1

A comparison of experimental and observational data analyses by J. L. Hill, J. P. Reiter, E. L. Zanutto

For obtaining inferences about causality, randomized experiments are the gold standard. Random assignment of treatments ensures, in large samples, that the background characteristics in the treatment groups are similar, so that comparisons of the groups' outcome variables measure primarily differences in the effects of the treatments. For some causal questions, however, it is not possible to assign treatments to units at random, perhaps for ethical or practical reasons. Typically, such observational studies involve collecting and comparing units from existing databases that have nonrandom treatment assignments. Unlike in randomized experiments, there is no assurance that background characteristics are similar across treatment groups, and simple comparisons of the outcome variables can be confounded by such differences.

Researchers use a variety of methods to deal with confounding in observational studies. One approach is to fit linear regressions that include causally-relevant background characteristics as covariates. Typically, such models include indicator variables for the treatments. Another approach, developed by Rosenbaum and Rubin (1983, 1984) specifically to deal with the problem of confounding in observational studies, is to use propensity scores. Here, the goal is to create two groups of units closely balanced on causally-relevant background characteristics. Importantly, both approaches can mitigate only confounding from observed background variables; the groups still may differ on variables not contolled for in the models.

In this paper, we illustrate the potential efficacy of these types of analyses. The causal question we address concerns the effects on intelligence test scores of a particular intervention that provided very high quality childcare for children with low birth weights. We have data from the randomized experiment performed to evaluate the causal effect of this intervention, as well as observational data from the National Longitudinal Survey of Youth on children not exposed to the intervention. Using these two datasets, we compare several estimates of the treatment effect from the observational data to the estimate of the treatment effect from the experiment, which we treat as the gold standard. This general strategy of evaluating the efficacy of competeting nonexperimental techiques by creating a "constructed" observational study using a randomized experiment was first used by Lalonde (1986). Other studies using the same or similar strategies include Lalonde and Maynard (1987); Fraker and Maynard (1987); Friedlander and Robins (1995); Heckman, Ichimura and Todd (1997); and, Dehejia and Wahba (1999). We also demonstrate the use of propensity scores with data that has been multiply imputed to handle pretreatment and post-treatment missingness. To our knowledge, these other constructed observational studies performed analyses using only units with fully observed data.

In the end, for these data we find that the propensity score approaches yield estimated treatment effects consistent with the effects in the experiment, whereas the regression approach does not. The analyses also illustrate the importance of matching on geographic characteristics, something which can be easily overlooked when using propensity score approaches.

1.1 Experimental sample

Low birth weight infants have elevated risks of cognitive impairment and academic failures later in life (Klebanov, Brooks-Gunn, and McCormick, 1994a,b). One approach to reduce these risks is to provide extraordinary support for the families of low birth weight infants, for example intensive early childhood education for the infants and access to trained specialists for the parents.

To assess the effectiveness of such interventions, in 1985 researchers designed the Infant Health Development Program (IHDP). The IHDP involved randomizing 985 low birth weight infants to one of two groups: 1) a treated group assigned to receive weekly visits from specialists and to attend daily childcare at child development centers, and 2) a control group that did not have access to the weekly visits or child development centers. There were 377 infants assigned to the treated group and 608 assigned to the control group. The IHDP provided transportation to the childcare centers to reduce the risk of noncompliance. More details on the design of the experiment can be found in IHDP (1990), Brooks-Gunn, Liaw, and Klebanov (1992), and Hill Brooks-Gunn, and Waldfogel (2003).

The outcome variable is the infant's score on the Peabody Picture Vocabulary Test Revised (PPVT-R) administered at age 3 or 4. Other outcome variables were analyzed in the experiment, but this is the only outcome measured at the same time point in the IHDP and NLSY. The PPVT-R scores are available for all but 173 infants (17.6%).

There are many background variables associated with PPVT-R scores. We limit the variables in our analyses to those measured in both datasets, but a rich set of variables remains. These include characteristics of the infant's mother measured at the time of the birth of her child: age, marital status, race (Hispanic, black or other), educational attainment (less than high school, high school, some college, completed college), whether she worked during her pregnancy, and whether she received prenatal care. They also include characteristics of the child: sex, whether the child was first born, the birth weight, age of the child in 1990, the number of weeks child was born preterm, and the number of days the child had to stay in the hospital after birth. In addition to these socio-demographic variables, we have geographic data: county level unemployment rates and state indicators. In the experimental data, these variables are all fully observed, except for whether or not the mother worked during pregnancy, which is missing for 50 infants (5.1%).

All missing data are handled using multiple imputation (Rubin, 1987). Imputation methods are described in detail in Section 1.2.

As expected, randomization balances the distributions of the background variables in the treated and control groups. This is evident in the first panel of Table 1.1, which displays the covariates' means and standard deviations across the five imputed datasets for both the treated and control groups. The second panel of Table 1.1 displays similar summaries for the observational study, which is discussed in Section 1.2.

The experimental estimate of the intention-to-treat effect for the intervention relative to the control is 6.39 with a standard error of 1.17. This suggests that the combination of intensive child care and home visits had a significant positive average effect on children's test scores.

1.2 Constructed observational study

We now construct an observational study to assess the same question that the experiment addressed: what is the impact of the IHDP treatment? We use the treated infants from the IHDP as the treatment group, and a sample of infants from the National Longitudinal Survey of Youth (NLSY) as the comparison group. This "constructed" observational study reflects the type of data researchers might have access to in the absence of a randomized experiment.

	Experimental Sample		Observational Sample	
	Control	Treated	Full NLSY	Treated
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Mother				
Age (yrs.)	24.74 (6.11)	24.39 (5.93)	23.76 (3.15)	24.59 (5.93)
Hispanic	0.12 (0.33)	0.10 (0.30)	0.21 (0.41)	0.10 (0.30)
Black	0.54 (0.50)	0.55 (0.50)	0.29 (0.45)	0.53 (0.50)
White	0.34 (0.47)	0.34 (0.48)	0.50 (0.50)	0.37 (0.48)
Married	0.46 (0.50)	0.41 (0.49)	0.69 (0.46)	0.42 (0.49)
No HS degree	0.40 (0.49)	0.45 (0.50)	0.30 (0.46)	0.43 (0.50)
HS degree	0.27 (0.44)	0.28 (0.45)	0.42 (0.49)	0.28 (0.45)
Some college	0.21 (0.41)	0.16 (0.37)	0.19 (0.39)	0.17 (0.37)
College degree	0.12 (0.33)	0.11 (0.31)	0.08 (0.27)	0.13 (0.33)
Working	0.57 (0.50)	0.57 (0.50)	0.62 (0.49)	0.59 (0.49)
Prenatal care	0.95 (0.21)	0.95 (0.22)	0.99 (0.11)	0.95 (0.22)
Child				
Birth weight	1769 (473)	1819 (436)	3314 (604)	1819 (439)
Days in hospital	26.6 (24.7)	23.4 (22.3)	4.47 (7.63)	23.7 (22.6)
Age 1990 (mos.)	56.8 (2.13)	56.8 (2.04)	56.3 (29.1)	56.8 (2.03)
Weeks preterm	7.04 (2.77)	6.91 (2.52)	1.24 (2.18)	6.96 (2.52)
Sex (1=female)	0.51 (0.50)	0.50 (0.50)	0.50 (0.50)	0.50 (0.50)
First born	0.43 (0.50)	0.47 (0.50)	0.42 (0.49)	0.47 (0.50)
Geography				
Unemployment	0.08 (0.05)	0.08 (0.06)	0.09 (0.04)	0.08 (0.05)
Lives in state 1	0.14 (0.35)	0.13 (0.34)	0.01 (0.11)	0.13 (0.33)
Lives in state 2	0.11 (0.32)	0.12 (0.33)	0.02 (0.14)	0.12 (0.33)
Lives in state 3	0.10 (0.30)	0.12 (0.33)	0.05 (0.21)	0.12 (0.32)
Lives in state 4	0.14 (0.35)	0.12 (0.32)	0.02 (0.12)	0.12 (0.32)
Lives in state 5	0.16 (0.37)	0.13 (0.34)	0.06 (0.23)	0.12 (0.33)
Lives in state 6	0.09 (0.28)	0.12 (0.32)	0.04 (0.19)	0.13 (0.33)
Lives in state 7	0.16 (0.37)	0.14 (0.35)	0.09 (0.28)	0.13 (0.34)
Lives in state 8	0.10 (0.30)	0.12 (0.30)	0.01 (0.12)	0.14 (0.34)

Table 1.1: Means and standard deviations for both the experimental and observational studies. Dichotomous variables equal one for "yes" answers and equal zero for "no" answers. Differences in the experimental and observational samples for the IHDP treateds reflect differences due to independent imputation of missing data.

The NLSY is a panel survey that began in 1979 with a sample of approximately 12,000 teenagers who, appropriately weighted, were nationally representative at that time. These participants were interviewed every year thereafter until 1994 and biannually after that. Children of women in the NLSY also have been followed since 1986. Given that the IHDP began in 1985, we restrict our NLSY sample to the 4,511 children born from 1981 to 1989. The IHDP treatment was very intensive and extraordinary, so that the NLSY controls are unlikely to have received similar treatments.

As in the experimental data, the observational data contain missing values. The outcome variable, PPVT-R scores, is missing for 870 infants (19.3%). Twelve of the covariates have missing data, ranging from a minimum of 4 infants (.1%) for mother's education to a maximum of 212 infants (4.7%) for child's birth weight. Most covariates have missing data rates around 4%. Missing data were handled using multiple imputation, as discussed later in this section.

Panel 2 of Table 1 displays the means and standard deviations of the potentially confounding covariates for the treatment group and full NLSY comparison group (we reserve the term "control" for experimental control group). The treated children and the NLSY comparison group look quite different on a number of the covariates measured.

Analyses

We can try to control for differences in the groups' socio-demographic background variables in several ways. One approach is to fit a multiple regression of PPVT-R scores on the background variables, including an indicator variable for treatment; we call this the "Regression" approach. In the Regression approach, when the model describes relationships in the data well, the resulting estimated coefficient of the treatment indicator is a reasonable estimate of the average causal effect of the treatment. However, the estimate can be badly biased when the model fits the data poorly. When the data in the treated and comparison groups have different characteristics, the fitted regression involves extrapolations over much of the multidimensional covariate space (Rubin, 1997). Such violations of model assumptions can be difficult to detect.

A second approach is to match units based on estimated propensity scores to attempt to construct groups balanced on the confounding covariates. Treatment effects can be estimated by differencing the sample averages of the treated and matched comparison groups; we call this the "P-score Direct" approach. Or, they can be estimated by using the treated and matched groups in a multiple regression of the outcome on the confounding covariates and an indicator for treatment; we call this the "P-score Regression" approach. Alternative propensity score approaches, not considered here, include subclassification on propensity scores (Rosenbaum and Rubin, 1984; D'Agostino, 1998; Dehejia and Wahba, 1999) and propensity score weighted estimation (Rosenbaum, 1987; Schneider, Cleary, Zaslavsky, and Epstein, 2001; Hirano, Imbens, and Ridder, 2003).

The P-score Direct approach avoids the specification of regression models for the relationship between the outcome and the covariates. Although models must be fit to estimate propensity scores, estimates of treatment effects are generally less sensitive to misspecification of the propensity score model than the Regression approach is to misspecification of the regression model (Drake, 1993; Rubin, 1997). With close matching on estimated propensity scores, the groups should be balanced on the observed background characteristics. Part of the model-fitting process is checking this balance so that the researcher can discern whether the groups are too different for resulting treatment effect estimates to be trustworthy. Assuming close balance, direct comparisons of the average for the treated group and the average for the matched comparison group should be mostly free of confounding due to the matched variables (Rosenbaum and Rubin, 1983, 1984).

The P-score Regression approach in a sense combines the other two approaches. It is less likely to be subject to extrapolations than the Regression approach, because the treated and matched comparison units are in similar regions of the covariate space. But, it adjusts for slight imbalances in the groups' background characteristics with a regression model, thereby potentially reducing bias and increasing precision (Rubin, 1973, 1979; Rubin and Thomas, 2000).

The Regression approach and the matched-sample approaches estimate different quantities. The Regression approach estimates the average treatment effect across the full sample, whereas the matched sample approaches estimate the effect of the treatment on the treated (IHDP) group. These estimands can differ when the treatment effect is a non-constant function of the covariates, in which case the estimated treatment effects can differ even if each method produces unbiased estimates. In this study, we seek to estimate the effect of the treatment on the IHDP-treated group.

Importantly, both the regression and propensity score approaches work well only when we have controlled for all confounding covariates. When there are important confounding variables that have not been controlled for, either method can lead to biased estimates of treatment effects.

Missing data

Many social science researchers handle missing outcome data by restricting analyses to complete cases, sometimes in conjuction with other fixes such as dummy variables for missing data. This strategy leaves analyses open to bias because there may be systematic differences between the treated and control units with observed outcomes (Little and Rubin, 2002). This is even true in randomized experiments, unless the outcome data are missing completely at random (Frangakis and Rubin, 1999). For the constructed observational study, we therefore do not use experimental complete case estimates as benchmarks when comparing the regression and propensity score matching strategies.

Instead, we handle missing values using multiple imputation (Rubin, 1987). This retains the full sample for calculating intention-to-treat estimates and, under appropriate assumptions, should yield unbiased estimates of the intention-to-treat effect with the experimental data. We note that the complete case estimate of the experimental estimate is 5.7, roughly half a standard error larger than the multiple imputation estimate of 5.1. We also note that, in the observational study, using only complete cases forces us to exlude large numbers of children when implementing the strategies (more than 3000 children for the most comprehensive strategy). As a result, when using only complete cases, all the strategies perform poorly and without distinction.

For the experimental data, we assume the missing PPVT-R scores and mother's work status are missing at random (Rubin, 1976). We believe that the number and breadth of the covariates measured makes this assumption plausible. We then generate multiple imputations from chained regression models (van Buuren, Boshuizen, and Knook, 1999; Raghunathan, Lepkowski, Van Hoewyk, and Solenberger, 2001). The models are fit with the MICE software (`www.multiple-imputation.com`) for `S-Plus`. The imputation model for PPVT-R scores is a linear regression, fit using main effects for all covariates and the treatment indicator, as well as interactions between the treatment variable and all covariates. The imputation model for mother's working status is a logistic regression, fit using all covariates, the treatment indicator, and the outcome variable as predictors. For both models, we include all the main effects and interactions to reduce the risk of generating imputations that are not consistent with the relationships in the data.[1] Five imputations are independently generated for each missing value.

For the observational data, we assume data are missing at random and use MICE to generate five imputations for each missing value, using chained linear, logistic, and polytomous logistic regression models as appropriate. For PPVT-R scores, the linear regression is fit using all covariates and the treatment indicator, as well as all interactions between covariates and the treatment indicator. For all other variables, predictors for the imputation models include all covariates, the treatment indicator, and the outcome variable. The missing at random assumption is more tenuous in the observational sample than in the experimental sample, because of the increase in the number of variables with missing data.

Propensity score analyses are performed in a two step process. First, within each of the five completed datasets, we estimate propensity scores, find a matched control group, and calculate treatment effect estimates and their

[1] Imputing missing data for the purpose of causal analyses is a bit more complicated than standard imputation but the discussion is too detailed for the confines of this paper and will be reserved for future work.

standard errors. Second, we combine these five estimates and their standard errors using Rubin's (1987) combining rules for multiple imputation. Other examples of propensity score analysis of multiply imputed data can be found in Hill, Waldfogel, and Brooks-Gunn (2002) and Hill *et al.* (2003), and its underlying assumptions and potential efficacy are discussed in Hill (2004). Analyses for the Regression strategy are performed in the standard way using Rubin's combining rules.

Results of analyses

We consider several model specifications for the regression and propensity scores, controlling for different background variables. All regression models are of the form $Y \sim \mathrm{N}(X\beta, \sigma^2)$, where X contains covariates. All propensity scores are estimated using the fitted values from the logistic regression of treatment on the same X included in the regressions. Matches for each treated child are determined by finding the NLSY child with the closest propensity score to that child. We use matching with replacement because evidence suggests it can lead to smaller bias than matching without replacement (Dehejia and Wahba, 2002).

The first set of models, labeled DE, controls only for the socio-demographic variables. The second set of models, labeled DE+U, controls for the socio-demographic variables and the unemployment rate of the county the infant resides. Adding unemployment rate should help to control for the economic conditions in which the child was raised. The third set of models, labeled DE+U+S, controls for the socio-demographic variables, the county unemployment rate, and the state the infant was born in. The state variable should help control for differences in the availability and quality of healthcare, childcare, and other services, as well as for differences in lifestyles, across states. Ideally, we would control for county-level effects; however, there are not sufficient numbers of children in our study to do so. The fourth set of models, labeled DE+U+X, controls for the same variables as in DE+U+S but, additionally, performs exact matching on state. That is, each treated child is required to be matched with an NLSY child from the same same state.

Many of the 4,511 children in the full NLSY sample reside in states other than the eight states from the IHDP. We exclude the children from these "other" states when fitting the logistic and linear regressions for the DE+U+S and DE+U+X analyses. This reduces the pool of potential matches to about 1500 children, which could make close matching more difficult. Including children from non-IHDP states, however, forces a linear dependency with the group of treatment children in the logistic regressions if we try to include indicator variables for all states but one, making these models inestimable. An alternative to excluding the children from non-IHDP states is to combine data from two arbitrarily selected states into one category. In this case the estimated propensity scores and the resulting treatment effect estimates depend critically on which states are selected for this combination, which is undesirable. We do not exclude children from the non-IHDP states for the corresponding Regression analyses because it seems unlikely that a researcher unaccustomed to matching would think to do this. Excluding the children from non-IHDP states in the Regression analyses changes the estimates by roughly one quarter of the standard error.

Table 1.2 displays summary statistics reflecting the balance in the covariates for the different logistic regression models. The entries in Table 1.2 are standardized differences between the treated and comparison group means, defined in the caption to Table 1.2. Large absolute values indicate that the means are far apart, whereas absolute values near zero suggest close balance. This metric was used by Rosenbaum and Rubin (1984, 1985) to display covariate balance.

When comparing the treated group to the full NLSY sample, without any matching, we see that the groups' means differ greatly, especially for birth weight and weeks preterm. Matching on socio-demographic variables through propensity scores improves balance considerably, reducing most standardized differences. Matching additionally on unemployment rate does not substantially change balance. Exact matching on state results arguably in the best balance across the spectrum of variables. Exact matching on state gives better balance than simply including state indicators in the propensity score model, which is done by including indicator variables for state in the logistic regression used to estimate the propensity scores.

We now turn to the analysis of PPVT-R scores for each of these models. The point estimates and standard errors of the treatment effects are summarized in Table 1.3 for each analysis. We calculate standard errors for P-score Direct estimates using $\sqrt{Var(y_t)/n_t + \sum_i (w_i/n_c)^2 Var(y_c)}$, where $Var(y_t)$ is the variance of the treated units, $Var(y_c)$ is the variance of the distinct matched control units, and w_i is the number of times matched control unit i is used. We calculate point estimates and standard errors for P-score Regression using weighted least squares, with weights equal to w_i. These variance estimates are somewhat *ad hoc*; however, there are no commonly accepted and statistically validated estimators of treatment effect variances when matching on propensity scores with replacement. This is a subject for future research. Approximate 95% confidence intervals based on these variances are displayed in Figure 1.1.

We treat the result from the IHDP experiment as the target for comparison, since the estimated treatment effect is unbiased with relatively small standard error and the resulting confidence intervals are inferentially valid. The Regression approach, which always uses the full NLSY sample as the comparison group, consistently results in biased estimates of the treatment effect and little overlap with the confidence intervals from the randomized experiment. As we saw in Panel 2 of Table 1.1, the treated group and full NLSY sample infants have very different covariate distributions, so that lin-

Variable	Full NLSY	DE	DE+U	DE+U+S	DE+U+X
Mother					
Age (yrs.)	0.17	0.19	0.23	0.14	0.25
Hispanic	-0.32	-0.07	-0.10	-0.39	-0.34
Black	0.52	0.13	0.04	0.31	0.40
White	-0.27	-0.08	0.04	0.01	-0.11
Married	-0.55	-0.19	-0.07	-0.23	-0.02
No HS degree	0.27	0.08	0.07	0.28	-0.19
HS degree	-0.32	-0.19	-0.20	-0.15	-0.06
Some college	-0.07	-0.03	0.00	-0.43	0.02
College degree	0.15	0.21	0.21	0.35	0.36
Working	-0.06	-0.04	0.01	0.27	0.10
Prenatal care	-0.22	-0.13	-0.17	-0.27	-0.25
Child					
Birth weight	-2.83	0.18	0.17	0.42	0.17
Days in hospital	1.14	0.01	-0.01	-0.69	-0.44
Age 1990 (mos.)	0.03	0.14	0.06	-0.06	-0.09
Weeks preterm	2.43	-0.09	-0.06	-0.90	-0.23
First born	0.10	0.15	0.07	0.03	0.20
Sex (1 = female)	0.00	-0.02	0.01	0.08	0.01
Geography					
Unemployment	-0.06	-0.08	-0.06	0.06	-0.08
Lives in state 1	0.47	0.50	0.48	0.33	0.00
Lives in state 2	0.40	0.42	0.39	0.06	0.00
Lives in state 3	0.26	0.16	0.21	-0.43	0.00
Lives in state 4	0.42	0.46	0.46	0.19	0.00
Lives in state 5	0.24	0.32	0.27	-0.24	0.00
Lives in state 6	0.34	0.30	0.31	0.12	0.00
Lives in state 7	0.14	0.23	0.22	-0.08	0.00
Lives in state 8	0.47	0.44	0.44	0.24	0.00
Method controls for:					
Demographics		X	X	X	X
Unemployment			X	X	X
States				X	X
Exact state match					X

Table 1.2: Summaries of covariate balance in treated and control groups. The entries equal $(\bar{x}_t - \bar{x}_c)/\sqrt{(s_t^2 + s_{0c}^2)/2}$, where $\bar{x}_t$ and $\bar{x}_c$ are the sample means of the treated and comparison groups' covariates, and s_t^2 and s_{0c}^2 are the sample variances of the 377 treated and 4,511 non-treated children's covariates. The common denominator facilitates comparisons of the balance in unmatched and matched comparison groups. DE+U+S includes state in the propensity score model, whereas DE+U+X forces state to be exactly balanced.

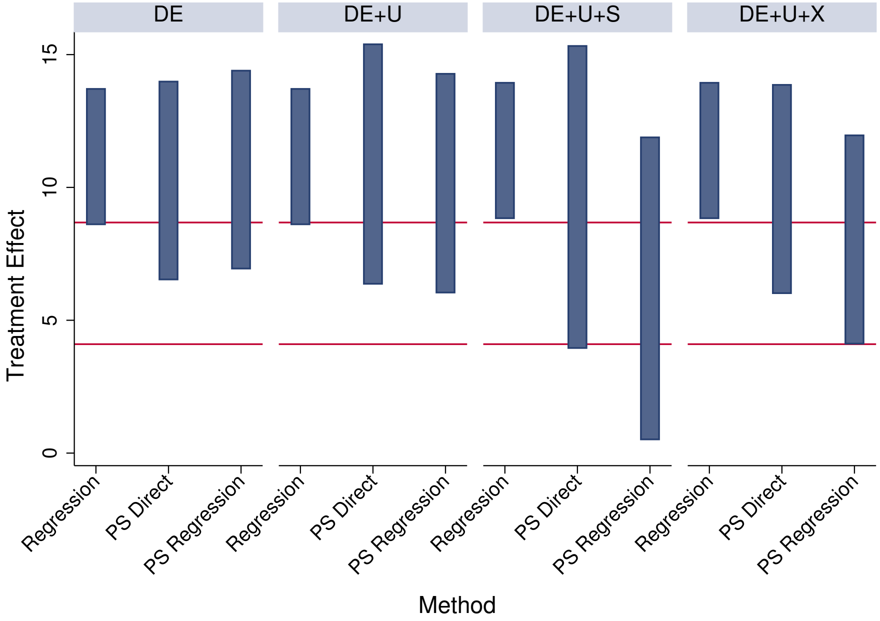

Figure 1.1: The bars represent approximate 95% confidence intervals for the average treatment effect using the various methods. The horizontal lines at 8.68 and 4.10 are the upper and lower limits, respectively, of the 95% confidence interval from the randomized experiment.

ear models fit using the full NLSY sample are especially prone to model misspecification caused by extrapolations. In contrast, once all socio-demographic and geographic variables are included in the matching, the P-score Direct or P-score Regression approaches result in estimates and intervals that more closely track those from the randomized experiment. The P-score Regression approach is better in this case than the P-score Direct approach, most likely because the regression model in the P-score Regression controls for residual imbalances in the covariates due to incomplete matching.

These analyses suggest that P-score Regression is the most effective for this study. However, generalizing this conclusion to say that propensity score matching is always the best approach, or always outperforms regression, would not be appropriate. We obtain reasonable estimates only after including the state variables in the propensity score models. If we had used the analyses based only on the socio-demographic variables, for example if the geographic variables were unavailable due to confidentiality restrictions, it would not

Method	DE	DE+U	DE+U+S	DE+U+X
Regression	11.16 (1.3)	11.16 (1.3)	11.39 (1.3)	11.39 (1.3)
P-score Direct	10.26 (1.9)	10.88 (2.3)	9.64 (2.9)	9.94 (2.0)
P-score Regression	10.67 (1.9)	10.16 (2.1)	6.20 (2.9)	8.04 (2.0)

Table 1.3: Point estimates and standard errors (in parentheses) of treatment effects. The treatment effect for the experiment equals 6.39 with a standard error of 1.17.

have been easy for us to detect that those inferences are so strongly biased, since the socio-economic variables are well balanced for the DE and DE+U propensity score analyses.

The analyses are sensitive to the specification of the model for the propensity scores, as illustrated by the similarities of the results in Table 1.3 until state is included in the models. Additionally, when we restrict the sample to the infants on the higher end of the range of birth weights, who presumably are easier to find matches for than infants on the lower end of the range, we do not find an identical ordering in terms of which method comes closest to the experimental estimate for this subgroup of 7.4. For DE+U+S, the P-score Direct estimate is 11.2, and the P-score Regression estimates is 6.0. For DE+U+X, the P-score Direct estimate of 8.8 is slightly more reliable than the P-score Regression estimate of 5.0. This contrasts with the results in Table 1.3 where the P-score Regression estimate did better across the board.

Since the propensity score analyses appear to outperform the unmatched regression analyses for these data, one might wonder to what extent bias is reduced by limiting the sample to only those control observations most similar to the treated observations, and to what extent bias is reduced by the "reweighting" of the control sample that occurs when matching units. To explore this issue, we perform a regression analysis on a sample that removes all control children who are from non-IHDP states or whose propensity scores are below the lowest propensity score among the treated units. The predictors include the demographic variables, unemployment rate and the state indicators. The estimate from this regression is 8.85 with a standard error of 1.89. This is closer to the experimental estimate than the regressions using the full sample, but not as close as the matched regression-adjusted results. Nonetheless, in these data, it appears that a large share of the bias reduction comes from reducing the sample space to observations that are more similar to each other.

Finally, we illustrate the effect of including the children from "other" states in the DE+U+S and DE+U+X models. If we arbitrarily combine the "other" states with state 8 (Washington) as the baseline for the dummy variables, the estimated treatment effects for the P-score Regression models are 9.1 for the

DE+U+S and 5.9 for the DE+U+X. If we instead combine “other” states with the second to last state (Texas), the P-score Regression treatment effect estimates are 5.8 for the DE+U+S and 8.0 for the DE+U+X. The exact match effects change because the propensity score estimates change, even though afterwards we force exact matches on state. This artificial dependence on the specification of the dummy variables led us to exclude the children in “other” states for the DE+U+S and DE+U+X models.

1.3 Concluding remarks

By comparing the results of an experiment and observational study, we have shown the potential advantage of propensity score approaches over regressions fit using the full comparison sample. Our study also revealed an important finding: it is useful to control for geographic variables. Doing so resulted in estimates from the observational study that more closely matched those from the experiment. This reinforces the importance of controlling for as many variables as possible in a propensity score analysis (Rosenbaum and Rubin, 1983). The sensitivity of these estimates to model specification—all of which led to reasonable balance on the included covariates—suggests that a range of treatment effect estimates should be presented when performing propensity score analyses.