLS and ridge regression, in each of the four AM strata, clockwise from top left, Texas middle schools, Texas high schools, Indiana high schools and Indiana middle schools.

Some matching designs may allow for a better—or, at least, less-biased—estimate of the MSE. As matched-sets grow in size, with greater numbers of untreated subjects matched to each treated subject, the power from resultant hypothesis tests increases. However, the marginal returns diminish: after each treated subject is already matched to four or five untreated subjects, adding more untreated

subjects to the matched sets does little to increase power. That being the case, researchers sometimes find it advantageous to limit the sizes of the matched sets, which increases the simplicity, and hence transparency, of the estimation technique. This situation offers analysts an unmatched sample that is statistically similar to the matched sample: the set of untreated subjects who would have been matched to treated subjects, but were not, due to the limit placed on the sizes of matched sets.

In other words, to construct this sample, a statistician would first match subjects without limiting the size of matched sets, and record which control subjects were included in the match. Next, she would match again, this time limiting the size of the matched sets. This divides the sample into four groups: (1) treated subjects, (2) untreated subjects that were included in the original, larger match but excluded from the final trimmed match, (3) untreated subjects included in the final match and (4) untreated subjects that could not be matched to any treated subjects. The untreated subjects that were included in the first match but not in the second (group 2) form a “matching validation” sample, which will presumably be more similar to the matched sample than the rest of the untreated sample. However, the second match will choose, from among the pool of matchable untreated subjects, those subjects which are most similar to the treated subjects; therefore, the matching validation sample will not be strictly representative of the matched sample. In summary, a researcher would fit the candidate models in group (4) and test their performance in group (2); as in the CV case, the actual prognostic model would then be fit using subjects from both groups (2) and (4), and generate predicted Y_C values for subjects in groups (1) and (3).

There is another disadvantage to validating models with the potential matches, relative to cross-validation: CV takes several random samples of the training set, and estimates the MSE from each of them. This maneuver serves to lessen the variance of CV estimates of MSE. In contradistinction, estimates from the matching validation result from only one sample, and will therefore vary more. For this reason, it may be best to use both CV and matching validation, and accounting for the standard error of the MSE estimate when comparing the respective results.

In the Texas strata, outcomes were only available by downloading individual files, one for each school, and copying values by hand from those files. For that reason, it was advantageous to limit the sizes of the matched sets; in this case, only up to 10 untreated schools could be matched to each treated school. The matched sets, however, could have been bigger, without within-set propensity-score differences exceeding the caliper. In the Texas middle school sample, out of a total of 1539 schools, 641 were included in the full match, which was pared down to only 57 schools, leaving 584 schools as a matching validation set. In the high school sample, the full match included 616 schools, and was pared down to 113 schools, out of a total of 1371. These larger matched sets allowed a matching validation exercise. First, all of the schools in the eventual matched

pseudo-experiment were excluded from the sample. Then, each model was fit to the set of schools that were never part of a match, and each model’s prediction was compared to actual values for the schools that were in the full match, but not in the final, pared-down match. The results are in Table 4.2.

	TX 8th			TX 9th		
	MSE	R^2	DF	MSE	R^2	DF
LASSO	38.10	0.73	81	146.32	0.65	81
PCR	41.99	0.70	81	160.45	0.61	81
Ridge	41.11	0.70	137	160.31	0.61	134
PLS	41.60	0.70	6	152.95	0.63	6
RF	38.81	0.72	50	141.46	0.66	500

Table 4.2 A potential-match validation of LASSO, PCR, Ridge regression and PLS models in the two Texas strata. The validation uses TAKS scores as outcomes, since EOC scores are not easily accessible for every school.

It is apparent from Table 4.2 that not much, in terms of prediction accuracy, is lost due to extrapolation; the MSE estimates, and resulting R^2 values, are basically the same.

4.6 Conclusion

Chapter three of this report discussed some advantages to predicting units’ Y_C values from pre-treatment covariates. Those advantages scale with the prediction accuracy of the model. This chapter gave some suggestions for how to choose a modeling strategy to maximize prediction accuracy in this case.

First, an inspection of some of the multivariate properties of the covariate matrix X can yield useful information for building prediction models; in particular, if a small number of covariates seems unusually highly correlated with outcomes, LASSO regression may be an effective modeling technique. Secondly, cross-validation can yield precise, if somewhat biased, estimates of each model’s prediction mean-squared error. Lastly, if the sizes of the matched sets in a matching design were artificially restricted, a researcher can use the units that would have been matched, but were excluded, as a test-set to estimate prediction MSE.

Somewhat surprisingly, there does not seem to be a large amount of extrapolation bias in the MSE estimates for the Texas strata.

The broader significance of this project is that it serves as a bridge between machine-learning,

one of the fastest-growing sub-disciplines in statistics, and causal inference. New machine learning algorithms can help boost traditional propensity-score designs, and possibly correct for residual bias if not all covariates could be included in propensity models.

Chapter 5

Multilevel Propensity and Prognostic Score Analysis: An Evaluation of A School-Wide Curricular Program with School- and Student-Level Data

5.1 Introduction

Quantitative researchers in general, and educational program evaluators in particular, often face multilevel or hierarchical data structures (Gelman and Hill, 2007; Raudenbush and Bryk, 1986). Further, researchers in possession of aggregated group-level data occasionally face the question of whether the collection of individual-level data (“microdata”) is worth the (often quite substantial) cost.

Some questions are impossible, or near-impossible, to answer with only group-aggregated data, such as some subgroup effects. However, given sufficient identification assumptions (such as random treatment assignment or strong ignorability (1.4)) overall average treatment effects are estimable with only group-aggregated variables. Indeed, Jacob et al. (2013) argues based, in part, on simulations, that “public-use aggregate school-level achievement data are, in fact, sufficient to address a wide range of evaluation questions.” That is, the bias and standard errors of effect estimates based on publicly available school-level data are quite similar to those based on individual-(student) level data.

Does this hold true using the setup presented in this report? In an observational study design, assessing a group-level treatment, that combines propensity-score matching with prognostic score/Peters-Belson adjustment, do individual-level data help? Can individual covariates and outcomes improve the accuracy of prognostic predictions in such a way to lessen the bias of an effect

estimate? Is there a way to use individual outcomes to design more powerful test statistics? This paper will attempt to answer both questions.

First, however, a more fundamental question arises: ought researchers, who are in possession of individual-level data, model a treatment assignment as group-level or individual-level? The next section will address this question, and argue that in many cases, a group-level assignment model is most appropriate. Section three will suggest a heuristic for developing a powerful test statistic for a causal effect of a group level treatment that exploits individual-level data, and discuss the use of individual-level covariates in a prognostic model. Section four will apply the methods developed to a preliminary evaluation of a school-level educational program, for which both school-level and individual-level variables are available.¹ The preliminary evaluation will illustrate the use of this report's statistical techniques, as well as a possible advantage to individual-level data.

5.2 Modeling a Multilevel Experiment from Observational Data

How can statisticians incorporate micro-data into an observational study of a group-level treatment? When not every individual in treated groups receives treatment, one possibility is to model the intervention as if it were randomized not at the group level, but at the individual level. This would significantly increase the sample size, and hence, the power of the study. In some cases, therefore, this will be an attractive option. Instead of matching similar groups to each other, and estimating a group-level effect, statisticians can match individuals to each other—possibly constraining the matches to be within-group—and calculate the average difference in outcomes between treated and untreated individuals within each match.

However, this approach involves a different set of assumptions, which in some scenarios may be implausible. First, strong ignorability 1.4 would have to hold for individuals, not for groups. In some cases, individual-level ignorability may be more plausible than group-level ignorability, but in some cases the opposite will be true. In some cases, for instance, it is reasonable to assume that both schools and students within schools choose to adopt a program so as to maximize their expected test scores; this would be a serious challenge to the ignorability assumption if schools can anticipate their potential outcomes with and without the intervention. However, strong ignorability seems more plausible at the school level than at the student level. Students, in consultation with their teachers, may have a reasonably good sense of what educational strategies work the best

¹To make clear that the evaluation is preliminary, and strictly for purposes of illustration, this report will omit the name and any unnecessary description of the treatment program.

for them, based heavily on unmeasured, or unmeasurable, self-knowledge. On the other hand, schools, which are made up of many students, may have a harder time predicting a program's treatment effect. Furthermore, schools' decisions are likely to be based more heavily on measured data than students' decisions; this suggests that the treatment assignment mechanism can be better modeled as a function of measured pre-treatment covariates for schools than for students. Finally, for the portion of the treatment decision that cannot be explained by covariates—that is, variation in treatment assignment within level sets of covariates or propensity scores—there is likely to be more haphazardness in school-level treatment decisions than in student-level decisions. Since schools are, to an extent, bureaucracies, it is likely that factors unrelated to outcomes played important roles in schools' decisions to opt in or out of an intervention.

Additionally, for valid statistical inference, it is necessary to model the mutual dependence of subjects' random treatment assignment. The simplest way to do so is to assume that, by design, one subject in each matched set is randomly selected for treatment, and that treatment assignment is independent between matched sets. Like the first half of this assumption, that treatment assignment is random within matched sets, the second part of this assumption, that treatment assignment is independent across matched sets, may be more plausible if applied to groups rather than to individuals. For instance, in an educational program in which not every student within treated schools participates, it is hardly unreasonable to suppose that students consult each other, and follow each others' examples, when deciding whether or not to opt in.

Following this reasoning, in many cases researchers will choose to model two-level observational data as emerging from a stratified, cluster-randomized experiment, rather than from individual-level randomization. However, individual data can still play a helpful role in three ways: microdata can help researchers design more powerful statistical tests, estimate more precise prognostic scores and account better for individual-level non- or partial-compliance.

5.3 Designing a Powerful Test Statistic Using Microdata

5.3.1 Multilevel Modeling of Group-Level Treatments with Individual-Level Outcomes

A common approach to analyzing multilevel data, especially in an educational context, is by using a multilevel model (Gelman and Hill, 2007) (such as a Hierarchical Linear Model, or HLM; Raudenbush and Bryk 2002; HLMs and multilevel models are types of mixed models; Hartley and

Rao 1967). A standard HLM evaluating a group-level treatment combines two models: one for individuals, and one for groups (of course, there can be more than two levels). Let i index groups, and s index individuals. Then a simple HLM individual-level model would be:

$$Y_{is} = \alpha_i + \beta X_{is} + \varepsilon_{is} \quad (5.1)$$

where X_{is} is a vector of individual-level covariates, and $\varepsilon_{is} \stackrel{iid}{\sim} N(0, \sigma_Y^2)$. The values α_i are group-level intercepts, themselves modeled as random in the group-level model:

$$\alpha_i = \alpha + \delta X_i + \gamma Z_i + v_i \quad (5.2)$$

where α , δ and γ are parameters to be estimated, X_i is a vector of group-level covariates, Z_i is the group's treatment assignment, and $v_i \stackrel{iid}{\sim} N(0, \sigma_\alpha)$. Here, the coefficient γ is the treatment effect.

HLM estimation can be a very useful modeling tool; however, it involves several parametric assumptions, such as parametric distributions of random effects or residuals, which cause some researchers pause. In addition, an HLM-based observational study is not an “experimental” approach, as discussed in Section 1.1.4, and does not share the advantages discussed there, such as small-sample robustness or the objectivity that results from separating design from analysis. However, HLMs can boast some other advantages: for instance, they can accommodate both group- and individual-level covariates, and they adaptively weight observations to maximize estimates' efficiency and tests' power. This section will show how, in an experimental-style propensity score observational study, researchers can adapt some of the power-maximizing properties of HLMs, without committing fully to the HLM modeling assumptions.

As discussed, this report's matching approach models the treatment as group-level, and conceives of a hypothetical group-level experiment. That being the case, the possible treatment effect will be an effect on group-level outcomes; that is, individual-level outcomes aggregated to the group level. The aggregation process is where HLM insights can play a role.

5.3.2 Adapting HLM Reasoning to Propensity-Score Designs: Aggregating Outcomes

Adapting HLM ideas to the Rubin Causal Model necessitates adapting HLM notation. Let Y_{Cis} be the control potential outcome of individual s (for “student”) in group i ; let \vec{Y}_{Ci} be the vector of Y_{Cis} values for a particular group i . Then, for some aggregation function $g(\cdot)$, the potential outcome

for group i is $Y_{Ci} = g(\vec{Y}_{Ci})$. So if, for instance, group-level outcomes are means of individual level outcomes, then $Y_{Ci} = (1/n_i) \sum_s Y_{Cis} = \overline{Y_{Ci}}$.

If only group-level data is available, researchers have little choice in aggregation function $g(\cdot)$; individual data allows for more flexibility. For instance, researchers can choose to estimate a program's average effectiveness on a group's median outcome, or on outcomes for some subgroup of students. These choices can be based on either substantive or statistical concerns; in particular, some aggregation functions $g(\cdot)$ may allow more powerful hypothesis tests, or more precise effect estimates.

We will offer a heuristic for choosing $g(\cdot)$, based on HLMs, that may maximize power. We will consider aggregating functions $g(\cdot)$ which are a school-level constant times the sum of Y values in a school, that is, $g(\vec{Y}_{Ci}) = w_i \sum_s Y_{Cis}$. Two possible values of w_i here are $w_i \equiv 1$, in which case Y_{Ci} would be the sum of individual outcomes in group i , or $w_i = 1/n_i$, in which case Y_{Ci} would be the mean of individual outcomes. It is likely that the function g of this class that maximizes power chooses weights w_i between $1/n_i$ and 1.

The simplest multilevel model for this data setting, with treatment administered at the group level, but outcomes measured at the individual level, excludes covariates X_{si} and X_i , and is

$$Y_{ij} = \alpha_i + \varepsilon_{si} \quad (5.3)$$

with $\varepsilon \stackrel{iid}{\sim} N(0, \sigma_Y^2)$, and level-2 model

$$\alpha_i = \alpha + \tau Z_i + v_i \quad (5.4)$$

with $v_i \stackrel{iid}{\sim} N(0, \sigma_\alpha^2)$. To estimate the treatment effect τ , when σ_Y and σ_α are known, let \bar{Y}_i be the mean outcome in group i . Then regress \bar{Y} on a constant and Z , weighting each observation with weight Δ_i , a function of the size of group i , n_i and the within- and between-group variances, σ_Y^2 and σ_α^2 . Then the treatment-effect estimate is (Raudenbush and Bryk, 2002)

$$\hat{\tau} = \frac{\sum (Z_j - p^*) (\bar{Y}_i - \bar{Y}^*) \Delta_i^{-1}}{\sum (Z_i - p^*)^2 \Delta_i^{-1}} \quad (5.5)$$

with weights

$$\Delta_i = \sigma_Y^2 / n_i + \sigma_\alpha^2 \quad (5.6)$$

and

$$\bar{Y}^* = \frac{\sum_i \bar{Y}_i \Delta_i^{-1}}{\sum_i \Delta_i^{-1}}; p^* = \frac{\sum_i \bar{Z}_i \Delta_i^{-1}}{\sum_i \Delta_i^{-1}}. \quad (5.7)$$

Given that Z is binary and only takes values 0 and 1, the estimate (5.5) becomes

$$\frac{\sum_{Z_i=1} \bar{Y}_i \Delta_i^{-1}}{\sum_{Z_i=1} \Delta_i^{-1}} - \frac{\sum_{Z_i=0} \bar{Y}_i \Delta_i^{-1}}{\sum_{Z_i=0} \Delta_i^{-1}} \quad (5.8)$$

a difference in weighted means between treated and untreated groups, with weights

$$\Delta_i^{-1} = \frac{1}{\sigma_\alpha^2 + \sigma_Y^2/n_i} = \frac{n_i/\sigma_Y^2}{1 + \frac{\sigma_\alpha^2}{\sigma_Y^2} n_i}. \quad (5.9)$$

Since the $1/\sigma_Y^2$ in the numerator is common to all terms in the numerators and denominators of (5.8), the weighted difference in means becomes

$$\frac{\sum_{Z_i=1} \frac{\bar{Y}_i n_i}{1 + \sigma_\alpha^2/\sigma_Y^2 n_i}}{\sum_{Z_i=1} \frac{n_i}{1 + \sigma_\alpha^2/\sigma_Y^2 n_i}} - \frac{\sum_{Z_i=0} \frac{\bar{Y}_i n_i}{1 + \sigma_\alpha^2/\sigma_Y^2 n_i}}{\sum_{Z_i=0} \frac{n_i}{1 + \sigma_\alpha^2/\sigma_Y^2 n_i}}. \quad (5.10)$$

This estimator treats groups as units, and estimates the difference between their aggregate outcomes. However, how their outcomes are aggregated depends the multilevel properties of the data, and, in particular, the ratio σ_α/σ_Y , which is, roughly, the group effect. When $\sigma_\alpha \ll \sigma_Y$, an individual's group membership is not terribly predictive of her individual outcome Y_{is} . Then, $n_i/(1 + \sigma_\alpha^2/\sigma_Y^2 n_i) \approx n_i$, and (5.10) is approximately a comparison of group sums, not means. The intuition is that as group membership matters less, the optimal estimate increasingly ignores grouping and simply compare treated and untreated individuals (though treatment is still modeled at the group level). When $\sigma_\alpha \approx \sigma_Y$, then (5.10) uses group means as group-level outcomes. When $\sigma_\alpha \gg \sigma_Y$, groups are relatively homogeneous, so group membership is highly predictive of an individual's outcome. In that case, conceptually, the multilevel model considers each group's outcome to be its theoretical mean $\mathbf{E}Y_{is}$; the groups' "sample" means are estimates of these expectations. The higher n_i , the better the group-level mean estimates for group i , so groups with low values for n_i are shrunk towards the grand mean more so than groups with high n_i .

Estimating Within- and Between-Group Variance, σ_Y^2 and σ_α^2

In an HLM, σ_Y^2 and σ_α^2 are unknown, and must be estimated; indeed, this is often one of the aims of multilevel modeling (Gelman, 2005). This is often accomplished using maximum-likelihood techniques with the EM algorithm (Dempster et al., 1977) or Fisher scoring (Harville, 1977). However, if a researcher has set aside part of the untreated sample—the subset that was unfit to match—estimation of σ_Y^2 and σ_α^2 is simpler. An estimate of σ_α^2 would be the empirical variance of group means in the unmatched sample, and an estimate of σ_Y^2 would be the average of groups' variances. Granted, for these estimates to be approximately unbiased for the variance components in the matched sample requires a uniformity assumption that may not be true. However, the stakes are relatively low: σ_Y^2 and σ_α^2 are useful for improving treatment effect estimates' efficiency. Miscalculating the variance components will not lead to bias, but may lead to somewhat sub-optimal efficiency.

However, if the outcomes are Peters-Belson adjusted using prognostic scores, as described in Section 3.3.3, then the variance components must take the adjustment into account. The between-group variance σ_α is the variance of the prediction errors $Y - \hat{Y}$, which may be estimated with cross-validation. Similarly, cross validation can be used to estimate within-group variance σ_Y , when the prediction model is designed to estimate individual-level outcomes.

5.3.3 Using Group- and Individual-Level Covariates

Another advantage of multilevel modeling is its ability to incorporate variables that are measured at both the individual and group levels. Without HLM technology, doing so in the regression-modeling paradigm can be vexing. One possible approach would be to combine both sets of covariates in the same OLS model:

$$Y_{is} = \alpha + \beta X_{is} + \delta X_i + \gamma Z_i + \varepsilon_{is} \quad (5.11)$$

This model, however, ignores the grouping structure, which can lead to severely biased standard errors. (Indeed, combining equations (5.1) and (5.2) gives a very similar expression, except that there is additional modeled error due to v_i .) A generalized-least-squares estimate of model (5.11) which allows for within-group correlation is roughly equivalent to a HLM design. Adding group dummy variables to (5.11) will not work either, since these will be exactly collinear with group-level covariates X_i .

Using propensity score matching with prognostic-score Peters-Belson adjustment, however, as in Section 3.3.3, avoids this problem. In prognostic score estimation, prediction accuracy is

important, but uncertainty estimation is unnecessary: inference for treatment effect estimators will emerge from permutation tests based on the modeled hypothetical experiment. Therefore, fitting equation (5.11) as a prognostic model, in the unmatched group, and using it to predict Y_C values for the matched groups, is not problematic. Of course, the propensity-score matching itself already adjusts for some group-level covariates.

5.4 Example Dataset: Preliminary Investigation of A School-Level Educational Program

The educational program that this report will attempt to evaluate, for illustrative purposes, is a new program recently adopted by six rural Kentucky high schools; as will be seen, these schools are fairly typical for Kentucky. Within each school, a subset of students participates in the program; this report will, however, model the treatment at the school level, for reasons discussed in Section 5.2. An instrumental variables approach (see, eg. Angrist et al., 1996; Imbens and Rosenbaum, 2005) would perhaps sharpen the analysis, but is ancillary to the illustrative purposes here, and will be omitted.

Not much is known about why each school chose to adopt the system; however, the choices were made by the schools' boards, and this illustration will assume that the factors that led to the choices were complex and haphazard enough to be modeled as random, conditional on, or within level sets of, some measured covariates. All covariate data came from the Kentucky Department of Education: there is a matrix X of publicly-available school-level data, with each row representing a school and each column a variable. The matrix X contains data on demographics, standardized-test scores, enrollment, and graduation and dropout rates. Many variables contain historical data, going back to as early as 2008. In addition, we have access to approximately 100 student-level covariates, including demographic information (ethnicity, school-lunch status, homelessness, special education status and language knowledge) as well as detailed prior achievement measures from previous years.

The causal modeling assumption is, then, that if we can identify groups of schools which are similar on all of the relevant measured covariates but not on treatment status, that within those groups selection into treatment is random or haphazard. Specifically, we will focus on prior achievement, the proportions of schools' students that are members of minorities, economically disadvantaged or specially-educated, and school location, rural versus urban. These variables have a fairly high potential to confound a causal estimate—for instance, if schools with low prior test scores are more likely than others to adopt the program, but also more likely to exhibit low test scores in the future,

then estimates of the program's effects will be biased downwards. However, our modeling strategy will, hopefully, alleviate those problems: treated schools will be compared only to schools with similar prior testing records. Another concern our strategy will explicitly address is differing test score trends between schools. For instance, if schools whose test scores have been increasing over the past few years are more likely than others to adopt the program, and if those trends continue after the program's adoption, causal estimates will be biased upwards. To address this problem, we will compare treated schools to untreated schools that not only recorded similar test scores immediately prior to treatment, but which also exhibited similar test score trends. The same reasoning holds for important student demographics. In addition, the high-dimensional multivariate techniques discussed in Chapters 3 and 4, and expanded here, will alleviate possible confounding from other measured variables, both student- and school-level. Finally, we will conduct a sensitivity analysis to check our results' sensitivity to confounding from variables which were not measured or included in our dataset.

No variables from later than the 2010-2011 school year are included, since any later variables may have been affected by the presence of the program. There is a small amount of missing data—though none in the most important variables—that was imputed with a random forest algorithm (Stekhoven, 2012). All in all, there are 760 variables for 214 high schools.

The outcomes we will study are student scores on a statewide standardized test.

5.4.1 Multidimensional Analysis: Principal Components Analysis and Fisher-Transformed Bivariate Correlations

We conducted a principal components analysis (PCA) of the school-level dataset to help the propensity scores account for all of the possible measured confounding variables. This process yields a side benefit: the PCA provides graphical information about the multivariate structure of the dataset.

Figure 5.1 shows the cumulative proportion of the total variation in the dataset (the sum of the variances of all of the data's variables) accounted for by each principal component. For instance, the first three components, together, account for a little less than 40% of the variation in the entire dataset. Also, the first 16 principal components account for about 60% of the total variance. After the 16th PC, the marginal contribution of individual PCs decreases rapidly.

Figures 5.2 and 5.3 plot Kentucky high schools according to their values for the first and second, and first and third PCs, respectively. The treated schools are labeled, and the top three positive and negative variables that define each of the plotted PCs are indicated with arrows. The first PC

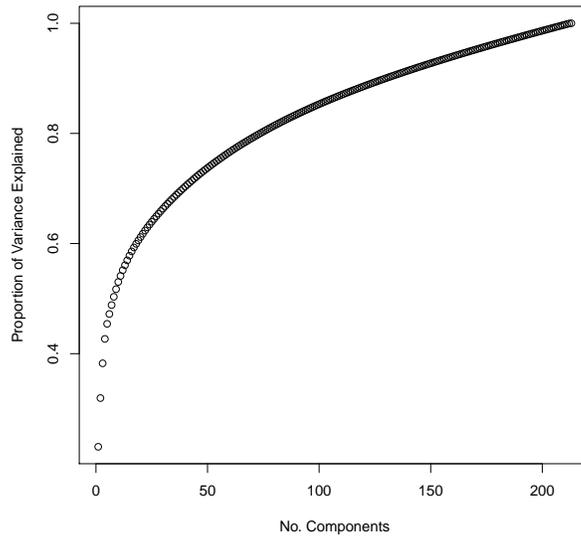


Figure 5.1 The cumulative proportion of the data's total variance accounted for by the first k principal components

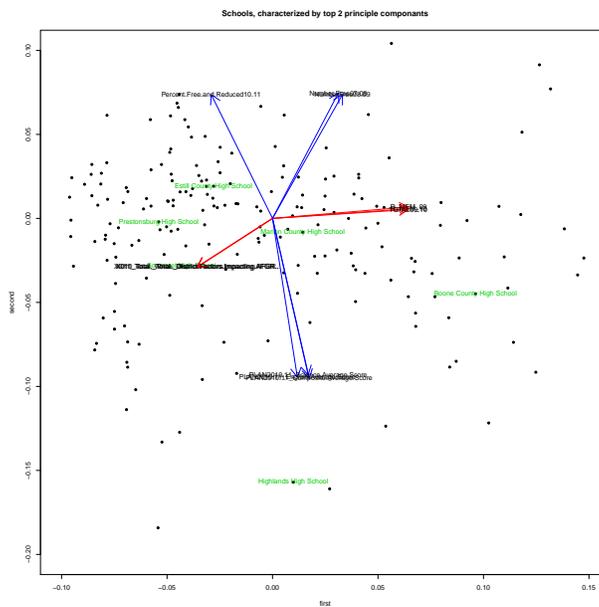


Figure 5.2 Kentucky High Schools, arranged according to the first two PCs. The treated schools are labeled, and the top three positive and negative variables that define the each PC are plotted as arrows. The top positive and negative variables for the first component are in red, and those for the second component are in blue. For clarity, some outlying schools were excluded.

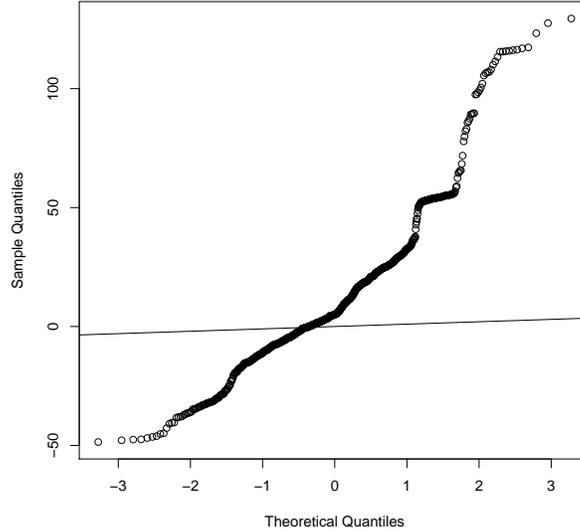


Figure 5.4 Fisher-transformed bivariate correlations between each school-level covariate and the school-mean tests scores.

accommodate anywhere near all of the 760 variables. In the logit model, we included a small subset of variables that seemed most important in the selection process, and for predicting outcomes. In addition, we included two principal components of the entire dataset in the logit model.

The dependent variable in the logit model was whether or not the school was treated (1= it was, 0 it wasn't). The pre-treatment variables included in the model were: percent free or reduced-price lunch (*FreeLunch*), percent White, a dummy variable indicating if the school was urban or suburban versus rural, and prior achievement measures. Prior achievement was measured by 2010 school mean ACT and PLAN scores (*ACTm* and *PLANm*) and estimated mean ACT trends (*ACT.Trend*). The ACT trends are the ordinary-least-squares slope estimates from the school-level regressions of school mean ACT scores from 2007–2010 on a linear time variable. In addition to these variables, we included the two principal components that most highly correlated with spring 2012 ACT composite averages. All told, the logistic model was

$$\begin{aligned} \log - odds(Treatment) = & \alpha + \beta_1 FreeLunch + \beta_2 White + \beta_3 ACTm + \beta_4 PLANm \\ & + \beta_5 SpecialEd + \beta_6 Urban + \beta_7 ACT.Trend + \beta_8 PC1 + \beta_9 PC2. \end{aligned} \quad (5.12)$$

Figure 5.5 shows the linear predictors from the logit model as a function of treatment assignment. The linear predictors are larger for treated schools, indicating that the model predicts treatment

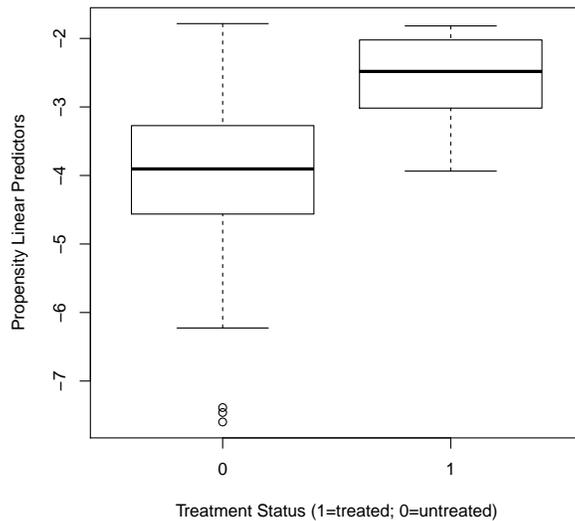


Figure 5.5 The distributions of linear predictors from model (5.12)

assignments, but the distributions of linear predictors overlap between the groups, so propensity score matches should be available (that is, assumption 3.5 holds).

With the fitted values (actually, linear predictors) of the logistic propensity model, we constructed an optimal match, using the `R optmatch` package (Hansen, 2007).

To achieve the covariate balance that we desired, we simultaneously matched on estimated propensity scores and on prior achievement measures: mean ACT scores and ACT trends. To do so, we computed the squared mahalanobis distance—computed using propensity scores and the two prior achievement measures—between each treated school and each untreated school.² The match was selected to minimize the sum of the squared mahalanobis distances within matched sets. Combining propensity scores with other variables in a mahalanobis distance has the effect of emphasizing certain important covariates; because prior achievement measures are the most important covariates to balance, this balancing algorithm puts extra effort into producing a match in which these covariates are balanced. In addition, we imposed a caliper on the match of 0.5 standard deviations of the propensity scores (linear predictors), ensuring that no two matched schools would differ in estimated propensity scores by more than half a standard deviation. We also instructed the matching algorithm to restrict the size of each matched set to five—one treated school matched

²The squared mahalanobis distance for several variables is the sum of squared distances on each of the variables, weighted by the variables' standard deviations and covariance. Weighting by covariance means that if a pair of variables (e.g. PLAN and ACT means) are highly collinear, the pair is down-weighted.

to up to four untreated schools—to ensure that a large enough reservoir was available for later steps in the analysis; specifically, the prognostic score model. For comparison purposes, we also constructed matches based only on the propensity scores and on the propensity scores and on mean ACTs. (Several other configurations were tried, but are not shown here.)

	PS	PS+ACTm	PS+ACTm+ACT.Trends
Marion County High School	2	3	5
Boone County High School	3	3	2
Somerset High School	4	5	3
Estill County High School	5	5	5
Prestonsburg High School	3	3	2
Highlands High School	5	3	5
Effective Sample Size	8.37	8.53	8.13

Table 5.1 Sizes of each of the optimal matches for the following matching schemes: matching on only the propensity scores and matching on combinations of the propensity scores and mean ACT composite scores and ACT trends. The last row displays the matches’ effective sample sizes.

Table 5.1 displays the sizes, and effective sample sizes, of each of these matches.

5.4.3 Evaluating the Match: Balance Tests

One of the objectives, and tests, of propensity score matching is covariate balance. To that end, we tested whether there was a difference in means of each covariate between treated and control members of each match, combined across matches and covariates.

The results of a balance test are on display in Table 5.2. This balance assessment includes several new variables, in addition to the covariates that were included in the propensity-score model: percent Black, Hispanic, Asian, Male, and English-Language-Learners, dropout rate, teachers’ average experience, teachers’ average base salary, the schools’ counties’ 2004 crime rates (FBI data from ICPSR, U.S. Dept. of Justice 2007) and indicators for whether the school is urban, suburban, rural, or located in a town. Overall, the p-value for the null hypothesis that all the balance variables have equal means between treated and untreated schools within each matched set is $p = 0.45$, so the null hypothesis is not rejected; there is no evidence of imbalance. The p-value for overall imbalance, including the entire dataset without stratifying by the propensity-score match, is $p = 0.65$, slightly higher. The match that we prefer, denoted PS+ACT+Slopes in Table 5.2, did produce one statistically significant imbalance, on the schools’ passing rates in the ACT English exam—the treated group was 0.35 standard differences lower than the untreated group. However, the match improved balance, over the unmatched set, on ACT mean scores and ACT trends.

Variable	Unmatched	PS	PS+ACT	PS+ACT+Slopes	
	std.diff	std.diff	std.diff	std.diff	
ACTm	-0.20	-0.26	-0.06	-0.17	
PLANm	0.01	-0.20	-0.08	-0.29	
ACT.Trend	-0.38	0.41	0.35	-0.16	
Urban.BinaryTRUE	0.17	0.00	0.04	0.33	
FreeLunch	-0.27	-0.06	-0.01	0.19	
Black	0.04	0.05	0.14	0.18	
Hispanic	0.20	0.24	0.63	0.30	
Asian	-0.25	-0.09	-0.02	0.05	
Male	-0.29	-0.23	0.02	-0.03	
ACTengB	-0.31	-0.60	-0.22	-0.35	*
ACTmathB	0.07	-0.08	0.17	0.16	
ACTreadB	-0.21	-0.21	-0.08	-0.20	
SpecialEd	-0.56	0.21	0.06	0.00	
ELL	-0.05	0.19	0.29	0.13	*
Urbancity	-0.41	-0.28	-0.14	0.00	
Urbanrural	0.03	-0.12	0.28	0.10	
Urbansuburb	-0.39	-0.24	-0.84	-0.52	
Urbantown	0.59	0.54	0.43	0.28	
dropout	-0.46	-0.78	-0.26	-0.56	
CrimeRate04	-0.09	-0.07	0.01	0.28	
teaching.experiance	0.58	-0.03	0.32	-0.14	
base.salary	0.40	-0.14	0.25	0.20	
PC1	-0.28	0.21	-0.15	-0.17	
PC2	0.01	-0.12	0.20	0.06	
CrimeRate04.NATRUE	-0.19	0.00	0.00	-0.28	

Table 5.2 Covariate balance (standardized differences) for matching, in four scenarios: the unmatched sample, the sample matched using only propensity scores (PS), the sample matched using both propensity scores and ACT prior achievement (PS+ACT) and the match using propensity scores, ACT prior achievement and schools' estimated ACT trends (PS+ACT+Slopes).

5.4.4 Prognostic Models

To test and correct for possible imbalance on the entire set of student- and school-level covariates, we constructed prognostic scores. As described in Section 3.3, the prognostic models were to be fit using data from the schools that were not either treated schools themselves, nor matched to treated schools; so for the model-fitting stage, data from all schools involved in the match were set aside.

In order to choose a model, we first conducted a five-fold cross validation (CV), as described in Section 4.4. We fit four different models: the LASSO, ridge regression, partial least squares

(PLS) and principal components regression (PCR). To illustrate the multilevel structure of the data, we fit each model to three different sets of data: a set containing just school-level covariates, one containing only student level covariates, and one containing all of the data. As described in Section 5.3.3, fitting a model to both student- and school-level covariates is not problematic when the goal is simply prediction accuracy, and when school-level fixed-effects are not included. The goal of the prognostic models is to model school-mean test scores, not individual test scores. That being the case, the CV was conducted in a somewhat unconventional manner. First, in a conventional CV, each case is sampled individually; that is, in a five-fold CV, the individual cases are randomly divided into five groups, each of which has its turn as a test-set, while the others serve to train the model. In this CV, however, the sampling was done on the schools: the schools were randomly divided into five groups to serve as each other’s training and testing sets. Next, the mean-squared-error (MSE) calculations were performed at the school-level: the MSE is the average squared difference between each school’s mean predicted score and actual mean score. Similarly, the prediction R^2 values considered represent the proportions of school-level variance the respective models explain.

	School-Level				Student-Level				Total			
	MSE	SE	R^2	DF	MSE	SE	R^2	DF	MSE	SE	R^2	DF
LASSO	0.56	0.07	0.66	15.00	1.09	0.06	0.34	101.00	0.27	0.03	0.83	101.00
PCR	0.71	0.13	0.57	41.00	1.11	0.07	0.32	101.00	0.44	0.03	0.73	101.00
Ridge	0.55	0.06	0.66	27.00	1.09	0.06	0.34	136.00	0.54	0.08	0.67	234.00
PLS	0.71	0.12	0.57	5.00	1.07	0.06	0.35	68.00	0.43	0.06	0.73	10.00

Table 5.3 Results of a school-level cross-validation study of four modeling strategies, LASSO, PCR, ridge regression and PLS, modeling school-level outcomes. The MSE estimates estimate the MSE of school mean test scores, not individual test scores; similarly, the R^2 s represent the proportions of school-level variation that the models explain.

The results of the CV appear in Table 5.3. In this case, it seems, school-level covariates are much more powerful for predicting school-aggregated outcomes than student-level covariates. For instance, the best prediction MSE using school-level covariates is 0.55 *points*² (SE 0.06), compared to 1.07 (SE 0.06) using student-level covariates—almost twice as high. However, models that combine both levels of data do the best, uniformly; in particular, the LASSO achieved a prediction MSE of 0.27 *points*² (SE 0.03), roughly half of the best MSE using only school-level data. While this somewhat dramatic result comes from one particular case, it is not far-fetched to expect similar gains in other contexts.

The MSEs reported in Table 5.3 are much lower than those in Table 4.1, which compared prediction strategies using data from Texas and Indiana for evaluating Agile Mind. The main explanation for this discrepancy is that Table 4.1 compared predictions of the percentage of students

passing an exam (i.e. meeting a benchmark) rather than average scale scores. This is partly an issue of scale: the ACT PLAN test scores range from 4–32, whereas passing rates range from 0–100. In addition, however, average scale scores tend to be more informative—passing rates rely on an artificial dichotomization—and therefore easier to predict. In some situations, this serves as an additional advantage of microdata: if average scale scores are not publicly available, they are certainly computable from individual outcomes (see Jacob et al., 2013).

	School-Level			Student-Level			Total		
	MSE	R ²	DF	MSE	R ²	DF	MSE	R ²	DF
LASSO	0.83	0.49	36.00	0.36	0.78	48.00	0.30	0.82	84.00
PCR	1.00	0.39	21.00	0.42	0.74	21.00	0.74	0.55	21.00
Ridge	0.84	0.49	39.00	0.40	0.76	126.00	0.73	0.56	273.00
PLS	0.99	0.40	6.00	0.41	0.75	14.00	0.38	0.77	13.00

Table 5.4 A comparison of modeling strategies LASSO, PCR, ridge regression and PLS, trained on school- or student-level data or both, using matching validation, as proposed in Section 4.2

Table 5.4 compares models across various levels of data using matching validation, the validation method proposed in Section 4.2: using schools that were potential, yet rejected, matches for treated schools. In contrast with traditional CV, matching validation estimates smaller MSEs using student-level data than using school-level data; combining the two levels still performs best, but now by a small margin. The reason for this discrepancy is unclear; since standard errors are not available for matching validation MSE estimates, the difference between matching validation and CV may be a result of sampling error. Another explanation is that extrapolation is a larger problem with school-level data than with student-level data. Since the propensity-score match was based on school-level covariates, it is on school-level, rather than student-level, covariates that matched and unmatched schools differ most. So too, the functional relationship between school-level covariates and outcomes may differ more between matched and unmatched schools than the analogous relationships at the student level. The best-performing predictive model was the same regardless of validation technique: the LASSO on the combined data. In fact, the estimated MSE for this model was roughly the same between validation techniques: 0.28 with CV and 0.30 with matching validation.

Following the validation results, the LASSO, with 101 degrees of freedom, using both student and school data, was chosen as the prognostic modeling technique.

5.4.5 Outcome Analysis

With the match selected, and prognostic scores computed (using the LASSO on combined student- and school-level covariates), the design stage of the analysis is complete. What remains is an outcome analysis: testing Fisher’s strict null hypothesis of no effects, estimating ATEs or ETTs and estimating confidence intervals.

The operative model here is stratified, cluster randomization: within matched sets of schools, one school was randomized to the treatment condition, and the rest of the schools were randomized to a control condition. As described in Section 1.1.2, randomness, in this model, is purely a function of Z , the treatment assignment variable, which is a stratified-random sample. For a given test statistic T , with a realized value of t , the p-value testing Fisher’s strict null hypothesis—that the treatment had no effect on any school’s average test score—would be the proportion of possible realizations of Z that would lead to a values of T greater than the realized value, t . T , of course, is a function of both Z and Y ; however, under Fisher’s strict null, Z has no effect on Y , so $T(Z, Y)$ may be computed exactly for any value of Z . In the model here, any one school could have been chosen, with equal probability, from each of the matched-sets. Since the sizes of the matched sets are $\{5, 2, 3, 5, 2, 5\}$, there are $5 \times 2 \times 3 \times 5 \times 2 \times 5 = 1500$ possible random draws of Z , each of which is equally likely. To estimate p-values, we drew randomly from the distribution of Z ; for each random draw z , we computed $T(z, Y)$; since every random draw was equally likely, the estimated p-value was the proportion of random draws whose values of T exceeded the realized test statistic t .

The test statistic $T(Z, Y)$ that we used is the statistic discussed in Section 5.3.2, equation (5.5). (In order to compute two-sided p-values, the test statistic was, more accurately, the absolute value of (5.5).) To compute the weights, estimates for σ_α and σ_Y were extracted from the cross-validation of the LASSO model. Since the model was approximately unbiased in its prediction of outcomes Y , the variance of the predictions is approximately equal to the estimated MSE, so $\sigma_\alpha \approx 0.3$. Within schools, the variance of the prediction error for the optimal LASSO model was calculated as $\sigma_Y \approx 5.4$. Therefore, the ICC is approximately $0.3/(5.4 + 0.3) = 1/19$, and the ratio $\sigma_\alpha^2/\sigma_Y^2 \approx 1/18$.

Table 5.5 displays results using these data and methods. The first row actually more appropriately belongs to the design stage: a hypothesis test treating prognostic scores, which are based entirely on pre-treatment covariates, as outcomes yields an approximate “effect” of -0.17 with a p-value of 0.62 and a 95% confidence interval of (-1.00, 0.69). There is no evidence of confounding from measured, but omitted variables.

The next row conducts a hypothesis test using raw outcomes. The estimate from this row is also insignificant: a point-estimate of an effect of -0.43, $p = 0.30$ and a 95% confidence interval

Outcome	Estimate	Effect Size	p-value	%95 CI
Prognostic (\hat{Y})	-0.17	-0.01	0.62	(-1.00,0.69)
Y (raw)	-0.43	-0.03	0.30	(-1.60, 0.64)
Y (PB adjusted)	-0.26	-0.02	0.42	(-1.1, 0.46)

Table 5.5 Estimates, effect sizes, permutation p-values and 95% confidence intervals for the three different outcomes: prognostic scores (which are not, technically, an outcome, so a null result upholds the model) and unadjusted and adjusted composite average test scores. The test statistic is (5.5). The effect size is the estimate, in points on the test, divided by the overall standard deviation of student scores among the matched group.

of (-1.60,0.64). The central estimate of the preliminary evaluation uses Peters-Belson adjusted outcomes; this estimate is also insignificant: a point estimate of -0.26, $p = 0.42$, with a 95% confidence interval of (-1.1,0.46).

The result from the confidence interval is the most substantively interesting: although it is impossible to tell, from these data, if the program has an effect, these results seem to rule out a large, substantively significant effect. In particular, the Peters-Belson adjusted outcomes give a confidence interval of (-1.1,0.46) points, which severely limits the size of an effect.

Sensitivity Analysis

Could these results be an artifact of hidden bias? To assess this possibility, as in Section 3.6.4, we followed the methodology of Hosman et al. (2010), which parametrizes hidden bias from an unobserved variable U in terms of two parameters: ρ , the partial correlation between U and the outcome (in this case, Peters-Belson adjusted) Y , and T_Z , the standardized association between U and Z . To roughly calibrate T_Z and U , we estimated these parameters for a set of measured covariates, with results displayed in Table 5.6. We estimated sensitivity intervals for omitted variables U with a variety of different ρ and T_Z parameters: a worst-case scenario, with the highest ρ and T_Z parameters among the benchmark variables, and scenarios in which variables similar to prior achievement, race, special-education and free or reduced-price lunch variables were omitted. The substantive conclusions remained roughly intact in each of these analyses: the average effect of the treatment seems to be quite limited.

	TZ	rho	Interval
Worst Case	-1.40	0.30	-0.20 ± 0.8
Prior ACT	-0.40	0.20	-0.20 ± 0.6
% White	-0.60	-0.01	-0.20 ± 0.6
% Free Lunch	0.3	0.2	-0.2 ± 0.6
% Special-Ed	-0.7	0.2	-0.2 ± 0.6

Table 5.6 The estimate’s sensitivity to hidden bias: sensitivity intervals, accounting for the possible omission of a variable similar to prior ACT scores, % White, % Free or Reduced-Price Lunch, % Special Education, in terms of two parameters: T_Z , which measures a variable’s relationship with treatment assignment, and ρ , which measures a variable’s conditional relationship with the outcome Y . We also considered a “worst case” scenario, with the highest values for T_Z and ρ observed among important covariates.

5.4.6 Did Individual-Level Data Help?

To examine the usefulness, in this case, of microdata, we repeated the analysis above using only school-level covariates (we continued to use the school-average scale scores computed from individuals’ scores, under the assumption that these data—the school-level averages—would be obtainable). Table 5.7 contains point estimates, p-values and confidence intervals using the LASSO prognostic model chosen in Table 5.3. Of course, the results remain insignificant; however, confidence intervals are wider: the Peters-Belson-adjusted method, using only school-level covariates, is unable to rule out effects of greater than one point.

Outcome	Estimate	Effect Size	p-value	%95 CI
Prognostic (\hat{Y})	-0.26	-0.02	0.40	(-0.92, 0.57)
Y (raw)	-0.52	0.04	0.36	(-1.76, 0.72)
Y (PB adjusted)	-0.26	-0.02	0.64	(-1.65, 1.18)

Table 5.7 With only school level data, Estimates, effect sizes, permutation p-values and 95% confidence intervals for the three different outcomes: prognostic scores (which are not, technically, an outcome, so a null result upholds the model) and unadjusted and adjusted composite average ACT PLAN test scores. The test statistic is (5.5). The effect size is the estimate, in points on the test, divided by the overall standard deviation of student scores among the matched group.

5.5 Summary

This chapter addressed a common question in educational evaluations: would microdata improve the evaluation? It did so by suggesting some ways that microdata may be exploited in a Peters-Belson

propensity-score design: by computing test statistics based on multilevel modeling logic, and by estimating more accurate prognostic scores.

These methods were illustrated in a preliminary analysis of a new school-level intervention. The analysis here suggests that these interventions have at most a small effect on student achievement, as measured by standardized tests. The microdata that were available slightly reduced the size of the confidence intervals, and ruled out some moderate effects.

Further work could bolster, or add nuance, to the methodological conclusions presented here. Analytical power calculations, along with simulations, could give precise results on the extent of the power gain from microdata. In addition, a dataset with an actual effect may be useful fodder for examining the effectiveness of these methods.

Bibliography

- Atila Abdulkadirođlu, Joshua D Angrist, Susan M Dynarski, Thomas J Kane, and Parag A Pathak. Accountability and flexibility in public schools: Evidence from boston's charters and pilots. *The Quarterly Journal of Economics*, 126(2):699–748, 2011.
- Joseph G Altonji, Todd E Elder, and Christopher R Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. Technical report, National Bureau of Economic Research, 2000.
- J Angrist and G Imbens. [covariance adjustment in randomized experiments and observational studies]: Comment. *Statistical Science*, 17(3):304–307, 2002.
- Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables (Disc: p456-472). *Journal of the American Statistical Association*, 91(434):444–455, June 1996.
- Peter C Austin and Muhammad M Mamdani. A comparison of propensity score methods: A case-study estimating the effectiveness of post-ami statin use. *Statistics in medicine*, 25(12):2084–2106, 2006.
- Ian Ayres. Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of "included variable" bias. *Perspectives in biology and medicine*, 48(1):68–S87, 2005.
- Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473), 2006.
- Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Lasso methods for gaussian instrumental variables models. 2011.

- Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection amongst high-dimensional controls. 2012.
- W.A. Belson. A technique for studying the effects of a television broadcast. *Applied Statistics*, pages 195–202, 1956.
- R.A. Berk and D. Rauma. Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association*, pages 21–27, 1983.
- George EP Box and Norman R Draper. Empirical model-building and response surfaces: Wiley series in probability and mathematical statistics. *Empirical model-building and response surfaces: Willey series in probability and mathematical statistics*, 1987.
- Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- Leo Breiman and Ross Ihaka. *Nonlinear discriminant analysis via scaling and ACE*. Department of Statistics, University of California, 1984.
- M. D. Cattaneo, B. Frandsen, and R. Titiunik. A randomization inference approach to regression discontinuity, 2010. Talk at Society for Political Methodology Summer Meeting.
- M. D. Cattaneo, B. Frandsen, and R. Titiunik. Randomization inference in the regression discontinuity design. Technical report, University of Michigan, 2012.
- Devin Caughey and Jasjeet S Sekhon. Elections and the regression discontinuity design: Lessons from close us house races, 1942–2008. *Political Analysis*, 19(4):385–408, 2011.
- Raghavendra Chattopadhyay and Esther Duflo. Women as policy makers: Evidence from a randomized policy experiment in india. *Econometrica*, 72(5):1409–1443, 2004.
- William G Cochran. Matching in analytical studies*. *American Journal of Public Health and the Nations Health*, 43(6_Pt_1):684–691, 1953.
- R.D. Cook and S. Weisberg. Diagnostics for heteroscedasticity in regression. *Biometrika*, 70(1): 1–10, 1983.
- Thomas D Cook. waiting for life to arrive: a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654, 2008.

- Thomas D Cook, William R Shadish, and Vivian C Wong. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of policy analysis and management*, 27(4):724–750, 2008.
- Richard Correnti, Bendek B. Hansen, and Brian Rowan. Early implementation and student achievement outcomes in texas schools using agile mind algebra 1 services. 2008.
- D.R. Cox. *The Planning of Experiments*. John Wiley, 1958.
- Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.
- S.L. DesJardins and B.P. McCall. The impact of the gates millennium scholars program on the retention, college finance-and work-related choices, and future educational aspirations of low-income minority students. *Unpublished Manuscript*, 2008.
- Terry E Dielman. *Applied regression analysis for business and economics*. Duxbury/Thomson Learning, 2001.
- B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.
- Bradley Efron and Robert Tibshirani. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.
- R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.
- R. A. Fisher. *Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.
- Sir Ronald Aylmer Fisher, Statistiker Genetiker, Ronald Aylmer Fisher, Statistician Genetician, Great Britain, Ronald Aylmer Fisher, and Statisticien Généticien. *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh, 1970.
- Ildiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.
- D. Freedman. From association to causation via regression. *Advances in applied mathematics*, 18: 59–110, 1997.

- David A Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL <http://www.jstatsoft.org/v33/i01/>.
- Gary L. Gadbury. Randomization inference and bias of standard errors. *The American Statistician*, 55(4):310–313, 2001. ISSN 0003-1305.
- Carl Freidrich Gauss. *Theorie der Bewegung*. 1809.
- Andrew Gelman. Analysis of variance why it is more important than ever. *The Annals of Statistics*, 33(1):1–53, 2005.
- Andrew Gelman. Causality and statistical learning¹. *American Journal of Sociology*, 117(3): 955–966, 2011.
- Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.
- Andrew Gelman and David Weakliem. Of beauty, sex, and power: statistical challenges in estimating small effects. *American Scientist*, 97:310–316, 2009.
- Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383, 2008.
- Alan S Gerber and Donald P Green. The effect of a nonpartisan get-out-the-vote drive: An experimental study of leafletting. *Journal of Politics*, 62(3):846–857, 2000.
- P.I. Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses*, volume 2. Springer New York, 2000.
- Sander Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 168(2):267–306, 2005. URL <http://www.blackwell-synergy.com/toc/rssa/168/2>.
- Sander Greenland and James M Robins. Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*, 6(1):4, 2009.

- Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pages 1–12, 1943.
- J. Hahn, P. Todd, and W. Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.
- B.B. Hansen. Optmatch: Flexible, optimal matching for observational studies. *New Functions for Multivariate Analysis*, page 18, 2007.
- B.B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008a.
- Ben B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008b. doi: 10.1093/biomet/asn004.
- Ben B. Hansen and Jake Bowers. Covariate balance in simple, stratified and clustered comparative studies. Technical Report 436, University of Michigan, Statistics Department, 2007.
- Ben B. Hansen and Jake Bowers. Attributing effects to a cluster randomized get-out-the-vote campaign. *Journal of the American Statistical Association*, 104(487):873–85, 2009. DOI: 10.1198/Journal of the American Statistical Association.2009.ap06589.
- Herman O Hartley and JNK Rao. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1-2):93–108, 1967.
- David A Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.
- Trevor Hastie and Brad Efron. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2011. URL <http://CRAN.R-project.org/package=lars>. R package version 0.9-8.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.
- James J Heckman and Richard Robb. Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1):239–267, 1985.
- James J. Heckman, Hidehiko Ichimura, and Petra E. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64 (4):605–654, October 1997.

- James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.
- Carolyn Heinrich, Alessandro Maffioli, and Gonzalo Vazquez. A primer for applying propensity-score matching. Technical report, Inter-American Development Bank, Office of Strategic Planning and Development Effectiveness (SPD), 2010.
- Georg Heinze and Peter Jüni. An overview of the objectives of and the approaches to propensity score analyses. *European heart journal*, 32(14):1704–1708, 2011.
- Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of epidemiology and community health*, 60(7):578–586, 2006.
- Daniel E Ho and Kosuke Imai. Randomization inference with natural experiments. *Journal of the American Statistical Association*, 101(475), 2006.
- A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- P. W. Holland. Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81:945–970, 1986.
- Joel L Horowitz and Charles F Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449): 77–84, 2000.
- Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.
- Carrie A Hosman. *Methods to Control for Overt and Hidden Biases in Comparative Studies*. PhD thesis, The University of Michigan, 2011.
- Carrie A. Hosman, Ben B. Hansen, and Paul W. Holland. The sensitivity of linear regression coefficients’ confidence limits to the omission of a confounder. *Annals of Applied Statistics*, 4(2): 849–870, 2010. ISSN 1932-6157. doi: 10.1214/09-[AnnalsOfAppliedStatistics](#)}315.
- H. Hotelling. The generalization of student’s ratio. *The Annals of Mathematical Statistics*, pages 360–378, 1931.

- G. Imbens and K. Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. Technical report, National Bureau of Economic Research, 2009.
- Guido Imbens and Donald Rubin. Rubin causal model. In Steven N. Durlauf and Lawrence E. Blume, editors, *The New Palgrave Dictionary of Economics*. Macmillan Publishers Ltd, 2008. doi: 10.1057/9780230226203.1466. URL http://www.dictionaryofeconomics.com/article?id=pde2008_R000247.
- Guido Imbens and Tristan Zajonc. Regression discontinuity design with multiple forcing variables. Technical report, Working Paper, September, 2011.
- Guido W. Imbens and Paul R. Rosenbaum. Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 168(1):109–126, 2005.
- G.W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.
- Robin T Jacob, Roger D Goddard, and Eun Sook Kim. Assessing the use of aggregate data in the evaluation of school-based interventions implications for evaluation research and state policy regarding public-use data. *Educational Evaluation and Policy Analysis*, 2013.
- Yevgeniya N Kleyman. *Testing for Covariate Balance in Comparative Studies*. PhD thesis, The University of Michigan, 2009.
- Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint Conference on artificial intelligence*, volume 14, pages 1137–1145. Lawrence Erlbaum Associates Ltd, 1995.
- David S. Lee. Randomized experiments from non-random selection in u.s. house elections. Technical report, Department of Economics, UC Berkeley, 2005. 40 pages; posted at Berkeley Econ website.
- David S Lee. Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697, 2008.
- D.S. Lee and D. Card. Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674, 2008.

- D.S. Lee and T. Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48:281–355, 2010.
- A. M. Legendre. *Nouvelles mthodes pour la dtermination des orbites des comtes*. 1805.
- H. Levene. Robust tests for equality of variances. *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*, pages 279–292, 1961.
- F Li, A Mattei, and F Mealli. Bayesian inference for regression discontinuity designs with application to the evaluation of italian university grants. 2013.
- Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3): 18–22, 2002. URL <http://CRAN.R-project.org/doc/Rnews/>.
- Winston Lin. Agnostic notes on regression adjustments to experimental data: reexamining freedmans critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.
- J.M. Lindo, N.J. Sanders, and P. Oreopoulos. Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics*, 2(2):95–117, 2010.
- J. Mandel. Use of the singular value decomposition in regression analysis. *The American Statistician*, 36(1):15–24, 1982.
- J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, 2008.
- Fabrizia Mealli and Carla Rampichini. Evaluating the effects of university grants by using regression discontinuity designs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(3):775–798, 2012.
- Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.
- Bjrn-Helge Mevik, Ron Wehrens, and Kristian Hovde Liland. *pls: Partial Least Squares and Principal Component regression*, 2011. URL <http://CRAN.R-project.org/package=pls>. R package version 2.3-0.
- James N Morgan and John A Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415–434, 1963.

- Romain Neugebauer and Mark van der Laan. Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference*, 129(1):405–426, 2005.
- J. Neyman, K. Iwaskiewicz, and S. Kolodziejczyk. Statistical problems in agricultural experimentation (with discussion). *Supplement to Journal of the Royal Statistical Society*, 2:107–180, 1935.
- Daryle Alan Olson. The efficacy of select nonparametric and distribution-free research methods: Examining the case of concomitant heteroscedasticity and effect of treatment. 2013.
- P. Oreopoulos. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *The American Economic Review*, pages 152–175, 2006.
- J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge U.P., 2000.
- Judea Pearl. A probabilistic calculus of actions. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 454–462. Morgan Kaufmann Publishers Inc., 1994.
- Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.
- C.C. Peters. A method of matching groups for experiment with no loss of population. *The Journal of Educational Research*, 34(8):606–612, 1941.
- Edwin James George Pitman. Significance tests which may be applied to samples from any populations: Iii. the analysis of variance test. *Biometrika*, 29(3/4):322–335, 1938.
- E.J.G. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.
- J. Porter. Estimation in the regression discontinuity model. *Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison*, pages 5–19, 2003.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL <http://www.R-project.org/>. ISBN 3-900051-07-0.
- Stephen Raudenbush and Anthony S Bryk. A hierarchical model for studying school effects. *Sociology of education*, pages 1–17, 1986.

- Stephen W Raudenbush and Anthony S Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. SAGE Publications, Incorporated, 2002.
- Ronald R Rindfuss, Larry Bumpass, and Craig St. John. Education and fertility: Implications for the roles women occupy. *American Sociological Review*, pages 431–447, 1980.
- J.M. Robins, A. Rotnitzky, and D.O. Sharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In M. E. Halloran and D. Berry, editors, *Statistical Models in Epidemiology, The Environment, and Clinical Trials*, pages 1–94. Springer-Verlag, 2000.
- Joseph P Romano. *Testing statistical hypotheses*. Springer, 2005.
- P. R. Rosenbaum. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society*, 147:656–666, 1984.
- P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39:33–38, 1985.
- Paul R. Rosenbaum. Sensitivity analysis for matching with multiple controls. *Biometrika*, 75:577–581, 1988.
- Paul R. Rosenbaum. Hodges-lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association*, 88:1250–1253, 1993.
- Paul R. Rosenbaum. *Observational Studies*. Springer-Verlag, second edition, 2002a.
- Paul R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002b.
- Paul R. Rosenbaum. *Design of Observational Studies*. Springer Verlag, 2010.
- Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983a.
- P.R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.
- P.R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3), 2002c.

- P.R. Rosenbaum. *Observational studies*. Springer, 2002d.
- P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983b.
- P.R. Rosenbaum and D.B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.
- D. B. Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29:185–203, 1973.
- D. B. Rubin. Assignment to treatment group on the basis of a covariate (corr: V3 p384). *Journal of Educational and Behavioral Statistics*, 2:1–26, 1977.
- D. B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328, 1979.
- D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology; Journal of Educational Psychology*, 66(5):688, 1974.
- D.B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.
- Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36, 2007.
- Donald B. Rubin. For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2(3):808–40, 2008.
- Sebastian Schneeweiss, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun, and M Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)*, 20(4):512, 2009.
- Jeffrey Smith and Petra Todd. Does matching overcome lalonde’s critique of nonexperimental methods. *Journal of Econometrics*, 125(1–2):305–353, 2005.
- Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. The MIT Press, 2000.
- J. Splawa-Neyman, DM Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

- Peter M Steiner, Thomas D Cook, William R Shadish, and MH Clark. The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, 15(3): 250, 2010.
- Daniel J. Stekhoven. *missForest: Nonparametric Missing Value Imputation using Random Forest*, 2012. URL <http://CRAN.R-project.org/package=missForest>. R package version 1.2.
- Elie Tamer. Partial identification in econometrics. *Annu. Rev. Econ.*, 2(1):167–195, 2010.
- Wan Tang and Xin M Tu. *Modern Clinical Trial Analysis*. Springer, 2013.
- D.L. Thistlethwaite and D.T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309, 1960.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- Robert Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.
- Jan Tinbergen. On a method of statistical business-cycle research. a reply. *The Economic Journal*, pages 141–154, 1940.
- Federal Bureau of Investigation U.S. Dept. of Justice. Uniform crime reporting program data [united states]: County-level detailed arrest and offense data, 2004. ICPSR04466-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 2007.
- Wilbert Van der Klaauw. Regression–discontinuity analysis: a survey of recent developments in economics. *Labour*, 22(2):219–245, 2008.
- S. Weisberg. *Applied linear regression*, volume 528. Wiley, 2005.
- P.H. Westfall and S.S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. Wiley-Interscience, 1993.
- H. Wold. Estimation of principal components and related models by iterative least squares. In Krishnaiah PR, editor, *Multivariate Analysis*, pages 391–420. Academic Press, 1966.

- G. Udny Yule. An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades (Part I.). *Journal of the Royal Statistical Society*, 62(2): 249–295, 1899.
- H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.
- Hui Zou, Trevor Hastie, and Robert Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.