WHEN TO CONTROL FOR COVARIATES? PANEL ASYMPTOTICS FOR ESTIMATES OF TREATMENT EFFECTS

Joshua Angrist and Jinyong Hahn*

Abstract—The problem of when to control for continuous or high-dimensional discrete covariate vectors arises in both experimental and observational studies. Large-cell asymptotic arguments suggest that full control for covariates or stratification variables is always efficient, even if treatment is assigned independently of covariates or strata. Here, we approximate the behavior of different estimators using a panel-data-type asymptotic sequence with fixed cell sizes and the number of cells increasing to infinity. Exact calculations in simple examples and Monte Carlo evidence suggest this generates a substantially improved approximation to actual finite-sample distributions. Under this sequence, full control for covariates is dominated by propensity-score matching when cell sizes are small, the explanatory power of the covariates conditional on the propensity score is low, and/or the probability of treatment is close to 0 or 1. Our panel-asymptotic framework also provides an explanation for why propensity-score matching can dominate covariate matching even when there are no empty cells. Finally, we introduce a random-effects estimator that provides finite-sample efficiency gains over both covariate matching and propensity-score matching.

I. Introduction

EVALUATION research typically begins with treatment-control comparisons. For example, estimates of the effect of training programs on earnings compare the earnings of those who receive training with a candidate control sample of untrained people. Because trainees are not chosen randomly, candidate control samples may not provide a very accurate picture of what would have happened to the trainees had they not been trained. This motivates attempts to reduce and perhaps even eliminate bias by controlling for covariates. Examples of econometric training program evaluations in this spirit include Ashenfelter and Card (1985), Card and Sullivan (1988), Dehejia and Wahba (1999), and Heckman, Ichimura, and Todd (1997), all of which estimate the effects of training programs on earnings or employment after conditioning on an array of personal characteristics, including earnings and/or employment histories. Similarly, Angrist (1998) estimates the effect of voluntary service on the earnings of military applicants by conditioning on the personal characteristics used by military recruiters to select soldiers.

A problem that often arises in studies of this type is how to control for continuously distributed or high-dimensional covariates. In many training evaluations, for example, the sample sizes are small, there are many covariates, and some of the covariates, such as past earnings, are continuous. This leads to small or missing covariate cells. A number of variations on exact covariate-matching schemes have been developed to deal with situations like this. These typically involve approximate matching or nonparametric smoothing of some kind.[1] A practical problem with strategies of this type is that even though different estimators may have very different properties, the existing theory provides little in the way of specific guidelines as to how to choose between them. Moreover, the (finite-sample) bias from approximate matching can be substantial (Rosenbaum & Rubin, 1985a).

An alternative strategy to control for covariates begins with Rosenbaum and Rubin's (1983) observation that bias can be eliminated by controlling for a scalar-valued function of the covariates, the propensity score. For a formal statement of this result, denote the covariate vector for person i by X_i and the treatment status by D_i, and define the conditional probability of treatment, or propensity score, as $p(X_i) \equiv \Pr[D_i = 1|X_i]$. Let Y_{0i} denote the potential or counterfactual earnings of a trainee if he or she had not been trained, and let Y_{1i} denote potential earnings as a trainee. The assumption that motivates covariate matching is that conditioning on X_i eliminates selection bias, that is,

$$Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i|X_i. \tag{1}$$

Rosenbaum and Rubin's propensity-score theorem states that if equation (1) is true, then it must also be true that conditioning on $p(X_i)$ eliminates selection bias, that is,

$$Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i|p(X_i). \tag{2}$$

The value of propensity-score matching is in the dimension reduction generated by regions in the support of X where $p(X_i)$ is constant but $E[Y_{1i}|X_i]$ and $E[Y_{0i}|X_i]$ are not constant. In a simple randomized trial, for example, $p(X_i)$ is constant, so there is no need to control for covariates to eliminate bias.

Propensity-score matching in empirical work is often based on an estimated propensity score. In models with discrete covariates and no parametric assumptions or restrictions on the score, matching with an estimated propensity score is the same as covariate matching (Hirano, Imbens, & Ridder, 1999). A distinction between propensity-score matching and covariate matching arises in practice, however, because applied researchers have more information or are willing to make stronger assumptions about treatment assignment than about the relationship between covariates and outcomes. A number of empirical examples using the propensity score suggest that this approach works reasonably well (see, for example, Rosenbaum & Rubin,

Received for publication January 9, 2002. Revision accepted for publication June 4, 2003.

* MIT and NBER; and UCLA, respectively.

We are grateful to Alberto Abadie, Gary Chamberlain, Guido Imbens, Whitney Newey, and seminar participants at Berkeley, Hebrew University, Maryland, MIT, Penn State, The University of St Gallen, and Wisconsin for helpful discussions and comments. This paper is a revised version of NBER Technical Working Paper 241.

[1] See, for example, Cochran (1965), Rubin (1973, 1979), or Rosenbaum (1995, Chapter 9) for discussions of caliper and nearest-neighbor matching, or Heckman et al. (1997) for examples of nonparametric smoothing.

1984, 1985b; Dehejia & Wahba, 1999; Imbens, Rubin, & Sacerdote, 2001; Heckman, Ichimura, & Todd, 1998).

This evidence of practical utility notwithstanding, from a theoretical point of view, propensity-score-based estimators present a puzzle when sampling variance is a consideration as well as bias. Hahn (1998) shows that the propensity score is ancillary for estimates of average treatment effects, in the sense that knowledge of the propensity score does not lower the semiparametric efficiency bound for this parameter. Moreover, covariate matching is asymptotically efficient, that is, attains the semiparametric efficiency bound, whereas propensity-score matching does not. Finally, these theoretical results include the case where exact matching is not feasible, as with continuously distributed covariates, and the relevant conditional mean functions must be approximated. In short, conventional asymptotic arguments would appear to offer no justification for anything other than full control for covariates in estimation of average treatment effects.[2]

The first purpose of this paper is to develop an analytical framework and present some examples to substantiate the intuition that, because covariate cells may be small or empty, in finite samples there is a cost to covariate matching, even if the covariates are discrete and exact matching is feasible. Our framework suggests that exact covariate matching with discrete covariates is, in many important cases, less efficient than propensity-score matching. These results come from an analogy to well-known finite-sample results for random-effects panel data models with nonstochastic regressors. In addition to providing some intuition for the finite-sample behavior of alternative estimators, the panel framework also provides an explanation for why propensity-score matching can dominate covariate matching even when there are no empty cells.

A second goal of our paper is to provide specific guidelines for the relative finite-sample performance of covariate matching and propensity-score matching estimators. These results are based on an alternative asymptotic approximation where cell sizes are fixed but the number of cells becomes infinitely large. Because this approximation is similar to the large-cross-section, small-time-series asymptotic approximation commonly used for panel models, we call this *panel asymptotics*. The panel-asymptotic sequence is similar to sequences used by Bekker (1994), Bekker and van der Ploeg (1996), and Angrist and Krueger (1995) to analyze the finite-sample behavior of instrumental variables estimators. Our panel-asymptotic analysis shows that propensity-score matching is more efficient than covariate matching when cell sizes are small, the explanatory value of the covariates is low conditional on the propensity score, and/or the probability of treatment is close to 0 or 1.

The paper is organized as follows. In the next section, we outline the basic setup and compare the finite-sample behavior of two types of matching estimators in a simple model. Section III develops the panel-data version of the treatment effects problem, and introduces an alternative asymptotic sequence based on increasing the number of cells of fixed size. That section also discusses the possibility of producing a more efficient random-effects estimator from a linear combination of covariate-matching and propensity-score-matching estimators. Section IV discusses the likely generality of these results and presents some Monte Carlo evidence which suggests that the new asymptotic sequence does indeed provide an accurate description of the relative finite-sample performance of matching and propensity-score estimators. Finally, Section V concludes and suggests some directions for further work. Technical derivations are presented in an appendix.

II. Notation and Motivation

The setup is as follows. We first presume that conditioning on covariates eliminates selection bias:

ASSUMPTION 1: Treatment is independent of potential outcomes (ignorable), that is, $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$.

Also, the distribution of regressors is characterized by:

ASSUMPTION 2: X_i is multinomial and takes K possible values, say $x^{(1)}, \ldots, x^{(K)}$, with probability $\frac{1}{K}$.

This is a modeling device that allows us to change the number of cells. We believe this is not really restrictive, for $x^{(1)}, \ldots, x^{(K)}$ can be anything. In particular, the multinomial assumption allows for a discrete approximation to any distribution for large K.[3]

The treatment-assignment mechanism is described next:

ASSUMPTION 3: The propensity score $\Pr[D_i = 1|X_i]$ is known to be fixed at π.

The motivation for this is that in models with discrete covariates, differences between covariate matching and propensity-score matching arise from the manner in which the covariates are handled when the propensity score is constant. A fixed propensity score allows us to capture this idea very simply. Inasmuch as observations are assumed to be independent across cells, the question of efficiency in a setting with a variable propensity score is also addressed by looking at a single score value. Finally, the assumption of a fixed score reflects our view that the imposition of restrictions on the propensity score lies at the heart of the propensity-score–covariate-matching distinction.[4]

The two most commonly discussed parameters in evaluation studies are the effect of treatment on the treated

[2] Robins and Ritov (1997) discuss a related problem.

[3] Chamberlain (1987) used a similar strategy to analyze a semiparametric model.

[4] As a practical matter, empirical work relies on an estimated propensity score, which may lead to errors in propensity-score matching. See Abadie and Imbens (2002) for an approach to correcting the bias induced by imperfect matching.

$E[Y_{1i} - Y_{0i}|D_i = 1]$, and the average treatment effect $E[Y_{1i} - Y_{0i}]$. Because $\Pr[D_i = 1|X_i] = \pi$ in our setup, $E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}]$. This equivalence allows us to sidestep the fact that knowledge of the propensity score can reduce the asymptotic variance bound for $E[Y_{1i} - Y_{0i}|D_i = 1]$. The propensity score weights covariate-specific comparisons underlying the effect-on-the-treated parameter, though the efficient estimator for this parameter still involves covariate matching and not matching on the propensity score (Hahn, 1998, Proposition 7). The fact that the propensity score is used for more than matching in estimates of effects on the treated affects the statistical propensities of alternative estimators (see, for example, Hirano, Imbens, & Ridder, 1999). We therefore leave the more complicated question of efficient estimation of effects on the treated for future work.

In most of the paper, we model cell size as fixed, so the sampling framework stratifies on X_i:

ASSUMPTION 2′: Each cell size is equal to M. We adopt the convention that the first n_{1k} individuals are treated in each cell, so that $n_{1k} \sim \text{Binomial}(M, \pi)$.

Stratified sampling is empirically relevant for some studies (for example, Card & Sullivan, 1988; Angrist, 1998), but we adopt this assumption for technical reasons: it simplifies the arguments and allows us to focus on the randomness in treatment status and outcomes within covariate cells. It should also be noted that the ignorability assumption makes X_i ancillary, so little would seem to be lost from stratification in this setting. Later, we substantiate this claim in a Monte Carlo comparison using random sampling. Although these assumptions are restrictive and highly stylized, we show below that allowing the number of fixed-size cells to approach infinity generates an accurate characterization of the relative finite-sample properties of the estimators we study.

For the next step, the following notation is useful:

Definition 1. Let y_{0ki} and y_{1ki} denote potential outcomes under control and treatment for the i^{th} individual in the k^{th} cell. Let y_{0k} and y_{1k} denote the expected potential outcomes under control and under treatment in the k^{th} cell. Also, let σ^2_{0k} and σ^2_{1k} denote the conditional variance of y_{0ki} and y_{1ki} in the k^{th} cell. Finally, let $\alpha_k \equiv y_{0k}$ and $\beta_k \equiv y_{1k} - y_{0k}$.

We now define the two estimators considered in this paper. The covariate matching estimator is

$$b_c \equiv \frac{1}{\sum_{k=1}^{K} 1(1 \le n_{1k} \le M-1)} \times \sum_{k=1}^{K}\left[1(1 \le n_{1k} \le M-1) \times \left(\frac{1}{n_{1k}}\sum_{i=1}^{n_{1k}} y_{1ki} - \frac{1}{n_{0k}}\sum_{i=n_{1k}+1}^{M} y_{0ki}\right)\right].$$

Because the propensity score is constant, matching on it is equivalent to ignoring covariates. The propensity-score matching estimator is therefore

$$b_p \equiv \left(\frac{1}{\sum_{k=1}^{K} 1(1 \le n_{1k}) n_{1k}} \times \sum_{k=1}^{K}\left[1(1 \le n_{1k}) n_{1k}\left(\frac{1}{n_{1k}}\sum_{i=1}^{n_{1k}} y_{1ki}\right)\right]\right) - \left(\frac{1}{\sum_{k=1}^{K} 1(n_{1k} \le M-1) n_{0k}} \times \sum_{k=1}^{K}\left[1(n_{1k} \le M-1) n_{0k}\left(\frac{1}{n_{0k}}\sum_{i=n_{1k}+1}^{M} y_{0ki}\right)\right]\right).$$

Note that both estimators are unbiased, and that there is some probability that matching on covariates and/or the propensity score cannot be implemented. For example, if all cells consist of only treated individuals, then matching on either covariates or the propensity score is infeasible.

A simple example can be used to illustrate important differences in the finite-sample behavior of b_c and b_p. Assume that:

1. $K = 2$.
2. The treatment effect is constant and equal to β_0, that is, $y_{0ki} = \alpha_k + \varepsilon_{ki}$ and $y_{1ki} = \beta_0 + \alpha_k + \varepsilon_{ki}$.
3. $E[\varepsilon^2_{ki}] = 1$.

The analysis of this example is facilitated by distinguishing three cases. First, both cells may contain treated and control observations. Second, one of the two cells may consist of treated or control observations only. Third, each cell might consist of treated or control observations only, in which case b_c cannot be computed. We therefore focus on variance comparisons conditional on the event that b_c exists. The efficiency of b_p relative to b_c is defined to be $\sqrt{\text{Var}(b_c)}/\sqrt{\text{Var}(b_p)}$, where $\text{Var}(b_c)$ and $\text{Var}(b_p)$ denote the *conditional* variance of b_c and b_p given the event that both are computable. It is also useful to define

$$\text{Var}(\alpha_k) \equiv \frac{1}{2}\sum_{k=1}^{2}(\alpha_k - \bar{\alpha})^2, \quad \text{where} \quad \bar{\alpha} \equiv \frac{1}{2}\sum_{k=1}^{2}\alpha_k. \quad (3)$$

Note that the R^2 in the theoretical regression of the outcomes on covariates (cell indicators) can be written $\text{Var}(\alpha_k)/[\text{Var}(\alpha_k) + 1]$.

We tabulated the relative efficiency of b_p for various (π, M, R^2) combinations. Tables 1 and 2 report the relative

TABLE 1.—ACTUAL RELATIVE PERFORMANCE [SE(b_c)/SE(b_p)] OF b_p IN A TWO-CELL EXAMPLE ($\pi = 0.1$)

Cell Size (M)	Covariate R^2								
	.40	.35	.30	.25	.20	.15	.10	.05	.00
2	0.90	0.94	0.98	1.02	1.06	1.10	1.13	1.17	1.21
3	0.84	0.88	0.92	0.95	0.99	1.02	1.06	1.09	1.13
4	0.82	0.85	0.89	0.92	0.96	0.99	1.02	1.06	1.09
5	0.81	0.84	0.88	0.91	0.94	0.97	1.01	1.04	1.07
6	0.80	0.84	0.87	0.90	0.93	0.96	0.99	1.02	1.05
7	0.80	0.83	0.86	0.90	0.93	0.96	0.99	1.02	1.05
8	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.04
9	0.79	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.04
10	0.79	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.04
11	0.79	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.04
12	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.04
13	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.04
14	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.04
15	0.80	0.83	0.86	0.89	0.93	0.95	0.98	1.01	1.04
16	0.80	0.83	0.87	0.90	0.93	0.96	0.98	1.01	1.04
17	0.80	0.83	0.87	0.90	0.93	0.96	0.99	1.01	1.04
18	0.80	0.84	0.87	0.90	0.93	0.96	0.99	1.02	1.04
19	0.81	0.84	0.87	0.90	0.93	0.96	0.99	1.02	1.05
20	0.81	0.84	0.87	0.90	0.93	0.96	0.99	1.02	1.05
21	0.81	0.84	0.87	0.91	0.94	0.97	0.99	1.02	1.05
22	0.81	0.84	0.88	0.91	0.94	0.97	1.00	1.02	1.05
23	0.81	0.85	0.88	0.91	0.94	0.97	1.00	1.03	1.05
24	0.81	0.85	0.88	0.91	0.94	0.97	1.00	1.03	1.06
25	0.82	0.85	0.88	0.91	0.94	0.97	1.00	1.03	1.06
26	0.82	0.85	0.88	0.91	0.95	0.97	1.00	1.03	1.06
27	0.82	0.85	0.88	0.92	0.95	0.98	1.01	1.03	1.06
28	0.82	0.85	0.89	0.92	0.95	0.98	1.01	1.04	1.06
29	0.82	0.85	0.89	0.92	0.95	0.98	1.01	1.04	1.06
30	0.82	0.86	0.89	0.92	0.95	0.98	1.01	1.04	1.07
80	0.80	0.84	0.87	0.90	0.93	0.96	0.98	1.01	1.04
90	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.03
100	0.80	0.83	0.86	0.89	0.92	0.95	0.97	1.00	1.03

Notes: The table reports relative standard errors for covariate-matching and propensity-score estimates of a constant-treatment-effect homoskedastic model with two covariate cells. Cell size is fixed at M. The standard errors are based on an exact calculation detailed in the Appendix. The probability of treatment in this case is 1/10.

efficiency of b_p measured by the ratio $\sqrt{\text{Var}(b_c)}/\sqrt{\text{Var}(b_p)}$ for $\pi = 0.1$ and 0.5. This is an exact finite-sample calculation, the details of which are discussed in the Appendix. The tables show that the relative efficiency of b_p increases as R^2 falls, M falls, or π falls, and that b_p is actually more efficient than b_c for some (π, M, R^2) combinations. These exact calculations suggest that conventional (large-cell) asymptotic approximations may provide a poor guide to the relative precision of these estimators in some applications.

We can also ask whether the relative efficiency of b_p for some (π, M, R^2) combinations is solely a consequence of the fact that b_p typically uses more observations than b_c. Consider the relative efficiency in case 1, where both estimators use the same number of observations. Any difference in variance in this case can therefore be attributed to the efficiency with which each estimator processes information. When $M = 2$, $b_p = b_c$ in case 1, so there is obviously no efficiency difference. But with $M = 3$ and $\pi = 0.5$, b_p is moderately more efficient than b_c for $R^2 \leq .16$. Thus, the relative finite-sample efficiency of b_p arises—at least in this example—for reasons beyond the fact that b_p uses more observations.

III. Panel Characterization of the Treatment-Effects Model

The panel analog of the model in the previous section allows us to draw on the econometric literature dealing with problems of this type. Examples include Wallace and Hussain (1969), Maddala (1971), Chamberlain and Griliches (1975), Mundlak (1978), Hausman and Taylor (1981), and Chamberlain (1984). We argue that covariate matching is a type of "within" estimator, while propensity-score matching is a "pooled" estimator. Standard results for random-effects panel models with nonstochastic regressors suggest that neither within estimators or pooled estimators are efficient, and that their precision cannot be ranked unambiguously.

A. *Random-Coefficients Notation*

The panel equivalent of the evaluation problem looks like this. Let D_{ki} denote a binary treatment indicator, and write the observed y_{ki} as

$$y_{ki} \equiv D_{ki}y_{1ki} + (1 - D_{ki})y_{0ki} = y_{0ki} + (y_{1ki} - y_{0ki})D_{ki}$$

$$= \alpha_k + \beta_k D_{ki} + \{(y_{0ki} - \alpha_k) + (y_{1ki} - y_{0ki} - \beta_k)D_{ki}\}.$$

TABLE 2.—ACTUAL RELATIVE PERFORMANCE [SE(b_c)/SE(b_p)] OF b_p IN A TWO-CELL EXAMPLE ($\pi = 0.5$)

Cell Size (M)	Covariate R^2								
	.40	.35	.30	.25	.20	.15	.10	.05	.00
2	0.91	0.95	0.98	1.02	1.05	1.08	1.11	1.14	1.17
3	0.87	0.90	0.94	0.97	1.01	1.04	1.07	1.10	1.14
4	0.84	0.88	0.91	0.95	0.98	1.01	1.05	1.08	1.11
5	0.83	0.86	0.90	0.93	0.96	1.00	1.03	1.06	1.09
6	0.82	0.85	0.89	0.92	0.95	0.98	1.02	1.05	1.08
7	0.81	0.85	0.88	0.91	0.94	0.98	1.01	1.04	1.06
8	0.81	0.84	0.87	0.91	0.94	0.97	1.00	1.03	1.06
9	0.80	0.84	0.87	0.90	0.93	0.96	0.99	1.02	1.05
10	0.80	0.83	0.86	0.90	0.93	0.96	0.99	1.01	1.04
11	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.04
12	0.79	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.03
13	0.79	0.82	0.86	0.89	0.92	0.95	0.97	1.00	1.03
14	0.79	0.82	0.85	0.88	0.91	0.94	0.97	1.00	1.03
15	0.79	0.82	0.85	0.88	0.91	0.94	0.97	1.00	1.02
16	0.79	0.82	0.85	0.88	0.91	0.94	0.97	0.99	1.02
17	0.79	0.82	0.85	0.88	0.91	0.94	0.97	0.99	1.02
18	0.78	0.82	0.85	0.88	0.91	0.94	0.96	0.99	1.02
19	0.78	0.82	0.85	0.88	0.91	0.94	0.96	0.99	1.02
20	0.78	0.82	0.85	0.88	0.91	0.93	0.96	0.99	1.02
21	0.78	0.81	0.85	0.88	0.91	0.93	0.96	0.99	1.02
22	0.78	0.81	0.85	0.88	0.90	0.93	0.96	0.99	1.01
23	0.78	0.81	0.84	0.88	0.90	0.93	0.96	0.99	1.01
24	0.78	0.81	0.84	0.87	0.90	0.93	0.96	0.99	1.01
25	0.78	0.81	0.84	0.87	0.90	0.93	0.96	0.99	1.01
26	0.78	0.81	0.84	0.87	0.90	0.93	0.96	0.99	1.01
27	0.78	0.81	0.84	0.87	0.90	0.93	0.96	0.98	1.01
28	0.78	0.81	0.84	0.87	0.90	0.93	0.96	0.98	1.01
29	0.78	0.81	0.84	0.87	0.90	0.93	0.96	0.98	1.01
30	0.78	0.81	0.84	0.87	0.90	0.93	0.96	0.98	1.01
80	0.78	0.81	0.84	0.87	0.90	0.92	0.95	0.98	1.00
90	0.78	0.81	0.84	0.87	0.90	0.92	0.95	0.98	1.00
100	0.78	0.81	0.84	0.87	0.90	0.92	0.95	0.98	1.00

Notes: The table reports relative standard errors for covariate-matching and propensity-score estimates of a constant-treatment-effect homoskedastic model with two covariate cells. Cell size is fixed at M. The standard errors are based on an exact calculation detailed in the Appendix. The probability of treatment in this case is 1/2.

With ε_{ki} defined as the residual in the above equation, we can write

$$y_{ki} = \alpha_k + \beta_k D_{ki} + \varepsilon_{ki}, \qquad k = 1, \ldots, K, \quad i = 1, \ldots, M, \tag{4}$$

where the parameter of interest is equal to $E[\beta_k]$. This is the random-coefficient panel model considered by Swamy (1970) and Chamberlain (1992).[5]

Consider first the simple model where β_k is fixed at β_0 and ε_{ki} is homoskedastic with variance σ^2_ε. Observe that α_k is independent of D_{ki} because of assumption 1, so $\text{Var}(\alpha_k|D_{ki}) = \text{Var}(\alpha_k) \equiv \sigma^2_\alpha$. Conditional on the realizations of D_{ki} and assuming that both b_p and b_c can be computed, it is easy to see that, under these assumptions, equation (4) is the traditional random-effects panel model with nonstochastic regressors

$$y_{ki} = \alpha_k + \beta_0 D_{ki} + \varepsilon_{ki}. \tag{5}$$

The efficient unbiased estimator for this model is a weighted average of between and within estimators, or equivalently, within and pooled estimators.[6] Note that

$$b_c = \frac{1}{K}\sum_{k=1}^{K} \hat{b}_k,$$

where $\hat{b}_k$ is the OLS estimator of β_k for the k^{th} cell. Because b_c is the sample average of estimators using only within-cell variation, b_c is also a within-type estimator, though it is not equal to the traditional within estimator, which can be written in this case as

$$b_w = \sum_{k=1}^{K} \hat{\pi}_k(1-\hat{\pi}_k)\hat{b}_k \Big/ \sum_{k=1}^{K} \hat{\pi}_k(1-\hat{\pi}_k),$$

where $\hat{\pi}_k = \frac{1}{M}\sum_{i=1}^{M} D_{ki}$.[7] Note also that b_p is the OLS coefficient from a regression of y on D, and is therefore the traditional pooled estimator that ignores group structure.

[5] Chamberlain and Imbens (1996) similarly treat unobserved covariate effects as random in a high-dimension instrumental variables problem.

[6] See, for example, Maddala (1971).

[7] See, for example, Angrist (1998).

This suggests that b_c and b_p cannot be ranked unambiguously, because neither within-type nor pooled estimators are efficient for random-effects panel models.

To substantiate this conjecture, we calculated the finite-sample variances of b_c and b_p, treating D_{ki} as nonstochastic (so n_{1k} is fixed) and α_k as random, assuming constant treatment effects and homoskedastic errors, and conditional on b_c and b_p both being computable. This is the stochastic environment typical of the panel-data literature. Because the variance is conditional on n_{1k}, there is no problem of missing covariate cells for the matching estimator.

In this setting, it can be shown that

$$\mathrm{Var}(b_c) = \frac{\sigma_\varepsilon^2}{K^2} \sum_{k=1}^{K} \left(\frac{1}{n_{1k}} + \frac{1}{M - n_{1k}} \right),$$

and

$$\mathrm{Var}(b_p) = \sigma_\alpha^2 \sum_{k=1}^{K} \left(\frac{n_{1k}}{\sum_{k=1}^{K} n_{1k}} - \frac{M - n_{1k}}{KM - \sum_{k=1}^{K} n_{1k}} \right)^2 + \sigma_\varepsilon^2 \left(\frac{1}{\sum_{k=1}^{K} n_{1k}} + \frac{1}{KM - \sum_{k=1}^{K} n_{1k}} \right).$$

For example, if $K = 2$, $M = 3$, $n_{11} = 1$, and $n_{12} = 2$, we have

$$\mathrm{Var}(b_c) = \frac{3}{4}\sigma_\varepsilon^2 \quad \text{and} \quad \mathrm{Var}(b_p) = \frac{2}{9}\sigma_\alpha^2 + \frac{2}{3}\sigma_\varepsilon^2.$$

Hence, the difference between them is

$$\mathrm{Var}(b_c) - \mathrm{Var}(b_p) = \frac{1}{12}\sigma_\varepsilon^2 - \frac{2}{9}\sigma_\alpha^2,$$

which is of ambiguous sign.

How can this ambiguity be reconciled with the conventional asymptotic result that b_c is more efficient than b_p? Traditional asymptotic arguments fix the data-generating process and let the number of observations grow to infinity. In our setting, this asymptotic sequence would have K fixed while $M \to \infty$. The efficient random-effects panel estimator converges to the fixed effects estimator under this asymptotic sequence. Fixed-effects estimation is also the efficiently weighted matching estimator under constant treatment effects, so there is no contradiction between finite-sample results for the nonstochastic panel and the asymptotic efficiency of covariate matching under a large-M asymptotic sequence. On the other hand, for small M a panel-type asymptotic sequence with M fixed while $K \to \infty$ may be more appropriate.

B. Panel Asymptotics

To provide more general results on the relative efficiency of covariate and propensity score matching, we use an alternative asymptotic approximation where cell sizes are fixed and the number of cells grows to infinity. As noted above, this corresponds to a large-cross-section, small-time-series asymptotic approximation for panel data. The analog of the cross-section dimension in our case is K, and the analog of the time-series dimension is M.

As a regularity condition, we assume that

ASSUMPTION 4: The sequence $\{(\alpha_k, \beta_k, \varepsilon_{k1}, \ldots, \varepsilon_{kM}); k = 1, 2, \ldots\}$ is i.i.d. Furthermore, for given (α_k, β_k), $(D_{k1}, \varepsilon_{k1}), \ldots, (D_{kM}, \varepsilon_{kM})$ are i.i.d.

Assumption 4 implies that we approximate sampling distributions without assuming any prior information on α_k and β_k. This seems consistent with the nonparametric spirit of matching procedures.

As before, our objective is to estimate the average treatment effect, $\beta \equiv E[y_{1ki} - y_{0ki}]$. The main theoretical result is given below:

Theorem 1. Under assumptions 1–5, we have

$$\sqrt{K}\,(b_c - \beta) \to \mathcal{N}(0, \omega_c^2),$$
$$\sqrt{K}\,(b_p - \beta) \to \mathcal{N}(0, \omega_p^2),$$

where

$$\omega_c^2 \equiv \frac{g(\pi, M)\,E[\sigma_{1k}^2] + g(1-\pi, M)\,E[\sigma_{0k}^2]}{\{1 - \pi^M - (1-\pi)^M\}^2} + \frac{\mathrm{Var}(y_{1k} - y_{0k})}{1 - \pi^M - (1-\pi)^M},$$

and

$$\omega_p^2 \equiv \frac{1}{M\pi}E[\sigma_{1k}^2] + \frac{1}{M(1-\pi)}E[\sigma_{1k}^2] + \mathrm{Var}(y_{1k} - y_{0k}) + \frac{1}{M}\mathrm{Var}\left(\sqrt{\frac{1-\pi}{\pi}}\,y_{1k} + \sqrt{\frac{\pi}{1-\pi}}\,y_{0k} \right),$$

where

$$g(\pi, M) \equiv \sum_{k=1}^{M-1} \binom{M}{k} \pi^k (1-\pi)^{M-k} \frac{1}{k}.$$

Proof. See Angrist and Hahn (1999).

The implications of this result for the relative sampling variance of b_c and b_p can be summarized using $\sqrt{\omega_c^2/\omega_p^2}$. This expression is complicated but can be tabulated, or simplified for special cases. Tables 3 and 4 report the relative efficiency of b_p for two values of π, assuming a

TABLE 3.—RELATIVE PERFORMANCE [SE(b_c)/SE(b_p)] OF b_p USING PANEL ASYMPTOTICS ($\pi = 0.1$)

Cell Size (M)	Covariate R^2								
	.40	.35	.30	.25	.20	.15	.10	.05	.00
2	1.10	1.14	1.18	1.23	1.27	1.30	1.34	1.38	1.41
3	0.95	0.99	1.03	1.06	1.10	1.13	1.16	1.19	1.23
4	0.90	0.94	0.97	1.01	1.04	1.07	1.10	1.13	1.16
5	0.88	0.91	0.95	0.98	1.01	1.04	1.08	1.10	1.13
6	0.87	0.90	0.94	0.97	1.00	1.03	1.06	1.09	1.12
7	0.86	0.90	0.93	0.96	1.00	1.03	1.06	1.09	1.11
8	0.86	0.90	0.93	0.96	0.99	1.03	1.06	1.08	1.11
9	0.86	0.90	0.93	0.96	0.99	1.03	1.06	1.08	1.11
10	0.86	0.90	0.93	0.96	1.00	1.03	1.06	1.09	1.11
11	0.86	0.90	0.93	0.97	1.00	1.03	1.06	1.09	1.12
12	0.87	0.90	0.94	0.97	1.00	1.03	1.06	1.09	1.12
13	0.87	0.90	0.94	0.97	1.00	1.03	1.06	1.09	1.12
14	0.87	0.91	0.94	0.97	1.01	1.04	1.07	1.10	1.12
15	0.87	0.91	0.94	0.98	1.01	1.04	1.07	1.10	1.13
16	0.88	0.91	0.95	0.98	1.01	1.04	1.07	1.10	1.13
17	0.88	0.91	0.95	0.98	1.01	1.04	1.08	1.10	1.13
18	0.88	0.92	0.95	0.98	1.02	1.05	1.08	1.11	1.14
19	0.88	0.92	0.95	0.99	1.02	1.05	1.08	1.11	1.14
20	0.88	0.92	0.95	0.99	1.02	1.05	1.08	1.11	1.14
21	0.88	0.92	0.95	0.99	1.02	1.05	1.08	1.11	1.14
22	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
23	0.89	0.92	0.96	0.99	1.02	1.06	1.09	1.12	1.14
24	0.89	0.92	0.96	0.99	1.02	1.06	1.09	1.12	1.15
25	0.89	0.92	0.96	0.99	1.03	1.06	1.09	1.12	1.15
26	0.89	0.92	0.96	0.99	1.03	1.06	1.09	1.12	1.15
27	0.89	0.92	0.96	0.99	1.03	1.06	1.09	1.12	1.15
28	0.89	0.92	0.96	0.99	1.03	1.06	1.09	1.12	1.15
29	0.89	0.92	0.96	0.99	1.03	1.06	1.09	1.12	1.15
30	0.89	0.92	0.96	0.99	1.03	1.06	1.09	1.12	1.15
80	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.04	1.07
90	0.82	0.85	0.89	0.92	0.95	0.98	1.00	1.03	1.06
100	0.81	0.85	0.88	0.91	0.94	0.97	1.00	1.02	1.05

Notes: The table reports relative standard errors for covariate-matching and propensity-score estimates of a constant-treatment-effect homoskedastic model. Cell size is fixed at M. The standard errors are based on the panel-asymptotic approximation in theorem 1. The probability of treatment in this case is 1/10.

constant treatment effect and homoskedastic errors. As before, we define the theoretical R^2 as $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$. The tables again show that the relative efficiency of b_p typically increases as R^2 falls, M falls, or π falls, and that b_p is actually more efficient than b_c for some (π, M, R^2) combinations. Interestingly, the case for ignoring covariates is even stronger in this set of tabulations than for the two-cell example. On the other hand, it is also noteworthy that the many-cell panel-asymptotic approximation captures important features of the actual finite-sample distributions for a two-cell example. In section IV, we turn to the question of whether theorem 1 captures the actual finite-sample behavior of b_c and b_p in samples with many cells and random cell sizes. First, however, we compare the many-cell and large-cell approximations for the model analyzed in the theorem.

C. Comparison with Large-Cell Asymptotics

How do panel-asymptotic results differ from conventional asymptotic results, where the number of cells is fixed and cell sizes are random and increasing? Let $\mathbb{N} = M^*K$ and M^* denote the total sample size and average cell size in a random sample. Using an $M^* \to \infty$ conventional asymptotic sequence, where $\mathbb{N}$ grows to ∞ as a consequence while K is fixed, it is straightforward to show that

$$\sqrt{\mathbb{N}}\,(b_c - \beta) \to \mathcal{N}\left(0, \frac{E[\sigma_{1k}^2]}{\pi} + \frac{E[\sigma_{0k}^2]}{1-\pi} + \mathrm{Var}(y_{1k} - y_{0k})\right)$$

and

$$\sqrt{\mathbb{N}}\,(b_p - \beta) \to \mathcal{N}\left(0, \frac{E[\sigma_{1k}^2]}{\pi} + \frac{E[\sigma_{0k}^2]}{1-\pi} + \mathrm{Var}(y_{1k} - y_{0k}) + \mathrm{Var}\left(\sqrt{\frac{1-\pi}{\pi}}\,y_{1k} + \sqrt{\frac{\pi}{1-\pi}}\,y_{0k}\right)\right).$$

So conventional asymptotics approximate finite sample variances as

$$\mathrm{CVar}(b_c) \equiv \frac{1}{\mathbb{N}}\left(\frac{E[\sigma_{1k}^2]}{\pi} + \frac{E[\sigma_{0k}^2]}{1-\pi} + \mathrm{Var}(y_{1k} - y_{0k})\right), \quad (6)$$

TABLE 4.—RELATIVE PERFORMANCE [SE(b_c)/SE(b_p)] OF b_p USING PANEL ASYMPTOTICS ($\pi = 0.5$)

Cell Size (M)	Covariate R^2								
	.40	.35	.30	.25	.20	.15	.10	.05	.00
2	1.10	1.14	1.18	1.23	1.27	1.30	1.34	1.38	1.41
3	0.95	0.99	1.03	1.06	1.10	1.13	1.16	1.19	1.23
4	0.90	0.94	0.98	1.01	1.04	1.08	1.11	1.14	1.17
5	0.88	0.92	0.95	0.99	1.02	1.05	1.08	1.11	1.14
6	0.87	0.90	0.94	0.97	1.00	1.03	1.06	1.09	1.12
7	0.86	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11
8	0.85	0.88	0.91	0.95	0.98	1.01	1.04	1.07	1.09
9	0.84	0.87	0.90	0.94	0.97	1.00	1.03	1.05	1.08
10	0.83	0.86	0.90	0.93	0.96	0.99	1.02	1.05	1.07
11	0.82	0.86	0.89	0.92	0.95	0.98	1.01	1.04	1.06
12	0.82	0.85	0.88	0.92	0.95	0.97	1.00	1.03	1.06
13	0.81	0.85	0.88	0.91	0.94	0.97	1.00	1.02	1.05
14	0.81	0.84	0.88	0.91	0.94	0.96	0.99	1.02	1.05
15	0.81	0.84	0.87	0.90	0.93	0.96	0.99	1.02	1.04
16	0.80	0.84	0.87	0.90	0.93	0.96	0.99	1.01	1.04
17	0.80	0.84	0.87	0.90	0.93	0.96	0.98	1.01	1.04
18	0.80	0.83	0.86	0.90	0.92	0.95	0.98	1.01	1.03
19	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.03
20	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.00	1.03
21	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.00	1.03
22	0.80	0.83	0.86	0.89	0.92	0.95	0.97	1.00	1.03
23	0.79	0.83	0.86	0.89	0.92	0.95	0.97	1.00	1.03
24	0.79	0.83	0.86	0.89	0.92	0.94	0.97	1.00	1.02
25	0.79	0.82	0.86	0.89	0.91	0.94	0.97	1.00	1.02
26	0.79	0.82	0.85	0.88	0.91	0.94	0.97	1.00	1.02
27	0.79	0.82	0.85	0.88	0.91	0.94	0.97	0.99	1.02
28	0.79	0.82	0.85	0.88	0.91	0.94	0.97	0.99	1.02
29	0.79	0.82	0.85	0.88	0.91	0.94	0.97	0.99	1.02
30	0.79	0.82	0.85	0.88	0.91	0.94	0.97	0.99	1.02
80	0.78	0.81	0.84	0.87	0.90	0.93	0.95	0.98	1.01
90	0.78	0.81	0.84	0.87	0.90	0.93	0.95	0.98	1.01
100	0.78	0.81	0.84	0.87	0.90	0.93	0.95	0.98	1.01

Notes: The table reports relative standard errors for covariate-matching and propensity-score estimates of a constant-treatment-effect homoskedastic model. Cell size is fixed at M. The standard errors are based on the panel-asymptotic approximation in theorem 1. The probability of treatment in this case is 1/2.

and

$$\mathrm{CVar}(b_p) \equiv \frac{1}{\mathbb{N}}\left(\frac{E[\sigma_{1k}^2]}{\pi} + \frac{E[\sigma_{0k}^2]}{1-\pi} + \mathrm{Var}(y_{1k} - y_{0k}) + \mathrm{Var}\left(\sqrt{\frac{1-\pi}{\pi}}\, y_{1k} + \sqrt{\frac{\pi}{1-\pi}}\, y_{0k}\right)\right). \tag{7}$$

The last term in CVar(b_p) can be interpreted as the penalty for failure to control for covariates under conventional asymptotics. Note that, with constant treatment effects, this term equals 0 if the between cell variance is 0.

Panel asymptotics approximate finite-sample variances as

$$\mathrm{PVar}(b_c) \equiv \frac{1}{\mathbb{N}}\left(\frac{Mg(\pi, M)E[\sigma_{1k}^2] + Mg(1-\pi, M)E[\sigma_{0k}^2]}{\{1-\pi^M-(1-\pi)^M\}^2} + \frac{M\,\mathrm{Var}(y_{1k} - y_{0k})}{1-\pi^M-(1-\pi)^M}\right) \tag{8}$$

and

$$\mathrm{PVar}(b_p) \equiv \frac{1}{\mathbb{N}}\left(\frac{1}{\pi}E[\sigma_{1k}^2] + \frac{1}{1-\pi}E[\sigma_{1k}^2] + M\,\mathrm{Var}(y_{1k} - y_{0k}) + \mathrm{Var}\left(\sqrt{\frac{1-\pi}{\pi}}\, y_{1k} + \sqrt{\frac{\pi}{1-\pi}}\, y_{0k}\right)\right), \tag{9}$$

where we have used the fact that $\mathbb{N} = KM$ in the panel-asymptotic sequence. The penalty term in CVar(b_p) remains in PVar(b_p), but now the terms

$$\frac{E[\sigma_{1k}^2]}{\pi} \quad \text{and} \quad \frac{E[\sigma_{0k}^2]}{1-\pi}$$

in CVar(b_c) become

$$\frac{Mg(\pi, M)E[\sigma_{1k}^2]}{\{1-\pi^M-(1-\pi)^M\}^2} \quad \text{and} \quad \frac{Mg(1-\pi, M)E[\sigma_{0k}^2]}{\{1-\pi^M-(1-\pi)^M\}^2}$$

under panel asymptotics in equation (8). This is because the first two terms partly reflect the fact that some cells may have to be dropped in the computation of b_c. Note also that the panel-asymptotic approximation inflates the third term in $\text{CVar}(b_c)$ and $\text{CVar}(b_p)$, which is

$$\text{Var}(y_{1k} - y_{0k})$$

in both expressions. This term becomes

$$\frac{M\ \text{Var}(y_{1k} - y_{0k})}{1 - \pi^M - (1-\pi)^M} \quad \text{and} \quad M\ \text{Var}(y_{1k} - y_{0k})$$

in equations (8) and (9). The inflation factor $M/[1 - \pi^M - (1 - \pi)^M]$ in equation (8) is larger than M in equation (9). This partly reflects the fact that the conventional asymptotic approximation is more optimistic about the precision with which realized cell differences are actually estimated. Note also that the inflation factor is larger for π close to 0 or 1.

To summarize the difference between the two approximations, we write

$$\begin{aligned}
&[\text{PVar}(b_c) - \text{PVar}(b_p)] - [\text{CVar}(b_c) - \text{CVar}(b_p)] \\
&\quad = \frac{1}{\mathbb{N}}\left(\frac{Mg(\pi, M)}{[1 - \pi^M - (1-\pi)^M]^2} - \frac{1}{\pi}\right)E[\sigma_{1k}^2]] \\
&\quad + \frac{1}{\mathbb{N}}\left(\frac{Mg(1-\pi, M)}{[1 - \pi^M - (1-\pi)^M]^2} - \frac{1}{1-\pi}\right)E[\sigma_{0k}^2] \\
&\quad + \frac{1}{\mathbb{N}}\left(\frac{M}{1 - \pi^M - (1-\pi)^M} - M\right)\text{Var}(y_{1k} - y_{0k}).
\end{aligned}$$

The first two terms on the right can easily be shown to be nonnegative (see Angrist and Hahn, 1999). Note that the third term is 0 if and only if $\text{Var}(y_{1k} - y_{0k}) = 0$. We therefore expect the finite-sample advantage of b_p to be larger with heterogeneous treatment effects.

D. Linear Combinations of b_c and b_p

Because neither b_c nor b_p is efficient, we now ask whether it is possible to construct a treatment-effects estimator that is more efficient than both under the panel asymptotic sequence. We look at linear combinations of b_c and b_p, because we know that the efficient estimator for random-effects panel models has this form when the coefficients (treatment effects) are constant. In principle it is possible that estimators outside the linear-combination class dominate; but we leave a more general exploration of this question for future work.

Consider a minimum variance linear combination of b_c and b_p of the form

$$b^* = \xi b_c + (1 - \xi) b_p.$$

The asymptotic variance of b^* is minimized by choosing

$$\xi = \frac{\text{PVar}_a(b_p) - \text{PCov}_a(b_c, b_p)}{\text{PVar}_a(b_c) - 2\text{PCov}_a(b_c, b_p) + \text{PVar}_a(b_p)},$$

where PVar_a and PCov_a denote asymptotic variance and asymptotic covariance under the asymptotic sequence in theorem 1. The variance terms are available from theorem 1. The covariance term is

$$\begin{aligned}
&\frac{1}{\pi M}E[\sigma_{1k}^2] + \frac{1}{M(1-\pi)}E[\sigma_{0k}^2] \\
&\qquad + \frac{1 - \pi^{M-1}}{1 - \pi^M - (1-\pi)^M}\text{Var}(y_{1k}) \\
&\qquad + \frac{1 - (1-\pi)^{M-1}}{1 - \pi^M - (1-\pi)^M}\text{Var}(y_{0k}) \\
&\qquad + \frac{-2 + \pi^{M-1} + (1-\pi)^{M-1}}{1 - \pi^M - (1-\pi)^M}\text{Cov}(y_{1k}, y_{0k}).
\end{aligned}$$

Note that, as in theorem 1, b^* is derived for M fixed, as in conventional panel models. It turns out that theorem 1 provides a good approximation to finite-sample behavior even when cell sizes are not fixed, but we have not yet derived an efficient estimator for this case.[8]

To develop intuition for the weighting formula, suppose as before that (i) the treatment effect is constant and equal to β_0, that is, $y_{0ki} = \alpha_k + \varepsilon_{ki}$ and $y_{1ki} = \beta_0 + \alpha_k + \varepsilon_{ki}$; (ii) ε_{ki} has variance equal to σ_ε^2; and (iii) α_k has mean μ_α, and variance σ_α^2. After some algebra, we obtain the following simplification:

$$\text{PVar}_a(b_p) - \text{PCov}_a(b_c, b_p) = \frac{1}{M\pi(1-\pi)}\sigma_\alpha^2$$

and

$$\begin{aligned}
&\text{PVar}_a(b_c) - 2\,\text{PCov}_a(b_c, b_p) + \text{PVar}_a(b_p) \\
&\quad = \left(\frac{g(\pi, M)}{[1 - \pi^M - (1-\pi)^M]^2} + \frac{g(1-\pi, M)}{[1 - \pi^M - (1-\pi)^M]^2}\right. \\
&\quad \left. - \frac{1}{\pi M} - \frac{1}{M(1-\pi)}\right)\sigma_\varepsilon^2 + \frac{1}{M\pi(1-\pi)}\sigma_\alpha^2.
\end{aligned}$$

Therefore, the optimal weight is equal to

$$\xi^* = \frac{\dfrac{1}{\pi(1-\pi)}\sigma_\alpha^2}{\left(\dfrac{Mg(\pi, M)}{[1 - \pi^M - (1-\pi)^M]^2} + \dfrac{Mg(1-\pi, M)}{[1 - \pi^M - (1-\pi)^M]^2}\right.} \cdot$$
$$\left. - \frac{1}{\pi} - \frac{1}{(1-\pi)}\right)\sigma_\varepsilon^2 + \frac{1}{\pi(1-\pi)}\sigma_\alpha^2$$

[8] Chamberlain (1992, section 4) presents a bound for a traditional panel model, but his bound does not impose independence of ε_{ki}, $i = 1, \ldots, M$.

TABLE 5.—MONTE CARLO RELATIVE PERFORMANCE [SE(b_c)/SE(b_p)] OF b_p ($\pi = 0.1$)

Avg. Cell Size (M^*)	Covariate R^2								
	.40	.35	.30	.25	.20	.15	.10	.05	.00
2	1.06	1.10	1.14	1.18	1.22	1.26	1.29	1.33	1.36
3	0.97	1.01	1.04	1.08	1.12	1.15	1.18	1.22	1.25
4	0.93	0.96	1.00	1.04	1.07	1.10	1.14	1.17	1.20
5	0.90	0.94	0.98	1.01	1.04	1.08	1.11	1.14	1.17
6	0.89	0.93	0.97	1.00	1.03	1.06	1.09	1.12	1.15
7	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
8	0.88	0.92	0.95	0.99	1.02	1.05	1.08	1.11	1.14
9	0.88	0.92	0.95	0.98	1.02	1.05	1.08	1.11	1.13
10	0.88	0.92	0.95	0.98	1.01	1.04	1.07	1.10	1.13
11	0.88	0.91	0.95	0.98	1.01	1.05	1.08	1.10	1.13
12	0.88	0.91	0.95	0.98	1.01	1.04	1.07	1.10	1.13
13	0.88	0.91	0.95	0.98	1.01	1.05	1.08	1.10	1.13
14	0.88	0.91	0.95	0.98	1.01	1.05	1.08	1.10	1.13
15	0.88	0.92	0.95	0.98	1.02	1.05	1.08	1.11	1.13
16	0.88	0.92	0.95	0.99	1.02	1.05	1.08	1.11	1.14
17	0.88	0.92	0.95	0.99	1.02	1.05	1.08	1.11	1.14
18	0.88	0.92	0.95	0.99	1.02	1.05	1.08	1.11	1.14
19	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
20	0.88	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
21	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
22	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
23	0.89	0.92	0.96	0.99	1.02	1.06	1.09	1.11	1.14
24	0.88	0.92	0.96	0.99	1.02	1.05	1.09	1.12	1.15
25	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
26	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
27	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
28	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
29	0.89	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14
30	0.88	0.92	0.96	0.99	1.02	1.05	1.08	1.11	1.14

Notes: The table reports relative standard errors for covariate-matching and propensity-score estimates of a constant-treatment-effect homoskedastic model. Cell size is random. The standard errors were calculated by Monte Carlo integration of analytic formulas that condition on cell sizes and number treated. The probability of treatment is 1/10.

Without loss of generality, we may normalize $\sigma_\varepsilon^2 = 1$. Note that (i) $\xi^* \to 1$ as $M \to \infty$ ($b^* \approx b_c$)[9] and (ii) $\xi^* \to 0$ as $\sigma_\alpha^2 \to 0$ ($b^* \approx b_p$). In other words, the linear-combination estimator converges to the covariate matching estimator as the cell size gets large and/or the between-cell variance gets large ($\sigma_\alpha^2 \to \infty$). On the other hand, the linear-combination estimator converges to the propensity-score matching estimator as the random-effects variance gets small ($\sigma_\alpha^2 \to 0$). This is analogous to the behavior of the random-effects GLS estimator for traditional panel models with constant treatment effects: GLS converges to the within estimator as the time series dimension gets large and/or the variance of the random individual effects gets small, whereas convergence to the pooled estimator occurs in the opposite case.[10]

[9] This uses the fact that

$$\frac{Mg(\pi, M)}{[1 - \pi^M - (1 - \pi)^M]^2} - \frac{1}{\pi} \to 0, \quad \frac{Mg(1 - \pi, M)}{[1 - \pi^M - (1 - \pi)^M]^2} - \frac{1}{(1 - \pi)} \to 0.$$

Proof is available upon request.

[10] Swamy (1970) derives the maximum likelihood estimator of β assuming normality of (α_k, β_k, ε_{ki}), known error variances, and nonstochastic regressors (D_{ki}). This estimator is efficient under panel asymptotics if the error variances are common across cells. Except under constant treatment effects, the Swamy estimator does not appear to simplify to a linear combination of b_p and b_c.

IV. Validity of the Approximation

The panel-asymptotic results in theorem 1 were derived under stratified sampling, for covariate cells of fixed size. Much of the discussion also relied on the simplifying assumption of constant treatment effects. In this section, we compare the finite-sample behavior predicted by theorem 1 with actual finite-sample behavior under random sampling. We begin with constant treatment effects, because this assumption allows an analytic derivation of the finite-sample variance conditional on cell sizes. We then do a Monte Carlo integration to allow for random cell sizes. Finally, we report the results from Monte Carlo experiments with heterogeneous treatment effects.

We again begin with a constant-treatment-effects model where $y_{0ki} = \alpha_k + \varepsilon_{ki}$ and $y_{1ki} = \beta_0 + \alpha_k + \varepsilon_{ki}$, with $\text{Var}(\varepsilon_{ki}) = \sigma_\varepsilon^2$ and $\text{Var}(\alpha_k) = \sigma_\alpha^2$. We assume that there are

TABLE 6.—MONTE CARLO RELATIVE PERFORMANCE [SE(b_c)/SE(b_p)] OF b_p ($\pi = 0.5$)

Avg. Cell Size (M^*)	Covariate R^2								
	.40	.35	.30	.25	.20	.15	.10	.05	.00
2	1.04	1.08	1.12	1.16	1.20	1.24	1.27	1.31	1.34
3	0.95	0.99	1.03	1.06	1.10	1.13	1.16	1.19	1.22
4	0.91	0.94	0.98	1.01	1.05	1.08	1.11	1.14	1.17
5	0.88	0.92	0.95	0.99	1.02	1.05	1.08	1.11	1.13
6	0.87	0.90	0.93	0.97	1.00	1.03	1.06	1.09	1.12
7	0.85	0.89	0.92	0.95	0.98	1.01	1.04	1.07	1.10
8	0.84	0.88	0.91	0.94	0.97	1.00	1.03	1.06	1.09
9	0.84	0.87	0.90	0.93	0.96	0.99	1.02	1.05	1.08
10	0.83	0.86	0.90	0.93	0.96	0.99	1.02	1.04	1.07
11	0.82	0.86	0.89	0.92	0.95	0.98	1.01	1.04	1.06
12	0.82	0.85	0.88	0.92	0.95	0.97	1.00	1.03	1.06
13	0.82	0.85	0.88	0.91	0.94	0.97	1.00	1.02	1.05
14	0.81	0.85	0.88	0.91	0.94	0.97	0.99	1.02	1.05
15	0.81	0.84	0.88	0.91	0.93	0.96	0.99	1.02	1.04
16	0.81	0.84	0.87	0.90	0.93	0.96	0.99	1.01	1.04
17	0.80	0.84	0.87	0.90	0.93	0.96	0.98	1.01	1.04
18	0.80	0.84	0.87	0.90	0.93	0.95	0.98	1.01	1.03
19	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.01	1.03
20	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.00	1.03
21	0.80	0.83	0.86	0.89	0.92	0.95	0.98	1.00	1.03
22	0.80	0.83	0.86	0.89	0.92	0.95	0.97	1.00	1.03
23	0.80	0.83	0.86	0.89	0.92	0.95	0.97	1.00	1.03
24	0.79	0.83	0.86	0.89	0.92	0.94	0.97	1.00	1.02
25	0.79	0.83	0.86	0.89	0.92	0.94	0.97	1.00	1.02
26	0.79	0.83	0.86	0.89	0.91	0.94	0.97	1.00	1.02
27	0.79	0.82	0.85	0.88	0.91	0.94	0.97	1.00	1.02
28	0.79	0.82	0.85	0.88	0.91	0.94	0.97	0.99	1.02
29	0.79	0.82	0.85	0.88	0.91	0.94	0.97	0.99	1.02
30	0.79	0.82	0.85	0.88	0.91	0.94	0.97	0.99	1.02

Notes: The table reports relative standard errors for covariate-matching and propensity-score estimates of a constant-treatment-effect homoskedastic model. Cell size is random. The standard errors were calculated by Monte Carlo integration of analytic formulas that condition on cell sizes and number treated. The probability of treatment is 1/2.

K cells, with cell sizes equal to $M_1, \ldots, M_K$. Let M^* denote the average cell size. Note that

$$b_c = \beta_0 + \frac{1}{\sum_{k=1}^{K} M_k 1(1 \le n_{1k} \le M_k - 1)} \times \sum_{k=1}^{K} \left[M_k 1(1 \le n_{1k} \le M_k - 1) \times \left(\frac{1}{n_{1k}} \sum_{i=1}^{n_k} \varepsilon_{ki} - \frac{1}{M_k - n_{1k}} \sum_{i=n_k+1}^{M_k} \varepsilon_{ki} \right) \right],$$

with conditional variance given $(M_1, \ldots, M_K, n_{11}, \ldots, n_{1K})$ equal to

$$\sigma_\varepsilon^2 \sum_{k=1}^{K} \left[\left(\frac{M_k}{\sum_{k=1}^{K} M_k 1(1 \le n_{1k} \le M_k - 1)} \right)^2 \times \left(\frac{1}{n_{1k}} + \frac{1}{M_k - n_{1k}} \right) 1(1 \le n_{1k} \le M_k - 1) \right]. \quad (10)$$

Also, note that

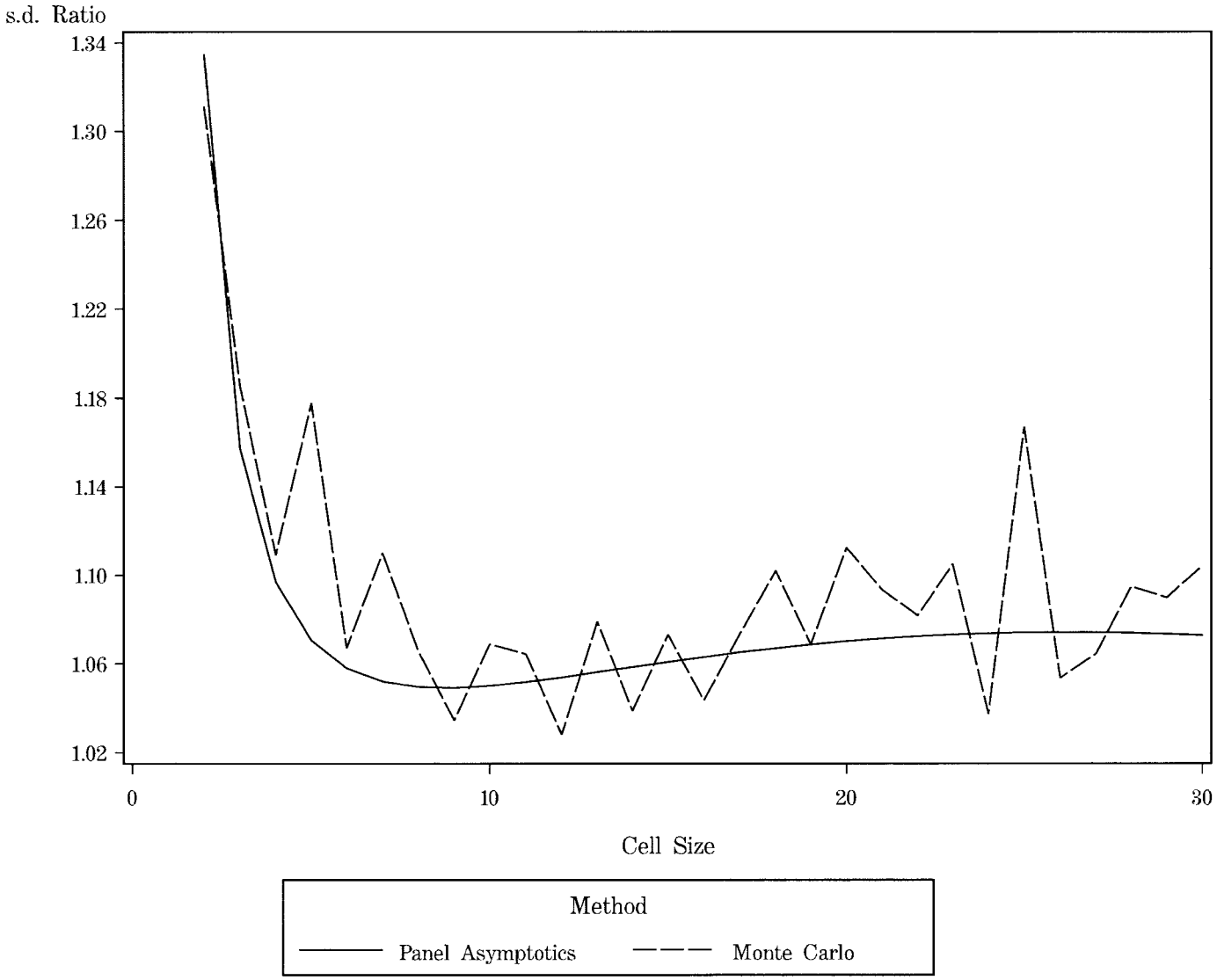
$$b_p = \beta_0 + \sum_{k=1}^{K} \left(\frac{n_{1k}}{\sum_{k=1}^{K} n_{1k}} - \frac{M_k - n_{1k}}{\sum_{k=1}^{K} (M_k - n_{1k})} \right) \alpha_k + \frac{1}{\sum_{k=1}^{K} n_{1k}} \sum_{k=1}^{K} \sum_{i=1}^{n_{1k}} \varepsilon_{ki} - \frac{1}{\sum_{k=1}^{K} (M_k - n_{1k})} \sum_{k=1}^{K} \sum_{i=n_{1k}+1}^{M_k} \varepsilon_{ki},$$

with conditional variance equal to

$$\sigma_\alpha^2 \sum_{k=1}^{K} \left[\left(\frac{n_{1k}}{\sum_{k=1}^{K} n_{1k}} - \frac{M_k - n_{1k}}{\sum_{k=1}^{K} (M_k - n_{1k})} \right) \right]^2 + \sigma_\varepsilon^2 \left(\frac{1}{\sum_{k=1}^{K} n_{1k}} + \frac{1}{\sum_{k=1}^{K} (M_k - n_{1k})} \right). \quad (11)$$

We set $K = 100$, and assume that $(M_1, \ldots, M_K)$ are generated by a multinomial distribution with equal

FIGURE 1.—PANEL-ASYMPTOTIC AND MONTE CARLO RATIO OF STANDARD ERRORS FOR ESTIMATORS WITH AND WITHOUT COVARIATES, RANDOM TREATMENT EFFECT



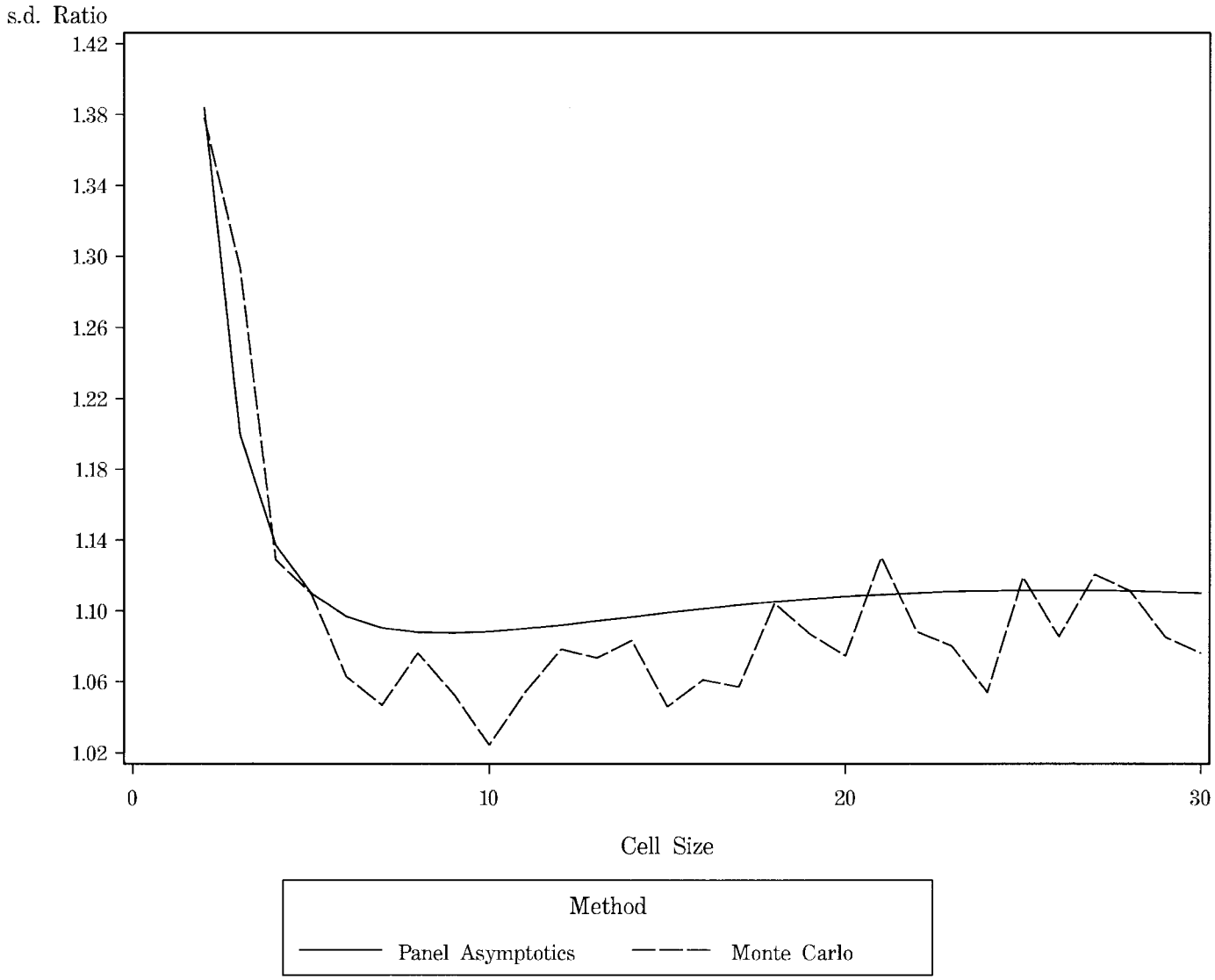
Notes: The figure plots the standard error of b_p divided by the standard error of b_c as a function of average cell size. The Monte Carlo design used 500 replications and 30 cells. The covariate R^2 was fixed at .1, and the treatment probability was fixed at .1. The model incorporates a heterogeneous treatment effect equal to 0 or 1 with probability $\frac{1}{2}$.

weights.[11] Results using 300 replications of ($M_1, \ldots, M_K$, $n_{11}, \ldots, n_{1K}$) to integrate equations (10) and (11) are reported in Tables 5 and 6. As before, we observe that the relative efficiency of b_p, measured by $\sqrt{\text{Var}(b_c)}/\sqrt{\text{Var}(b_p)}$, typically increases as R^2 falls, M^* falls, or π falls, and that b_p is actually more efficient than b_c for some (π, M^*, R^2) combinations. The relative efficiency calculated allowing for random cell sizes is remarkably close to the ratio in tables 3 and 4, calculated using our panel-asymptotic sequence.

[11] We fix the total sample size, then break the sample up into K subsamples with expected size $(\sum_{k=1}^{K} M_k)/K$. This is equivalent to random sampling from a multinomial distribution where $\Pr(X_i = x^{(k)}) = p$ for all j.

We also conducted a small Monte Carlo study of a model with heterogeneous treatment effects. Figures 1 and 2 compare Monte Carlo sampling distributions under random sampling with heterogeneous treatment effects with the corresponding panel-asymptotic approximation. Again, we consider a model where $y_{0ki} = \alpha_k + \varepsilon_{ki}$ and $y_{1ki} = \beta_k + \alpha_k + \varepsilon_{ki}$, with $\text{Var}(\varepsilon_{ki}) = \sigma_\varepsilon^2$ and $\text{Var}(\alpha_k) = \sigma_\alpha^2$. Both figures set $K = 30$, the covariate $R^2 = .1$, and the propensity score = 0.1. We used 500 Monte Carlo replications. Figure 1 shows results from a model where $\beta_k \sim \text{Binomial}(1, \frac{1}{2})$ independent of α_k and ε_{ki}. Figure 2 shows results from a model where $\alpha_k \sim \mathcal{N}(\alpha, \sigma_\alpha^2)$ and $\beta_k = 1(\alpha_k < \alpha)$. In this case, treatment effects are negatively correlated with untreated out-

FIGURE 2.—PANEL-ASYMPTOTIC AND MONTE CARLO RATIO OF STANDARD ERRORS FOR ESTIMATORS WITH AND WITHOUT COVARIATES, CORRELATED TREATMENT EFFECT



Notes: The figure plots the standard error of b_p divided by the standard error of b_c as a function of average cell size. The Monte Carlo design used 500 replications and 30 cells. The covariate R^2 was fixed at .1, and the treatment probability was fixed at .1. The model incorporates a heterogeneous treatment effect equal to 0 or 1, negatively correlated with y_{0k}.

comes. The panel-asymptotic approximation predicts the Monte Carlo efficiency ratio reasonably well in both figures.

V. Summary and Conclusions

Asymptotic theory provides a powerful and flexible tool for the analysis of the theoretical properties of alternative estimators, but empirical researchers have become increasingly aware that conventional asymptotic results can be misleading. In this paper, we develop a framework that improves on conventional asymptotic results about whether to control for covariates in the estimation of treatment effects. In cases that seem likely to be of practical importance, matching on the propensity score, which suffices to eliminate bias, is also more efficient than full covariate matching. Our panel data framework shows that there is more to this result than the possibility of missing cells in covariate matching schemes. The results presented here, based on an analogy with random-effects models for panel data, provide some general guidelines for when full covariate matching is counterproductive. In future work we hope to make these guidelines more specific, and to develop sharper results on efficiency bounds for random-effects estimators of the type introduced here.

As a caveat, it should be emphasized that our asymptotic results were derived under the assumption of equal-sized covariate cells. Although this assumption appears to be harmless for comparing various point estimators, we have not established that the asymptotic variance estimators derived under this sequence are accurate for the construction

of confidence intervals or hypothesis testing. Establishing this is also a task we leave for future work.

REFERENCES

Abadie, A., and G. Imbens, "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," Kennedy School of Government mimeograph (2002).

Angrist, J. D., "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica* 66 (1998), 249–288.

Angrist, J. D., and J. Hahn, "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," NBER technical working paper 241 (1999).

Angrist, J. D., and A. B. Krueger, "Split-Sample Instrumental Variables Estimates of the Returns to Schooling," *Journal of Business and Economic Statistics* 13 (1995), 225–235.

Ashenfelter, O. A., and D. E. Card, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," this REVIEW, 67 (1985), 648–660.

Bekker, P. A., "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica* 62 (1994), 657–681.

Bekker, P. A., and J. van der Ploeg, "Instrumental Variables Estimation Based on Grouped Data," University of Groningen mimeograph (1996).

Card, D. E., and D. Sullivan, "Measuring the Effect of Subsidized Training on Movements In and Out of Employment," *Econometrica* 56 (1988), 497–530.

Chamberlain, G., "Panel Data" in Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, vol. II (Amsterdam: Elsevier, 1984).

——— "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34 (1987), 305–334.

——— "Efficiency Bounds for Semiparametric Regression," *Econometrica* 60 (1992), 567–596.

Chamberlain, G., and Z. Griliches, "Unobservables with a Variance-Components Structure: Ability, Schooling, and the Economic Success of Brothers," *International Economic Review* 16 (1975), 422–449.

Chamberlain, G., and G. Imbens, "Hierarchical Bayes Models with Many Instrumental Variables," National Bureau of Economic Research technical working paper 204 (1996).

Cochran, W. G., "The Planning of Observational Studies of Human Populations (with Discussion)," *Journal of the Royal Statistical Society, Series A* 128 (1965), 234–266.

Dehejia, R. H., and S. Wahba, "Propensity Score Matching Methods for Non-experimental Causal Studies," *Journal of the American Statistical Association* 94 (1999), 1053–1062.

Hahn, J., "On the Role of the Propensity Score in the Efficient Estimation of Average Treatment Effects," *Econometrica* 66 (1998), 315–332.

Hausman, J. A., and W. E. Taylor, "Panel Data and Unobservable Individual Effects," *Econometrica* 49 (1981), 1377–1398.

Heckman, J. J., H. Ichimura, and P. E. Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64 (1997), 605–654.

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

Hirano, K., G. Imbens, and G. Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," UCLA mimeograph (1999).

Imbens, G. W., D. B. Rubin, and B. I. Sacerdote, "Estimating the Effect of Unearned Income on Labor Supply, Earnings, Savings and Consumption: Evidence from a Sample of Lottery Players," *American Economic Review* 91 (2001), 778–794.

Maddala, G. S., "The Use of Variance Components Models in Pooling Cross Section and Time Series Data," *Econometrica* 39 (1971), 341–358.

Mundlak, Y., "On the Pooling of Time Series and Cross Section Data," *Econometrica* 46 (1978), 69–85.

Robins, J., and Y. Ritov, "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine* 16 (1997), 285–319.

Rosenbaum, P. R., *Observational Studies* (New York: Springer-Verlag, 1995).

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

——— "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association* 79 (1984), 516–524.

——— "The Bias Due to Incomplete Matching," *Biometrics* 41 (1985a), 103–116.

——— "Constructing a Control Group using Multi-variate Matching Methods That Include the Propensity Score," *American Statistician* 39 (1985b), 33–38.

Rubin, D. B., "Matching to Remove Bias in Observational Studies," *Biometrics* 29 (1973), 159–183.

——— "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association* 74 (1979), 318–328.

Swamy, P. A. V. B., "Efficient Inference in a Random Coefficient Regression Model," *Econometrica* 38 (1970), 311–323.

Wallace, T. D., and A. Hussain, "The Use of Error Components Models in Combining Cross Section with Time Series Data," *Econometrica* 37 (1969), 55–72.

APPENDIX

Finite-Sample Variance in the Two-Cell Example

This calculation begins with the bias and variance of the two estimators, conditional on n_{11} and n_{12}, for cases where both estimators are defined. b_p is conditionally biased, though b_c is not.

1. b_c

Case 1. In this case, both cells contain treated and control observations. The conditional distribution of b_c given (n_{11}, n_{12}) has bias 0 and variance

$$\frac{1}{4}\left(\frac{1}{n_{11}}+\frac{1}{M-n_{11}}+\frac{1}{n_{12}}+\frac{1}{M-n_{12}}\right).$$

Therefore, the conditional mean squared error is

$$w_1(n_{11}, n_{12}) \equiv \frac{1}{4}\left(\frac{1}{n_{11}}+\frac{1}{M-n_{11}}+\frac{1}{n_{12}}+\frac{1}{M-n_{12}}\right).$$

Case 2. In this case, one of the two cells consists of treated or controls only. If the first cell is discarded but the second one is not, the conditional distribution of b_c is such that the bias is 0 and the variance is equal to $1/n_{12} + 1/(M - n_{12})$. Therefore, the conditional mean squared error is

$$w_2(n_{12}) \equiv \frac{1}{n_{12}}+\frac{1}{M-n_{12}}.$$

Similar comments apply when the second cell is discarded.

Case 3. In this case, each cell consists of treated or controls only. Because this happens in each cell with probability $\Pr(n_{1k} = 0 \text{ or } M) = \pi^M + (1 - \pi)^M$, with probability $[\pi^M + (1 - \pi)^M]^2$ the covariate-matching estimator is undefined.

Now, integrate over the distribution of n_{11} and n_{12}, using the fact that they are independent Binomial(M, π) random variables:

$$\sum_{n_{11}=1}^{M-1}\sum_{n_{12}=1}^{M-1} w_1(n_{11}, n_{12})\binom{M}{n_{11}}\pi^{n_{11}}(1-\pi)^{M-n_{11}}\binom{M}{n_{12}}\pi^{n_{12}}(1-\pi)^{M-n_{12}12}$$

$$+ [\pi^M + (1-\pi)^M]\sum_{n_{12}=1}^{M-1} w_2(n_{12})\binom{M}{n_{12}}\pi^{n_{12}}(1-\pi)^{M-n_{12}}$$

$$+ [\pi^M + (1-\pi)^M]\sum_{n_{11}=1}^{M-1} w_2(n_{11})\binom{M}{n_{11}}\pi^{n_{11}}(1-\pi)^{M-n_{11}}.$$

Dividing the above expression by $1 - [\pi^M + (1 - \pi)^M]^2$, we obtain the variance of interest.

2. b_p

We consider the finite-sample distribution of b_p for cases where b_c can be computed.[12]

Case 1. In this case,

$$b_p = \frac{1}{\sum_{k=1}^{2} n_{1k}} \sum_{k=1}^{2} \sum_{i=1}^{n_{1k}} (\alpha_k + \beta + \varepsilon_{ki}) - \frac{1}{2M - \sum_{k=1}^{2} n_{1k}} \sum_{k=1}^{2} \sum_{i=n_{1k}+1}^{M} (\alpha_k + \varepsilon_{ki})$$

with conditional bias $[M(n_{11} - n_{12})/(n_{11} + n_{12})(2M - n_{11} - n_{12})](\alpha_1 - \alpha_2)$, and variance $2M/(n_{11} + n_{12})[2M - (n_{11} + n_{12})]$. Therefore, the conditional mean squared error is given by

$$\left(\frac{M(n_{11} - n_{12})}{(n_{11} + n_{12})(2M - n_{11} - n_{12})} (\alpha_1 - \alpha_2) \right)^2 + \frac{2M}{(n_{11} + n_{12})(2M - (n_{11} + n_{12}))}.$$

Case 2(i). Consider the case where the first cell consists of all treated, but the second cell is not dropped by b_c. We then have

$$b_p = \frac{1}{M + n_{12}} \left(\sum_{i=1}^{M} (\alpha_1 + \beta_0 + \varepsilon_{1i}) + \sum_{i=1}^{n_{12}} (\alpha_2 + \beta + \varepsilon_{2i}) \right) - \frac{1}{M - n_{12}} \sum_{i=n_{12}+1}^{M} (\alpha_2 + \varepsilon_{2i}),$$

which has conditional bias $[M/(M + n_{12})](\alpha_1 - \alpha_2)$, and variance $1/(M + n_{12}) + 1/(M - n_{12})$. Therefore, the conditional mean squared error is given by

$$\left(\frac{M}{M + n_{12}} (\alpha_1 - \alpha_2) \right)^2 + \frac{1}{M + n_{12}} + \frac{1}{M - n_{12}}.$$

Similar comments apply when the second cell consists of all treated.

Case 2(ii). Consider the case where the first cell consists of all controls, but the second cell is not dropped by the covariate estimator. We then have

$$b_p = \frac{1}{n_{12}} \sum_{i=1}^{n_{12}} (\alpha_2 + \beta_0 + \varepsilon_{2i}) - \frac{1}{2M - n_{12}} \left(\sum_{i=1}^{M} (\alpha_1 + \varepsilon_{1i}) + \sum_{i=n_{12}+1}^{M} (\alpha_2 + \varepsilon_{2i}) \right),$$

which has the conditional bias $-[M/(2M - n_{12})](\alpha_1 - \alpha_2)$, and variance $1/(2M - n_{12}) + 1/n_{12}$. Therefore, the conditional mean squared error is given by

$$\left(- \frac{M}{2M - n_{12}} (\alpha_1 - \alpha_2) \right)^2 + \frac{1}{2M - n_{12}} + \frac{1}{n_{12}}.$$

Similar comments apply when the second cell consists of all controls.

Again, the mean squared error of interest is computed by integrating the mean squared error with respect to the distribution of (n_{11}, n_{12}) and dividing by $1 - [\pi^M + (1 - \pi)^M]^2$.

[12] We ignore the case where the first cell consists of all treated and the second cell consists of all control observations.