

Causal Inference of Repeated Observations: A Synthesis of the Propensity Score Methods and Multilevel Modeling*

Yu-Sung Su[†]

July 7, 2008

Abstract

The *fundamental problem of causal inference* is that an individual cannot be simultaneously observed in both the treatment and control states (Holland 1986). The propensity score methods that compare the treatment and control groups by discarding the unmatched units are now widely used to deal with this problem. In some situations, however, it is possible to observe the same individual or unit of observation in the treatment and control states at different points in time. The data has the structure that is often referred to as time-series-cross-sectional (TSCS) data. While multilevel modeling is often applied to analyze TSCS data, this paper proposes that synthesizing the propensity score methods and multilevel modeling is preferable.

The paper conducts a Monte Carlo simulation with 36 different scenarios to test the performance of the two combined methods. The result shows that synthesizing the propensity score matching with multilevel modeling performs better in that such method yields less biased and more efficient estimates. An empirical case study that reexamines the model of Przeworski et al. (2000) on democratization and development also shows the advantage of this synthesis.

Keywords: causal inference, balancing score, multilevel modeling, propensity score, time-series cross-sectional data

*Prepared for delivery at the PolMeth XXV, 25th Annual Summer Meeting, University of Michigan, Ann Arbor, July 10–12, 2008. Copyright by the Society for Political Methodology.

[†]Ph.D candidate of the Graduate School and University Center, City University of New York. E-mail: ys463@columbia.edu

1 Introduction

Casual inference concerns what would have happened to an outcome if it was treated with an intervention (treatment). Causal effect is the difference of the responses between the states of being treated and non-treated. Hence, causation involves counterfactuals (Lewis 1973, 557). As the treated or the control units serve as the counterfactual to each other, to gauge the causal effect, we encounter the *fundamental problem of causal inference* that an unit cannot be simultaneously observed in both the treatment and control states (Holland 1986, 947). This is also a missing data problem because either response states of treated or of control are missing in the data (Rosenbaum and Rubin 1983, 41). The best solution to this problem is to have identical twins and give treatment to one of them. Thus the effect of treatment is the difference between the responses of the two groups. Suffices to say that the goal of making a valid causal inference is to find identical twins. Nonetheless, in reality it is very unlikely and costly to achieve this goal.

Matching method that compare the treatment and control groups by discarding the unmatched units is now widely used to as a way to deal with this problem. The basic idea is that since we cannot compare treatment and control outcomes for the same units, we try to compare them with similar units. To attain this, we find control units that are similar to the treated units based on some pre-treatment characteristics (covariates). In turns, we can claim to have a comparable sets or have quasi-twins if our treated and control units have similar covariates. Nonetheless, it could be computational intensive to match on each covariate if the number of covariates is huge (high dimensions). Rosenbaum and Rubin (1983) introduce a propensity score matching that matches units with a vector of scalars (propensity score), which simplifies a high dimensional matching into a uni-dimensional one.

Nevertheless, causal study inherently involves dealing with multilevel data structure. For instance, in some situations, it is possible to observe the same individual or unit of observa-

tion in the treatment and control states at different points in time. Thus we compare treatment and control outcomes for the same units. This kind of data is of multilevel and is often refer to as time-series cross sectional (TSCS) data, longitudinal data or panel data. If the assignment of treatment has nothing to do with the time, then it is plausible to use this kind of data to predict the counterfactual outcome of a unit. Additionally, research designs such as blocking, stratification or subclassification involve grouping similar observations with different treatment assignment within groups. The data also has the multilevel structure.

Henceforth, while multilevel modeling is often applied to analyze this type of data (Goldstein 1995; Western 1998; Steenbergen and Jones 2002; Gelman and Hill 2006), synthesizing propensity score methods and multilevel modeling in the study of casual effect shall be preferable; yet relevant works are rare. In this paper, I propose that we should combine the propensity score methods with multilevel modeling. I extend the current methods of estimating propensity score (Rosenbaum and Rubin 1983; Joffe and Rosenbaum 1999; Imbens 2000; Lu, Zanutto, Hornik, and Rosenbaum 2001; Hirano and Imbens 2004; Zanutto, Lu, and Hornik 2005) to the data with multilevel structure by incorporating multilevel modeling. Since balancing confounding covariates, which relies on the quality of propensity score, is a crucial condition to obtain a valid causal effect, I argue that applying multilevel modeling in the stage of estimating propensity score is preferable because it reduces bias of the estimates (propensity score). Moreover, multilevel modeling is also advantageous in the stage of estimating the treatment effect because such method pools variance components from subgroups; hence the estimate of treatment effect is more efficient than that of mean comparison between subgroups or that of a simple regression.

The paper is organized as follows. In Section 2, I describe the basic setting of multilevel modeling. Next I review methods of estimating propensity score and incorporate the methods with multilevel modeling. In Section 3, I conduct Monte Carlo experiments on

simulated multilevel data under 36 different scenarios. I demonstrate the estimators obtained by the synthesis of propensity score matching and multilevel modeling yields more efficient and less biased estimates. In Section 4, I reexamine the relationship between the development (the treatment) and democratization (the outcome) using the data of Przeworski et al. (2000). In addition to the proclaimed properties in Section 3, I illustrate the way in which multilevel modeling is useful in obtaining varying causal effects over subclasses as the function of propensity score. Finally I make some concluding remarks in Section 5.

2 Methodology and Theory

2.1 The Basic Framework of Causal Inference

Since the series of introductory works by Rubin (1974, 1978, 1980), Rubin constructed a formal framework of causal inference. The basic idea is as follows: Let Y denote the outcome variable and let the binary variable Z indicate whether a unit is treated ($Z = 1$) or non-treated ($Z = 0$). In an experiment study where the assignment of treatment Z is random, the treatment effect τ is just the difference in the responses between the treated and non-treated, $E(\tau) = E(Y|Z = 1) - E(Y|Z = 0)$. However, in most of observational studies, the treatment assignment to units is not random. As a result, treatment groups may differ systematically with respect to some confounding covariates \mathbf{X} . Therefore, the treatment group may not be directly comparable. Methods that group units into subclasses based on these confounding covariates \mathbf{X} can ameliorate these systematic differences and attain the assumption of ignorable treatment assignment (Rosenbaum and Rubin 1983, 43). This is to say that the distribution of the potential outcomes, (Y^0, Y^1) , is the same across levels of treatments, Z ,

once we condition on confounding covariates \mathbf{X} (Gelman and Hill 2006, 183),

$$Y^0, Y^1 \perp Z \mid \mathbf{X}$$

Nonetheless, if the number of covariates is large (high dimensions), it could be computational intensive and difficultly to find matches or to do subclassification.

2.2 The Propensity Score Functions

Rosenbaum and Rubin (1983) propose to estimate the effect of treatment based on a single-index variable—the propensity score. They define propensity score as the conditional probability of receiving a treatment given pre-treatment characteristics \mathbf{X} . The rationale of this approach is to create a comparison group of non-treated units that resemble the group of treated unites. And the criterion of resemblance is that both groups have similar distribution of a scalar function of covariates—the propensity score $e(\mathbf{X})$.

$$Y^0, Y^1 \perp Z \mid e(\mathbf{X})$$

2.2.1 The Case of Binary Treatment

If the treatment is a binary variable, the propensity score can be estimated through a logistic regression as:

$$\Pr(Z_i = 1) = \text{logit}^{-1}(\mathbf{X}_i\beta), \quad i = 1, \dots, n$$

where $Z_i = \{0, 1\}$ is the indicator of exposure to treatment and \mathbf{X} is the multidimensional vector of pre-treatment characteristics. Rosenbaum and Rubin (1983) show that if the exposure to treatment is random within cells defined by \mathbf{X} , it is also random within cells defined by the values of the mono-dimensional variable $e(\mathbf{X})$. As result, given a population of units

denoted by i , if the propensity score $e(\mathbf{X})$ is known, the average effect of treatment on the treated (τ_{ATT}) can be estimated by comparing means of groups, formalized as follows:

$$\begin{aligned}\tau_{\text{ATT}} &= E(Y_i^1 - Y_i^0 \mid Z_i = 1) \\ &= E\{E[Y_i^1 - Y_i^0 \mid Z_i = 1, e(\mathbf{X}_i)]\} \\ &= E\{E[Y_i^1 \mid Z_i = 1, e(\mathbf{X}_i)] - E[Y_i^1 \mid Z_i = 0, e(\mathbf{X}_i)] \mid Z_i = 1\}\end{aligned}$$

where the outer expectation is the distribution of $(e(\mathbf{X}_i) \mid Z_i = 1)$; and Y_i^1 and Y_i^0 are the potential outcomes in the two counterfactual situations of treatment and no treatment respectively. Similarly, we can derive the average effect of treatment on the controlled (τ_{ATC}). The average treatment effect (τ) can be computed by the weighted average of the τ_{ATT} and the τ_{ATC} .

$$\tau = \frac{n_1 \tau_{\text{ATT}} + n_0 \tau_{\text{ATC}}}{n_1 + n_0}$$

Equivalently, we can estimate the average treatment effect τ in the regression as:

$$Y_i = \tau Z_i + \mathbf{X}_i' \boldsymbol{\beta} + \epsilon_i$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients of \mathbf{X} .

Accordingly, Rosenbaum and Rubin (1983) argue that for a given propensity score, exposure to treatment is random and therefore treated and control units should be on average observationally identical. Matching can be treated as a data pre-processing procedure. We discard unmatched data and only draw a causal inference within the matched ones. In other words, we compare cases that are more or less alike on the confounding covariates. So the only plausible reason for difference between treated and non-treated cases is the treatment. In addition, Cochran (1968) and Rosenbaum and Rubin (1984, 1985) demonstrate in their simulation studies that matching and subclassification with the propensity score can effec-

tively remove bias and balancing with the data.

2.2.2 The Cases of Non-Binary Treatment

Estimating the propensity score of non-binary treatment is similar to that of binary treatment. Nonetheless, the method relies on the notion of the balancing score. Rosenbaum and Rubin (1983) define the balancing score $b(\mathbf{X})$ as “a function of the observed covariates \mathbf{X} such that the conditional distribution of \mathbf{X} is the same for treated ($Z = 1$) and control ($Z = 0$) units” (42). In fact, because by definition, a balancing score must be finer than the propensity score; thus the propensity score is also a balancing score (43).

In general, the propensity score for a non-binary treatment can be expressed as:

$$e(\mathbf{X}_i) = \phi_\psi(Z_i|\mathbf{X}_i)$$

where ψ parameterizes this distribution.

With a continuous treatment, we can estimate $e(\mathbf{X}_i)$ with a linear regression:

$$\phi_\psi(Z_i|\mathbf{X}_i) \sim \mathcal{N}(\mathbf{X}'_i\boldsymbol{\beta}, \hat{\sigma}^2)$$

where $\psi = (\boldsymbol{\beta}, \hat{\sigma}^2)$, the coefficients and the residual errors are estimated from the linear regression.

With an ordinal treatment, we can estimate $e(\mathbf{X}_i)$ with an ordered logistic regression (McCullagh 1980; Congdon 2005):

$$\phi_\psi(Z_i|\mathbf{X}_i) = \text{logit}^{-1}(\mathbf{X}'_{i,k}\boldsymbol{\beta} - c_k) - \text{logit}^{-1}(\mathbf{X}'_{i,k-1}\boldsymbol{\beta} - c_{k-1})$$

where k is the level of treatment and $\psi = (\boldsymbol{\beta}, c)$, the coefficients and cut points are estimated

from the ordered logistic regression.

With a categorical (unordered) treatment, we can estimate $e(\mathbf{X}_i)$ with a multinomial logistic regression (Congdon 2005):

$$\phi_\psi(Z_i|\mathbf{X}_i) \sim \text{Multinomial}\left(\frac{\exp(\mathbf{X}'_i\boldsymbol{\beta})}{\sum \exp(\mathbf{X}'_i\boldsymbol{\beta})}, 1\right)$$

where $\psi = \boldsymbol{\beta}$, the coefficients are estimated from the multinomial logistic regression.

In any case, we can group our data into subclasses with similar distribution of the propensity score $e(\mathbf{X}_i)$. Hence, we can either compare means of subclasses or apply regression models to obtain the average treatment effect τ . In particular, since a balancing score is finer than the propensity score, we can simply use the balancing score to subclassify the data. In all cases, including the case with binary treatment, the balancing score is $b(\mathbf{x}) = \mathbf{X}'_i\boldsymbol{\beta}$ (Joffe and Rosenbaum 1999; Imbens 2000; Lu et al. 2001; Hirano and Imbens 2004; Zanutto et al. 2005).¹

2.3 Multilevel Modeling

Causal inference with repeated observations has a multilevel data structure, which is often referred to time-series-cross-sectional data (TSCS). In fact, as Gelman and Hill (2006) put it, causal inference using regression has an inherent multilevel structure because although we compare data between units, we are interested in causal inference within units. Nevertheless, Beck and Katz (1995) point out three potential problems of this data structure: (a) serial correlation: the response of a unit in a current state most likely affects the response of the same unit in the next state; (b) contemporaneous correlation: an outcome of a unit in a time point is very like to relate to outcomes of other units in the same time points; (c)

¹ Imbens (2000) and Hirano and Imbens (2004) use a generalized propensity score \hat{R}_i to subclassify the data with continuous treatment, where $\hat{R}_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{1}{2\hat{\sigma}^2} (Z_i - \mathbf{X}'_i\boldsymbol{\beta})\right]$

heteroscedasticity: there are some unobserved difference between groups.

In the past several years, scholars have proposed regression with “panel corrected standard errors” (PCSE) for performing TCSC data analysis (Beck and Katz 1995, 2001). But PCSEs cannot account for unit effects. Fixed effects (FE) estimators can easily deal with unit effects, but Beck and Katz (2001) point out that FE specifications will often have the consequence of dropping substantially significant time-invariant (or that vary slowly over time) predictors. FE is also problematic because they are quite costly in terms of degrees of freedom. Moreover, Gelman and Hill (2006) and Beck and Katz (2007) advocates that scholars should acknowledge the heterogeneity among units and should not be so naive as to assume that all countries (or units) are identical.

In their simulation study, Shor, Bafumi, Keele, and Park (2007) show that multilevel modeling (MLM) offers a more general and flexible approach to modeling TSCS data that can address its characteristic nuisances. It also outperforms other methods in terms of bias reduction and estimate efficiency. Moreover, Western (1998) shows that MLM provides: (1) more accurate forecasts than other models; (2) more accurate estimates of time-series effects than un-pooled analysis; and (3) more realistic accounting of uncertainty than conventional pooled analysis. In fact, as Gelman and Hill (2006) put it, MLM is a generalization of standard regression techniques (e.g, with a simple linear regression, we just assume the group level variance to be null), and can generally expected to perform as well, or better. Henceforth, throughout the paper, I will utilize MLM to estimate causal inference with repeated observations.

Nevertheless, we still need two assumptions to estimate the causal effect. The first assumption is that there is no interference between units that the treatment assignment for one unit does not affect the outcome for another. This is also called stable unit treatment values (SUTVA) (Rubin 1980). The second assumption is that conditional on confounding

covaraitees up to a particular time, the treatment assignment is random by time. This is also called sequential ignorability assumption (Rubin 1978; Segal et al. 2007). Under these two assumption, we also assume that there is no serial correlation and contemporaneous correlation in the data. Henceforth, the basic setting for a MLM with J groups, T times is:

$$Y_i \sim \mathcal{N}(\alpha_{s[i]} + \tau_s Z_i + \mathbf{X}'_i \boldsymbol{\beta}, \sigma_y^2), \text{ for } i = 1, \dots, n$$

$$\begin{pmatrix} \alpha_s \\ \tau_s \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu_\alpha \\ \mu_\tau \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho \sigma_\alpha \sigma_\tau \\ \rho \sigma_\alpha \sigma_\tau & \sigma_\tau^2 \end{pmatrix} \right) \text{ for } s = 1, \dots, S$$

This MLM model is a varying intercepts and varying slopes (treatment effects) model. α_j here capture unobserved characteristics of the J groups (unit heterogeneity). This model also allows α_j to be correlated with the treatment effects τ_j , where the correlation is estimated by the parameter ρ . This setting is useful when we try to get a smooth (varying) treatment effects from subclasses. In addition, no modeling on time aspect in this setting as we make the assumptions of the SUTVA and sequential ignorability.

Since the data has a multilevel structure, we shall estimate the propensity score with a MLM, too. In the case of binary Treatment:

$$\Pr(Z_i = 1 | \mathbf{X}_i) = \text{logit}^{-1}(\alpha_{j[i]} + \mathbf{X}_i \boldsymbol{\beta})$$

In the case of continuous treatment:

$$\phi_\psi(Z_i | \mathbf{X}_i) \sim \mathcal{N}(\alpha_{j[i]} + \mathbf{X}'_i \boldsymbol{\beta}, \hat{\sigma}^2)$$

In the case of ordinal treatment:

$$\phi_\psi(Z_i|\mathbf{X}_i) = \text{logit}^{-1}(\mathbf{X}'_{i,k}\boldsymbol{\beta} - c_{j[k]}) - \text{logit}^{-1}(\mathbf{X}'_{i,k-1}\boldsymbol{\beta} - c_{j[k-1]})$$

In the case of categorical (unordered) treatment:

$$\phi_\psi(Z_i|\mathbf{X}_i) \sim \text{Multinomial}\left(\frac{\exp(\alpha_{j[i]} + \mathbf{X}'_i\boldsymbol{\beta})}{\sum \exp(\alpha_{j[i]} + \mathbf{X}'_i\boldsymbol{\beta})}, 1\right)$$

Similarly, We can use the balancing score $\hat{b}(\mathbf{X})$ to match or subclassify the data. In all cases, the balancing score is $\hat{b}(\mathbf{X}_i) = \alpha_{j[i]} + \mathbf{X}'_i\boldsymbol{\beta}$, or $\hat{b}(\mathbf{X}_i) = \mathbf{X}'_i\boldsymbol{\beta}$ for the case of ordinal treatment.

3 Monte Carlo Experiments

In this section, I conduct several Monte Carlo experiments to demonstrate that using MLM in estimating the propensity score and the causal effect is preferable because the estimates are more efficient with less bias. I compare the results of four different methods: (1) MLM without the propensity score matching; (2) the propensity score matching (estimated from a simple logistic regression) with a MLM in estimating the treatment effect; (3) the propensity score matching (estimated from a MLM logistic regression) with a simple linear regression in estimating the treatment effect; (4) use MLM in estimating both the propensity score and the treatment effect, which is also the method this paper proposes.

To simplify the simulation, the response surfaces of the treated and control units are linear and parallel and the distribution of the treated and control units' covariates are completely overlapped. All the simulated response surfaces are generated from linear regressions of the outcome Y on two covariates (X_1, X_2) . Both X_1 and X_2 are drawn independently from $\mathcal{N}(1, 1)$. The linear, parallel responses are generated from $Y^0 \sim \mathcal{N}(X_1 + 2X_2, 1)$ and

$Y^1 \sim \mathcal{N}(X_1 + 2X_2 + 4, 1)$. So the treatment effect is 4. In addition, to test the sensitivity and robustness of the propensity score matching, I assign additional control units' covariates (distractors) which drawn from distributions that do not overlap with the distributions of the treated units' covariates (Hill and Reiter 2006). The distractors are generated from $Y^{*0} \sim \mathcal{N}(X_1^* + 2X_2^* + 4, 1)$ where X_1^* and X_2^* are drawn from $\mathcal{N}(3, 1)$. Hence, without matching, these distractors are supposed to offset the treatment effect and make the effect be zero.

Each simulation has $N = 1000$ observations and is subclassified into $J = 50$ groups with $T = 20$ time points in each group. I assign each group with specific intercepts drawn from $\alpha_j \sim \mathcal{N}(2, 1)$. Thus the data structure is of time-series-cross-sectional with specific group intercepts as the unobservable group characteristics. There are three different distractor assignment mechanism: (1) distractors are randomly assigned to the groups; hence distractors exist in the group level (thus these groups have no treatment assignment); (2) distractors are randomly assigned to the units; hence distractors exist in the individual level; (3) distractors are randomly assigned such that some groups have distractors (distractors in the individual levels) and some have distractors without treatment (distractors in the group levels). Thus the third scenario is the case between the first and second scenarios (see Figure 1 for the display of the data of the first and second scenarios). For instance, in Figure 1a (the second scenario), in the first six sampled groups, the treated and control units are strongly overlapped while the two sampled distractor groups contain only distractors with no treatment units. Likewise, in Figure 1b (the second scenario), distractors are sparse over eight sampled groups.

Additionally, I design three scenarios where the numbers of treated, control and distractors vary (see Table 1 for the outline of the assignment).

Henceforth, there are total of 36 different simulation scenarios. Since the distribution of the treated and control units' covariates are completely overlapped, I use matching without

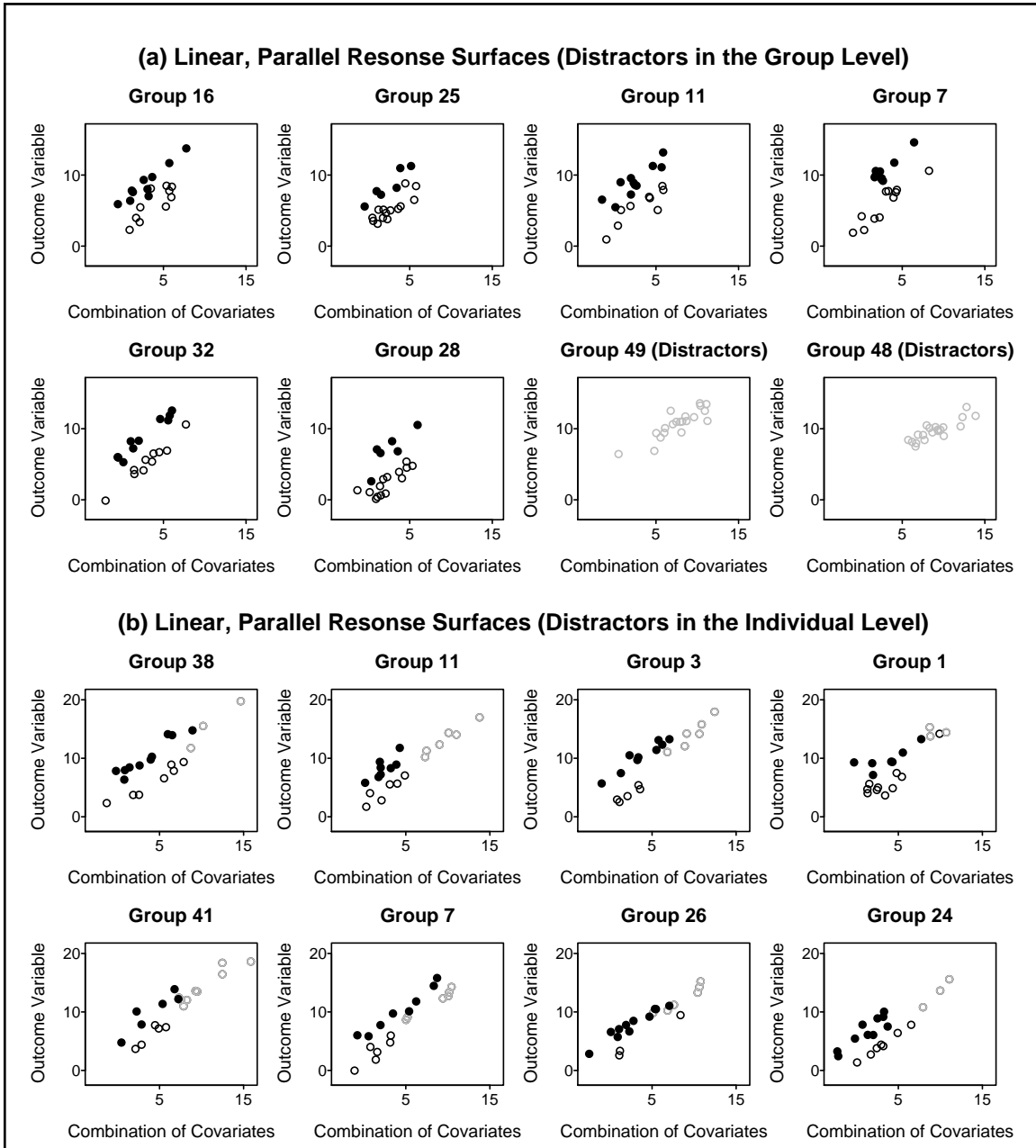


Figure 1: Simulated linear, parallel response surfaces against covariates. Total number of data points is $N = 1000$, where the number of treated (the solid dots) and control units (the dark open circles) are both 350 and the number of distractors is 300 (the light open circles).

	# treated	# true controls	# distractors
Few	450	450	100
Some	350	350	300
Many	100	100	800

Table 1: Outline of three different simulation scenarios of the numbers of treated, control and distractors

replacement to estimate the treatment effects. I run $n_{\text{sim}} = 1000$ simulations on 36 scenarios. The average treatment effect of each scenario is the average of 1000 regression coefficients of the treatment variables such that:

$$\widehat{\tau} = \frac{\sum_{s=1}^{n_{\text{sim}}} \widehat{\tau}_s}{n_{\text{sim}}}$$

Similarly, the standard error of such estimate is the average of 1000 standard errors for $\widehat{\tau}_s$ such that:

$$\widehat{\sigma}_{\tau} = \frac{\sum_{s=1}^{n_{\text{sim}}} \widehat{\sigma}_{\tau,s}}{n_{\text{sim}}}$$

Root mean square error (RMSE) is also calculated for each simulation to see degrees of deviation of the estimated treatment effects, $\widehat{\tau}_s$, away from the true treatment effect τ such that:

$$\widehat{\text{RMSE}} = \sqrt{\frac{\sum_{s=1}^{n_{\text{sim}}} (\widehat{\tau}_s - \tau)^2}{n_{\text{sim}}}}$$

Figure 2 summaries the 36 simulation scenarios. The most important aspect Figure 2 conveys is the comparison among four different methods (the four different rows). Overall, multilevel modeling without matching adjustment on the data performs the poorest in that the estimates are more biased and RMSE's are larger than other three methods. On the other end, using multilevel modeling to estimate both the propensity scores and the treatment effects yields less biased estimates and hence RMSE's are smaller than other methods. Nonetheless, if we only apply multilevel modeling in at least one stage of the analysis (the second and the third row), either in estimating the propensity scores or in estimating the

treatment effects, we get very similar results that the difference is very trivial. Moreover, The simulation results reveal that as long as we do some matching adjustment, the results are not going to differ a lot. Still, applying multilevel modeling to analyze causal inference with repeated observations is recommended because it yields more efficient estimates.

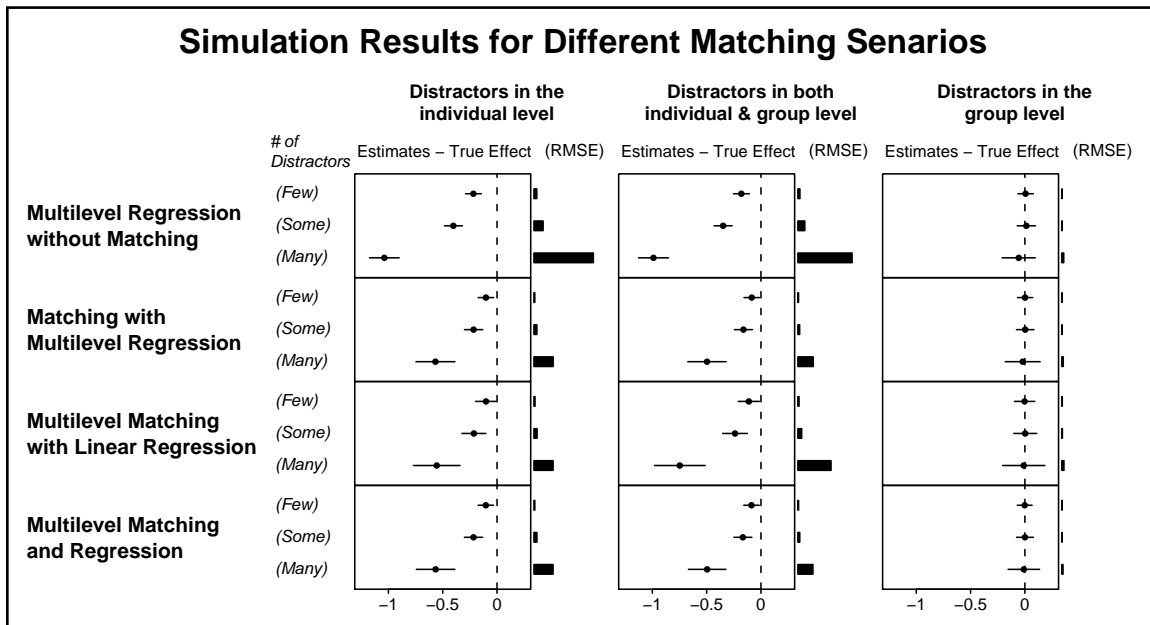


Figure 2: Simulation results of 36 scenarios. Overall, using multilevel modeling both in estimating the propensity score and the treatment effect is better than other alternatives. The estimates of MLM are more efficient and less bias.

The number of distractors (unmatched units) in the data is also a key factor that affects the accuracy of the estimation. Overall, as the number of distractors increases, we tend to get more biased estimates with bigger RMSE. Still, the method of multilevel matching and regression outperforms other alternatives.

Finally, the way in which distractors are assigned also matters. If the distractors are assigned randomly to each group (the column on the left), all four methods fail to yield unbiased estimates of the treatment effects. On the other end (the column on the right), if the distractors are assigned randomly to certain groups such that those groups have no treatment unit but contain only distractors, then all four methods perform similarly that they

yield unbiased estimates. Where if the distractor assignment mechanism lies in between the two aforementioned situations (the column in the middle), the estimates of the four methods are still biased. One way to redress this issue could be to do a two-stage matching. That is to match the data within each group first and match it again between groups.

4 Empirical Application: Democratization and Economic Development Revisited

In this section, I reexamine the model of Przeworski et al. (2000) in estimating the effect of economic development on democratization. In particular, I replicate the model of Table 2.17 of Przeworski et al. (2000). The specific question Przeworski et al. (2000) ask is that do the levels of GDP affect the rise and decline of political regimes (122–123)? To be sure, they did not try to make a causal link between economic development and democratization. I use their data to see if the result would change if we reexamined the data with the propensity score method and multilevel modeling.

The outcome variable is the type of regime of a country in a certain year, coded 1 means dictatorship and 0 otherwise. The treatment variable is GDP per capital in 1000 US dollars. I take log of the treatment variable to facilitate computation and interpretation (Fox 1997; Gelman and Hill 2006). Some confounding covariates are: GDP growth, the sum of number of leadership turnover, religious fractionalization, the percentages of population are Catholic, Protestant or Moslem, an indicator for whether a country was established after 1945, an indicator whether a country was a British colony, the number of previous political transitions, and the percentage of world democracies in a certain year.

Before proceeding to the causal model, I have to check that the data complies with the two basic assumptions of the SUTVA and sequential ignorability. The formal one is unveri-

fiable and thus we have to make the assumption of it. Huntington (1991) claims that there are waves of democratization which implies that a country being a democracy in certain year might have something to do with other country being a democracy in the same time. This is clearly a claim for contemporaneous correlation between units and that there is some kind of interference between units. Hence it is a violation of the SUTVA. Nonetheless, Doorenspleet (2000) shows that Huntington's false claim was merely a result of erroneous data analysis that there is actually no wave of democratization. Accordingly, I will assume the SUTVA hold in this case.

The assumption of sequential ignorability is complied by the model setting. Przeworski et al. (2000) assume that the data obeys a first-order Markov process that the present regime depends only on the regimes during the preceding year, but not beyond. Hence all the information about the lagged values of the exogenous variables is summarized by the lagged values of the outcome variable. In other words, we are estimating the impacts of the GDP on the current regime states that are conditional on the previous regime states. This assumption eliminates the potential problem of serial correlation and thus complies with the assumption of sequential ignorability.

First, I apply multilevel modeling without matching to reexamine model. The multilevel dynamic probit models for transition to dictatorship (Equation 1) and transition to democracy (Equation 2) are formalized as follows:

$$\Pr(y_{ij,t} | y_{ij,t-1} = 0) = \prod [\Phi(\alpha_{j[i]} + \tau^{DA} Z_i + \mathbf{X}'_i \boldsymbol{\beta})]^{y_i} [1 - \Phi(\alpha_{j[i]} + \tau^{DA} Z_i + \mathbf{X}'_i \boldsymbol{\beta})]^{1-y_i} \quad (1)$$

$$\Pr(y_{ij,t} | y_{ij,t-1} = 1) = \prod [\Phi(\alpha_{j[i]} + \tau^{AD} Z_i + \mathbf{X}'_i \boldsymbol{\beta})]^{y_i} [1 - \Phi(\alpha_{j[i]} + \tau^{AD} Z_i + \mathbf{X}'_i \boldsymbol{\beta})]^{1-y_i} \quad (2)$$

where τ^{AD} and τ^{DA} are the treatment effects of log(GDP) on regime types; j is the subscript for country and t is for year. I utilize varying intercepts models (α_j) to capture the country specific effects.

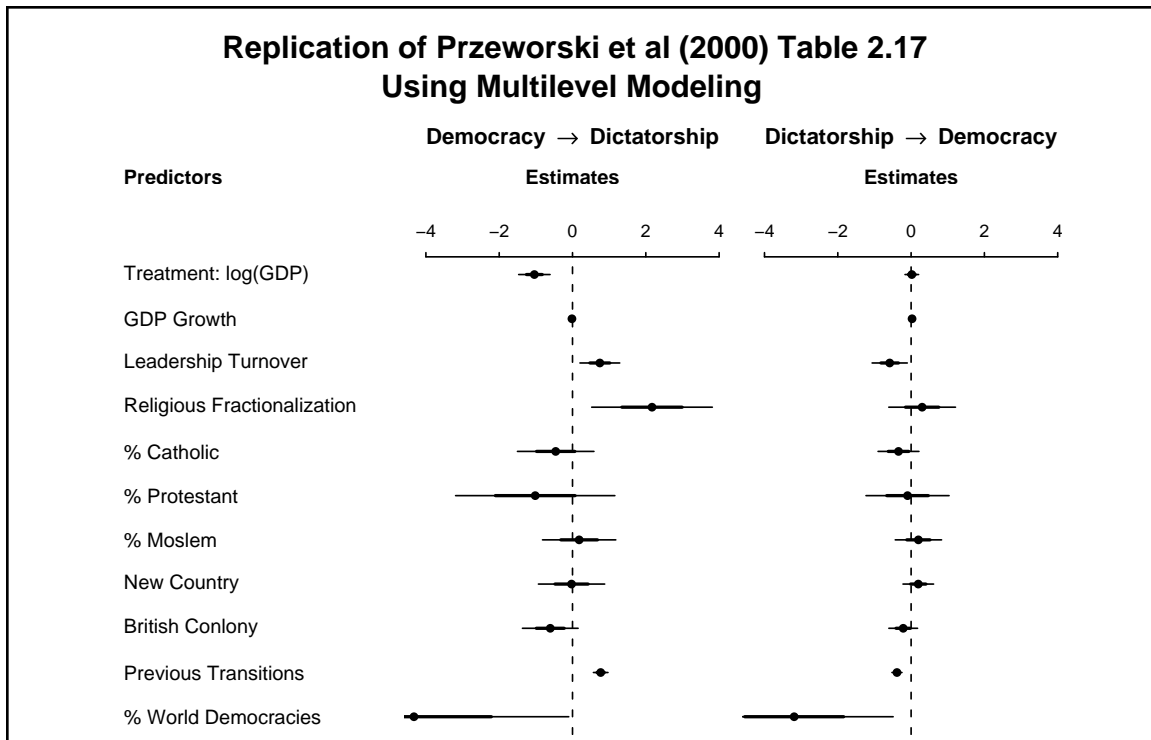


Figure 3: Replicated results of Table 2.17 of Przeworski et al. (2000). The dots indicate the estimates of regression coefficients. The thick and thin bars are ± 1 and ± 2 standard errors, respectively. Overall, the results are very similar to that of Przeworski et al. (2000).

Figure 3 summarizes the regression results (intercepts are omitted to save spaces). The dots represent the point estimates of the predictors, and the thick and thin bars represent ± 1 and ± 2 standard errors of the estimates respectively. Overall, the inference is consistent with that of Przeworski et al. (2000). On average, controlling for other covariates, one percent increase in the treatment (every 100 dollar increase in GDP per capita) results in 16% decrease in transition probabilities to dictatorship. This estimate is statistically significant that the 95% intervals do not cover the zero. Likewise, one percent increase in the treatment results in 0.3% increase in transition probabilities to democracy. But the estimate is statistically insignificant.²

² Epstein et al. (2006) point out the standard errors of the coefficients of the models are not correctly estimated in Przeworski et al. (2000). In this paper, the standard errors are corrected accordingly.

Next, to reduce the potential bias induced by the covariates, I estimate the balancing score of the treatment using multilevel linear regression,

$$Z_i \sim \mathcal{N}(\alpha_{j[i]} + \mathbf{X}'_i \boldsymbol{\beta}, \sigma_z^2)$$

where $b(\mathbf{X}_i) = \alpha_j + \mathbf{X}'_i \boldsymbol{\beta}$ and α_j is the intercept of each country. I subclassify the data into six subgroups with respect to the balancing score (Cochran 1968; Rosenbaum and Rubin 1984), where the first group means the group of the lowest $b(\mathbf{X})$ and hence predicts lower level of the treatment. To evaluate the balance of the covariates, I regress each covariate on the treatment variable, ($Z = \log(\text{GDP})$), using logistic and linear regression for indicator and continuous covariates respectively (Kosuke and van Dyk 2004). Figure 4 shows the standard normal quantile plots of t -statistics for the coefficient of the treatment variable in each regression that predicts each covariate. Figure 4a and Figure 4c demonstrate the lack of balance because the magnitudes of the t -statistics are large. On the other hand, Figure 4b and Figure 4d show the great improvement of balance after controlling for the balancing score. Henceforth, subclassification using the balancing score improve the balance of the distribution of the covariates.

Finally, I apply multilevel dynamic probit models to estimate the treatment effects, (τ^{AD} and τ^{DA}),

$$\begin{aligned} \Pr(y_{ijt} | y_{j,t-1} = 0) &= \prod [\Phi(\alpha_{s[i]} + \tau_s^{DA} Z_i + \mathbf{X}'_i \boldsymbol{\beta})]^{y_i} [1 - \Phi(\alpha_{j[i]} + \tau_s^{DA} Z_i + \mathbf{X}'_i \boldsymbol{\beta})]^{1-y_i} \\ \Pr(y_{ijt} | y_{j,t-1} = 1) &= \prod [\Phi(\alpha_{s[i]} + \tau_s^{AD} Z_i + \mathbf{X}'_i \boldsymbol{\beta})]^{y_i} [1 - \Phi(\alpha_{j[i]} + \tau_s^{AD} Z_i + \mathbf{X}'_i \boldsymbol{\beta})]^{1-y_i} \end{aligned}$$

where $s = 1, 2, \dots, 6$ and is subscripted for the subgroups of the balancing score. The models are the varying intercepts and varying slopes (treatment effects) models (also called the smooth coefficient model). Multilevel modeling is useful here because a simple mean comparison between subgroups cannot be used in a binary outcome case to obtain the treatment

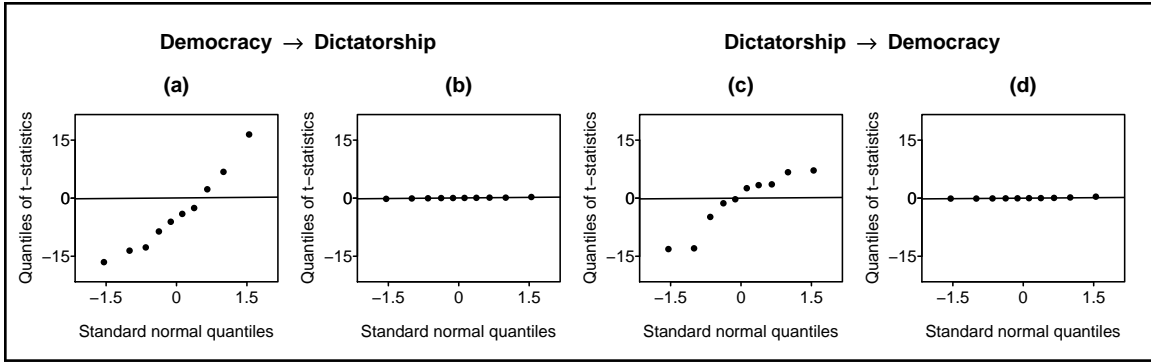


Figure 4: Standard normal quantile plots of t -statistics for the coefficient of $\log(\text{GDP})$ in each regression that predicts each covariate, (a) and (c) without controlling for the balancing score $b(\mathbf{X})$. The magnitude of t -statistics in (a) and (c) demonstrate the lack of balance; where (b) and (d) show the improvement of balance after controlling for $b(\mathbf{X})$. The dots represent the t -statistics for the coefficients and the lines are lines to standard normal quantile plots which pass through the first and third quartiles.

effect. In addition, neither is using separate regression within each subgroup preferable here because in binary outcome cases, there is often a problem of separation and a problem of lack of degrees of freedom. Henceforth, using multilevel modeling that pool the variance components between the individual level and the group level is preferable that the estimates are more efficient (Gelman and Hill 2006).

Figure 5 summarizes the regression results of the two models. For the model of transition to dictatorship (the left panel), on average, after controlling the confounding covariates, 1% increase in the treatment results in 14–21% decrease in the transition probability to dictatorship. The estimates are statistically significant. Once a democratic regime has the balancing score that falls above the fourth subgroup, it is very unlikely for it to decline to dictatorship. Two exceptions are Chile in 1973 of the 5th subgroup and Thailand in 1976 of the 6th subgroup. Likewise, for the model of transition to democracy (the right panel), on average, after controlling the confounding covariates, 1% increase in the treatment results in 0.3% increase in the transition probability to democracy. The estimates have great uncertainty that they are not statistically significant. Nonetheless, once a dictatorship is in the fourth subgroup, it

is very unlikely for it to stay in dictatorship. One exception is Sri Lanka in 1989. Overall, the results of the models with the adjustment of the balancing score do not differ much of those without the adjustment (see Figure 3).

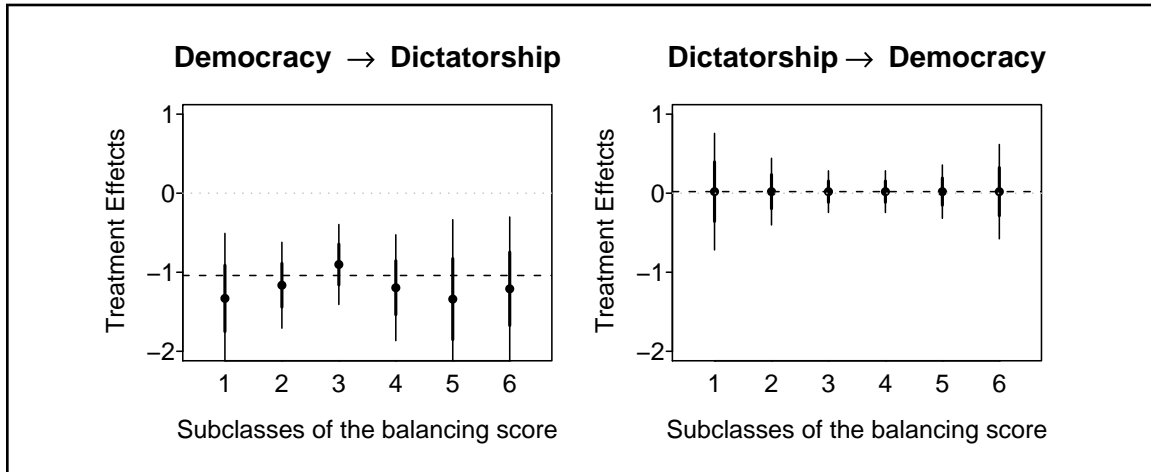


Figure 5: Estimated treatment effects from the multilevel dynamic probit models where the intercepts and the slopes (treatment effects) vary by the six subgroups. The dots represent the estimates of the treatment effects. The thick and the thin indicate the 50% and 95% intervals. The dashed reference lines are the treatment effects of Figure 3.

5 Concluding Remarks

This paper proposes the synthesis of the propensity score methods and multilevel modeling in dealing with causal inference with repeated observations. The Monte Carlo experiment with 36 scenarios demonstrates that applying multilevel modeling both in estimating the propensity score and the treatment effect is proved to be less biased and more efficient. The simulation results also show that using multilevel modeling in one of the two stages, either in estimating the propensity score or in estimating the treatment effect, gives similar results to those of using multilevel modeling in the both stages. The cost of not doing multilevel modeling in the both stages is to have less efficient estimates.

Additionally, although the simulation set up in this paper is in the simplest form that the

relationship of the response surfaces between the treated and control units are linear and parallel, its implication shall hold in more complex cases. Investigators can add in interactions and quadric forms of the covariates and treatment in terms of estimating both the propensity score and the treatment effect.

Nevertheless, the synthesis of the two methods cannot reduce the bias induced by the unmatched units that distribute idiosyncratically within groups. It performs well if certain groups in the population are the unmatched groups. One way to deal with this bias is to perform a two-stage matching, i.e., first matching within groups and secondly matching between groups.

Moreover, the paper also demonstrates the way in which multilevel modeling is preferable in cases where outcomes are binary or categorical. In these cases, mean comparison methods are either inappropriate or unable to yield an estimation. A varying coefficient multilevel model, on the other hand, not only is able to obtain the estimates, but also gives more efficient ones.

References

- Beck, Nathaniel, and Jonathan N. Katz. 1995. "What to Do (and Not to Do) With Time-Series Cross-Section Data." *American Political Science Review* 89 (3): 634–647.
- Beck, Nathaniel, and Jonathan N. Katz. 2001. "Throwing out the Baby with the Bath Water: a Comment on Green, Kim, and Yoon." *International Organization* 55 (2): 487–495.
- Beck, Nathaniel, and Jonathan N. Katz. 2007. "Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments." *Political Analysis* 15 (2): 182–195.
- Cochran, W. G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24 (2): 295–313.
- Congdon, P. 2005. *Bayesian Models for Categorical Data*. Chichester, England; Hoboken, NJ: Wiley.
- Doorenspleet, Renske. 2000. "Reassessing the Three Waves of Democratization." *World Politics* 52 (3): 384–406.
- Epstein, David L., Robert Bates, Jack Goldstone, Ida Kristensen, and Sharyn O'Halloran. 2006. "Democratic Transitions." *American Journal of Political Science* 50 (3): 551–569.
- Fox, John. 1997. *Applied Regression Analysis, Linear Models and Related Models*. Thousand Oaks, C.A.: Sage Publications.
- Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. UK: Cambridge University Press.
- Goldstein, Harvey. 1995. *Multilevel Statistical Models*. 2nd ed. London: Arnold.
- Hill, Jennifer, and Jerome P. Reiter. 2006. "Interval Estimation for Treatment Effects Using Propensity Score Matching." *Statistics in Medicine* 25 (13): 2230–2256.
- Hirano, Keisuke, and Guido W. Imbens. 2004. "The Propensity Score with Continuous Treatments." In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, edited by Donald B. Rubin, Andrew Gelman, and Xiao-Li Meng. Chichester, West Sussex, England ; Hoboken, NJ: John Wiley, pp. 73–84.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–960.
- Huntington, Samuel P. 1991. *The Third Wave: Democratization in the Late Twentieth Century*. Norman: University of Oklahoma Press.

- Imbens, Guido W. 2000. "The Role of the Propensity score in Estimating Dose-Response Functions." *Biometrika* 87 (3): 706–710.
- Joffe, Marshall M., and Paul R. Rosenbaum. 1999. "Invited Commentary: Propensity Scores." *American Journal of Epidemiology* 150 (4): 327–333.
- Kosuke, Imai, and David A. van Dyk. 2004. "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99: 854–866.
- Lewis, David. 1973. "Causation." *The Journal of Philosophy* 70 (17): 556–567.
- Lu, B., E. Zanutto, R. Hornik, and P. R. Rosenbaum. 2001. "Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse." *Journal of the American Statistical Association* 96: 1245–1253.
- McCullagh, Peter. 1980. "Regression Models for Ordinal Data." *Journal of the Royal Statistical Society. Series B (Methodological)* 42 (2): 109–142.
- Przeworski, Adam, Michael E. Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Material Well-Being in the World, 1950-1990*. New York, NY: Cambridge University Press.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.
- Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516–524.
- Rosenbaum, Paul R., and Donald B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39 (1): 33–38.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6 (1): 34–58.
- Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment." *Journal of the American Statistical Association* 75 (371): 591–593.

- Segal, Jodi B., Michael Griswold, Achy-Brou Aristide, Herbert Robert, Sydney M. Dy, Anne E. Millman, Albert W. Wu, and Frangakis Constantine E. 2007. "Using Propensity Scores Subclassification to Estimate Effects of Longitudinal Treatments: An Example Using a New Diabetes Medication." *Medical Care* 45 (10 Supplement 2): S149–S157.
- Shor, Boris, Joseph Bafumi, Luke Keele, and David Park. 2007. "A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data." *Political Analysis* 15 (2): 165–181.
- Steenbergen, Marco R., and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46 (1): 218–237.
- Western, Bruce. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modeling Approach." *American Journal of Political Science* 42 (4): 1233–1259.
- Zanutto, Elaine, Bo Lu, and Robert Hornik. 2005. "Using Propensity Score Subclassification for Multiple Treatment Doses to Evaluate a National Antidrug Media Campaign." *Journal of Educational and Behavioral Statistics* 30 (1): 59–73.