CHAPTER 32

Making Effects Manifest in Randomized Experiments

Jake Bowers

Experimentalists want precise estimates of treatment effects and nearly always care about how treatment effects may differ across subgroups. After data collection, concern may turn to random imbalance between treatment groups on substantively important variables. Pursuit of these three goals – enhanced precision, understanding treatment effect heterogeneity, and imbalance adjustment – requires background information about experimental units. For example, one may group similar observations on the basis of such variables and then assign treatment within those blocks. Use of covariates after data have been collected raises extra concerns and requires special justification. For example, standard regression tables only approximate the statistical inference that experimentalists desire. The standard linear model may also mislead via extrapolation. After providing some general background about how covariates may, in principle, enable pursuit of precision and statistical adjustment, this chapter presents two alternative approaches to covariance adjustment: one using modern matching techniques and another using the linear model – both use randomization as the basis for statistical inference.

1. What Is a Manifest Effect?

A manifest effect is one we can distinguish from zero. Of course, we cannot talk formally about the effects of an experimental treatment as manifest without referring to probability: a scientist asks, "Could this result have occurred merely through chance?" or "If the true effect were zero, what is the chance that we'd observe an effect as large as this?" More formally, for a frequentist, saying that a treatment effect is manifest means that the statistic we observe casts a great deal of doubt on a hypothesis of no effects. We are most likely to say that some observed effect casts doubt on the null hypothesis of no effect when we have a large sample and/or when noise in the

Many thanks to Jamie Druckman, Kevin Esterling, Don Green, Ben Hansen, Don Kinder, Jim Kuklinski, Thomas Leeper, Costas Panagopoulos, and Cara Wong. Parts of this work were funded by National Science Foundation (NSF) Grant Nos. SES-0753168 and SES-0753164.

outcome that might otherwise drown out the signal in our study has been well controlled. Fisher (1935) reminds us that although randomization alone is sufficient for a valid test of the null hypothesis of no effect, specific features of a given design allow equally valid tests to differ in their ability to make a treatment effect manifest:

> *With respect to the refinements of technique [uses of covariates in the planning of an experiment], we have seen above that these contribute nothing to the validity of the experiment, and of the test of significance by which we determine its result. They may, however, be important, and even essential, in permitting the phenomenon under test to manifest itself.* (24)

Of course, one would prefer a narrow confidence interval to a wide confidence interval, even if both excluded the hypothesis of no effects. As a general rule, more information yields more precision of estimation. One may increase information in a design by gathering more observations and/or gathering more data about each observation. This paper considers covariates as a refinement of technique to make treatment effects manifest in randomized studies. I focus first on simple uses of covariates in design and then offer some ideas about their use in post-treatment adjustment.

What Are Covariates? How Should We Use Them in Experiments?

A covariate is a piece of background information about an experimental unit – a variable unchanged and unchangeable by the experimental manipulation. Such variables might record the groups across which treatment effects ought to differ according to the theory motivating and addressed by the experimental design (e.g., men and women ought to react differently to the treatment) or might provide information about the outcomes of the experiment (e.g., men and women might be expected to have somewhat different outcomes even if reactions to treatment are expected to be the same across both groups). Covariates may be used profitably in experiments either *before* treatment is assigned (by creating subgroups of units within which treatment will be randomized during the recruitment and sample design of a study) and/or *after* treatment has been assigned and administered and outcomes measured (by creating subgroups within which outcomes ought to be homogeneous or adjusting for covariates using linear models).

Common Uses (and Potential Abuses) of Covariates in the Workflow of Randomized Experimentation

Every textbook on the design of experiments is, in essence, a book about the use of covariates in the design and analysis of experiments. This chapter does not substitute for such sources. For the newcomer to experiments, I summarize in broad strokes, and with minimal citations, the uses to which covariates may be put in the design and analysis of randomized experiments. After this summary, I offer a perspective on the use of covariates in randomized experimentation that, in fundamental ways, is the same as that found in Fisher (1925, 1935), Cochran and Cox (1957), Cox (1958), and Cox and Reid (2000). I differ from these previous scholars in hewing more closely and explicitly to 1) the now well-known potential outcomes framework for causal inference (Rubin 1974, 1990; Nayman 1990; Brady 2008; Sekhon 2008) and 2) randomization as the basis for statistical inference.

COVARIATES ALLOW PRECISION ENHANCEMENT

Blocking on background variables before treatment assignment allows the experimenter to create subexperiments within which the units are particularly similar in their outcomes, and adjustment using covariates after data have been collected may also reduce nontreatment-related variation in outcomes. In both cases, covariates can reduce noise that might otherwise obscure the effects of the treatment.

Of course, such precision enhancements arrive with some costs: implementing a blocking plan may be difficult if background information on experimental units is not available before recruitment/arrival at the lab (but see Nickerson 2005); care must be taken to reflect the blocking in the estimation of treatment effects to avoid bias and to take advantage of the precision enhancements offered by the design; and, in principle, analysts can mislead themselves by performing many differently adjusted hypothesis tests until they reject the null of no effects, even when the treatment has no effect.

COVARIATES ENABLE SUBGROUP ANALYSES

When theory implies differences in treatment effects across subgroups, subgroup membership must be recorded, and, if possible, the experiment ought to be designed to enhance the ability of the analyst to distinguish group differences in treatment effects. Covariates on subgroups may also be quite useful for post hoc exploratory analyses designed not to cast doubt on common knowledge, but to suggest further avenues for theory.

COVARIATES ALLOW ADJUSTMENTS FOR RANDOM IMBALANCE

All experiments may display random imbalance. Such baseline differences can arise even if the randomization itself is not suspect: recall that one of twenty unbiased hypothesis tests will reject the null of no difference at the predetermined error rate of $\alpha = .05$ merely due to chance. An omnibus balance assessment such as that proposed by Hansen and Bowers (2008) is immune from this problem, but any unbiased one-by-one balance assessment will show imbalance in 100α percent of the covariates tested. Thus, I call this problem "random imbalance" to emphasize that the imbalance could easily be due to chance and need not cast doubt on the randomization or administration of a study (although discovery of extensive imbalance might suggest that scrutiny of the randomization and administration is warranted).

Random imbalance in a well-randomized study on substantively important covariates may still confuse the reader. In the presence of random imbalance, comparisons of treated to controls will contain *both* the effect of the treatment *and* the differences due to the random imbalance. One may attempt to remove the effects of such covariates from the treatment effect by some form of adjustment. For example, one may use the linear regression model as a way to adjust for covariates, or one may simply group together observations on the basis of the imbalanced covariates. Adjustment, however, raises concerns that estimates of treatment effects may come to depend more on the details of the adjustment method rather than on the randomization and design of the study. Thus, the quandary: either risk known confusions of comparisons or bear the burden of defending and assessing an adjustment method. An obvious strategy to counter concerns about cherry-picking results or modeling artifacts is to present both adjusted and unadjusted results and to specify adjustment strategies before randomization occurs.

Randomization Is the Primary Basis for Statistical Inference in Experiments

A discussion of manifest effects is also a discussion of statistical inference: statistical tests quantify doubt against hypotheses, and a manifest effect is evidence that casts great doubt on the null of no effects. On what basis can we justify statistical tests for experiments?

In national surveys, we draw random samples. Statistical theory tells us that the mean in the sample is an unbiased estimator of the mean in the population as long as we correctly account for the process by which we drew the sample in our estimation. That is, in a national survey, our *target of inference* is often (but not always) the population from which the sample was drawn, and we are *justified* in so inferring by the sampling design.

In other studies, we may not know how a sample was drawn (either we have no well-defined population or no knowledge of the sampling process or both). But we may know how our observed outcome was generated: say we know that at the microlevel our outcome was created from discrete events occurring independently of each other in time. In that case, we would be justified in claiming that our population was created via a Poisson process: in essence, we have a population-generating machine, data-generating process, or model of outcomes as the target of our inference.

Now, what about randomized experiments? Although we might want to infer to a model, or even to a population, the strength of experiments is inference to a counterfactual. The primary targets of inference in a randomized experiment are the experimental treatment groups: we infer from one to another. Randomization makes this inference meaningful. But, randomization can also justify the statistical inference as well. The mean in the treatment group is a good estimator for what we would expect to observe if all of the experimental units were treated. The treatment group in a randomized study is a random sample from the finite "population" of the experimental pool.[1]

All standard textbooks note this fact, but they also point out that estimating causal effects using randomization-based theory can be mathematically inconvenient or computationally intensive and that, thus, using the large-sample sampling theory (and/or normal distribution models) turns out to provide very good approximations to the randomization-based results most of the time. Since the 1930s, the computational apparatus of randomization-based inference has expanded, as has its theoretical basis and applied reach. In this chapter, the statistical inference I present is randomization-based, even if most of it also uses large-sample theory. For example, it takes no more time to execute a randomization-based test of the null hypothesis of no effect using mean differences than it does using the linear regression model-based approximation.[2]

Recently, Freedman (2008a, 2008b, 2008c) reminded us that the target of inference in randomized experiments was the counterfactual, and he noted that linear regression and logistic regression were not theoretically justified on this basis. Green (2009) and Schochet (2009) remind us that linear regression can often be an excellent approximation to the randomization-based difference of means. The need for this exchange is obvious, even if it is echoed in the early textbooks. Those early authors moved quickly to the technicalities of the approximations rather than dwelling on the then uncomputable but theoretically justifiable procedures. As experimentation explodes as a methodology in political science, we are seeing many more small experiments, designs where randomization is merely a (possibly weak) instrument, and theories implying very heterogeneous treatment effects. I expect we will find more of these along with many more large studies with fairly normal-looking outcomes where treatment plausibly just shifts the normal curve of the treated away from the normal curve of the controls. Rather than hope that the linear model approximation works well, this chapter presents analyses that do not require that approximation and thereby offers others the ability to check the approximation.

2. Strategies for Enhancing Precision before Assignment

> *[W]e consider some ways of reducing the effect of uncontrolled variations on the error of the treatment comparisons. The general idea is the common sense one of grouping the units into sets, all the units in a set being as alike*

---

[1] See Rosenbaum (2002b, ch. 2) and Bowers and Panagopoulos (2009) for accessible introductions to randomization inference; a mode of inference developed in different yet compatible ways by Fisher (1935, ch. 2) (as a method of testing) and Neyman (1990) (as a method of estimating mean differences). In this paper, I follow the Fisher-style approach in which causal effects are inferred from testing hypotheses rather than estimated as points. Both methods (producing a plausible interval for causal effects using a point ± a range or testing a sequence of hypotheses) often produce identical confidence intervals.

[2] This chapter is written in the mixture of R and $\LaTeX$ known as Sweave (Leisch 2002, 2005) and, as such, the code required to produce the results, tables, and figures (as well as additional analyses not reported) are available for learning and exploration at http://jakebowers.org. Thus, I spend relatively little time discussing the details of the different methods, assuming that those interested in learning more will download the source code of the chapter and adapt it for their own purposes.

> *as possible, and then assigning the treatments so that each occurs once in each set. All comparisons are then made within sets of similar units. The success of the method in reducing error depends on using general knowledge of the experimental material to make an appropriate grouping of the units into sets. (Cox 1958, 23)*

We have long known that covariates enhance the precision of estimation to the extent that they predict outcomes. This section aims to make this intuition more concrete in the context of a randomized experiment.

An Example by Simulation

Imagine that we desire to calculate a difference of means. In this instance, we are using a fixed covariate x, and the purpose of this difference in means is to execute a placebo test or a balance test, not to assess the causal effects of a treatment. Imagine two scenarios, one in which a binary treatment $Z_{ib} = 1$ is assigned to $m_b = 1$ subject within each of B pairs $b = 1, \ldots, B$; $i = 1, \ldots, n_b$, $B \leq n; n = \sum_{b=1}^{B} n_b$ (for pairs $n = 2B$ and, thus, $n_b = 2$), and another in which a binary treatment $Z_i = 1$ is assigned to $m = n - m = (n/2)$, subjects $i = 1, \ldots, n$ without any blocking. Consider the test statistics

$$\begin{aligned} d_{\text{pairs}} &= \frac{1}{B}\sum_{b=1}^{B}\left(\sum_{i=1}^{nb} Z_{ib}x_{ib}/m_b - \sum_{i=1}^{n_b}(1 - Z_{ib})x_{ib}/(n_b - m_b)\right) \\ &= \frac{1}{B}\sum_{b=1}^{B}\left(\sum_{i=1}^{2}(Z_{ib}x_{ib} - (1 - Z_{ib})x_{ib})\right), \end{aligned} \tag{1}$$

which reduces to the difference in means of x between treated and control units within pairs summed across pairs and

$$d_{\text{no pairs}} = \sum_{i=1}^{n} Z_i x_i/m - \sum_{i=1}^{n}(1 - Z_i)x_i/(n - m), \tag{2}$$

which sums across all units within control and treatment conditions. These two quantities are the same, even if one is written as an average over B pairs, because pairs are blocks of equal size and, therefore, each block-specific quantity ought to contribute equally to the sum.

The theory of simple random sampling without replacement suggests that the variances of these statistics differ. First,

$$\begin{aligned} \text{Var}(d_{\text{no pairs}}) &= \frac{n}{m(n-m)}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n - 1} \\ &= \left(\frac{4}{n}\right)\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2. \end{aligned} \tag{3}$$

Second,

$$\begin{aligned} \text{Var}(d_{\text{pairs}}) &= \left(\frac{1}{B}\right)^2 \sum_{b=1}^{B}\frac{n_b}{m_b(n_b - m_b)} \times \sum_{i=1}^{n_b}\frac{(x_{ib} - \bar{x}_b)^2}{n_b - 1} \\ &= \left(\frac{2}{B^2}\right)\sum_{b=1}^{B}\sum_{i=1}^{2}(x_{ib} - \bar{x}_b)^2. \end{aligned} \tag{4}$$

If pairs were created on the basis of similarity on x, then $\text{Var}(d_{\text{no pairs}}) > \text{Var}(d_{\text{pairs}})$ because $\sum_{i=1}^{n}(x_i - \bar{x})^2 > \sum_{b=1}^{B}\sum_{i=1}^{2}(x_{ib} - \bar{x}_b)^2$. Any given x_i will be farther from the overall mean ($\bar{x}$) than it would be from the mean of its pair ($\bar{x}_b$). Note also that the constants multiplying the sums are $(4/n(n-1))$ in the unpaired case and $8/n^2$ (because $B = (n/2)$) in the paired case. As long as $n > 2$, $(4/(n^2 - n)) < (8/n^2)$, and this difference diminishes as n increases. Thus, benefits of pairing can diminish as the size of the experiment increases as long as within-pair homogeneity does not depend on sample size.

To dramatize the benefits possible from blocking, I include a simple simulation study based roughly on real data from a field experiment of newspaper advertisements and turnout in U.S. cities (Panagopoulos 2006). The cities in this study differed in baseline turnout (with baseline turnout ranging from

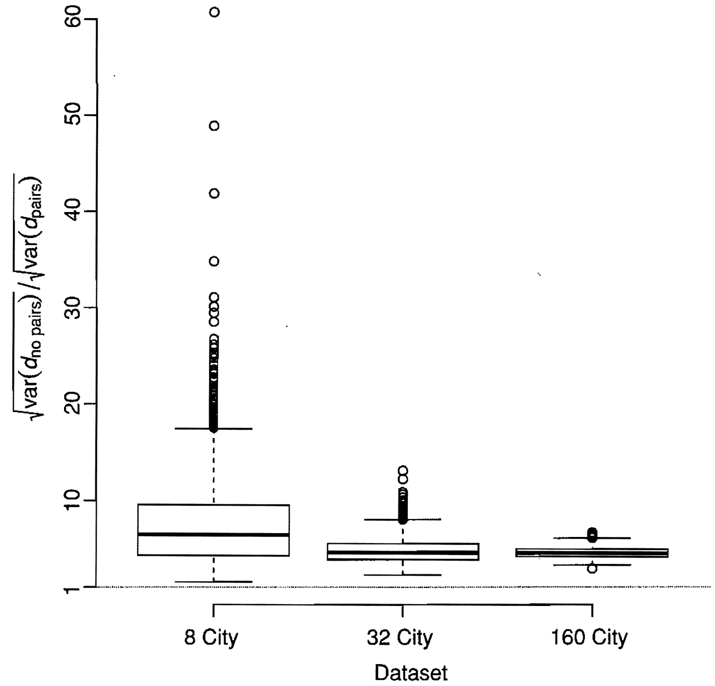

Figure 32.1. Efficiency of Paired and Unpaired Designs in Simulated Turnout Data

roughly ten percent to roughly fifty percent). Panagopoulos paired the cities before randomly assigning treatment, with baseline turnout differences within pair ranging from one to seven percentage points in absolute value. The simulation presented here takes his original 8-city dataset and makes two fake versions, one with 32 cities and another with 160 cities. The expanded datasets were created by adding small amounts of uniform random noise to copies of the original dataset. These new versions of the original dataset maintain the same general relationships between treatment and outcomes and covariates and pairs as the original, but allow us to examine the effects of increasing sample size. In this simulation, the covariate values are created so that the average difference of means within pair is zero. For each of the 1,000 iterations of the simulation and for each dataset ($n = 8, 32, 160$), the procedure was as follows:

Create fake data based on original data *Each pair receives a random draw from a normal distribution with mean equal to the baseline outcome of its control group and standard deviation equal to the standard deviation of the pair. For the original dataset, the within pair differences on baseline turnout of 1, 4, and 7 translate into standard deviations of roughly 0.7, 3, and 5. The "true" difference is required to be zero within every pair, but each pair may have a different baseline level of turnout and a different amount of variation – mirroring the actual field experiment.*

Estimate variance of treatment effects under null *Apply Equations (3) and (4).*[3]

Figure 32.1 shows that estimates of $\sqrt{\text{Var}(d_{\text{no pairs}})}$ were always larger than the estimates for $\sqrt{\text{Var}(d_{\text{pairs}})}$, although the ratio $\sqrt{\text{Var}(d_{\text{no pairs}})}/\sqrt{\text{Var}(d_{\text{pairs}})}$ diminished as the size of the experiment increased. The standard deviation across the simulations for the paired 8-city example (1.8) is close to the average standard deviation for the nonpaired tests in the 160-city example (2.3). That is, these simulations show that a paired experiment of 8 units could be more precise than an unpaired experiment of 160 units (although, of course, a paired experiment of 160 units would be yet more precise [the average null SD in that case is 0.5]). Notice that, in this case, the advantages of the paired design diminish as the size of the experiment increases but do not disappear.

---

[3] The actual computations used the `xBalance` command found in the `RItools` library for R (Bowers, Fredrickson, and Hansen 2009), as described in Hansen and Bowers (2008).

Pairing provides the largest opportunity for enhancement of precision in the comparison of two treatments – after all, it is hard to imagine a set more homogeneous than two subjects nearly identical on baseline covariates (Imai, King, and Nall 2009). And, even if it is possible for an unpaired design to produce differences of means with lower variance than a paired design, it is improbable given common political science measuring instruments.

What do we do once the experiment has run? One can enhance precision using covariates either via post-stratification (grouping units with similar values of covariates) or covariance adjustment (fitting linear models with covariates predicting outcomes). In the first case, analyses proceeds as if prestratified: estimates of treatment effects are made conditional on the chosen grouping. In the second case, linear models "adjust for" covariates – increased precision can result from their inclusion in linear models under the same logic as that taught in any introduction to statistics classes.

3. Balance Assessment: Graphs and Tests

We expect that the bias reduction operating characteristics of random assignment would make the baseline outcomes in the control group comparable to the baseline outcomes in the treatment group. If the distribution of a covariate is similar between treated and control groups, then we say that this covariate is "balanced" or that the experiment is balanced with respect to that covariate. Yet, it is always possible that any given randomization might make one or more observed (or unobserved) covariates imbalanced merely by chance. If the imbalanced covariates seem particularly relevant to substantive interpretations of the outcome (as would be the case with outcomes measured before the treatment was assigned), then we would not want such differences to confuse treatment-versus-control differences of post-treatment outcomes.

One Graphical Mode of Assessment

Figure 32.2 provides both some reassurance and some worry about balance on the baseline outcome in the thirty-two-city data example. The distributions of baseline turnout are nearly the same (by eye) in groups 2 and 3, but differ (again, by eye) in groups 1 and 4. Should these differences cause concern?

The top row of Table 32.1 provides one answer to the question about worries about random imbalance on baseline outcome. We would not be surprised to see a mean difference of $d_{\text{blocks}} = -1.2$ if there were no real difference given this design ($p = .2$). Of course, questions about balance of a study are not answerable by looking at only one variable. Table 32.1 thus shows a set of randomization-based balance tests to assess the null of no difference in means between treated and control groups on covariates one by one and also all together using an omnibus balance test. We easily reject the null of balance on the linear combination of these variables in this case ($p = .00003$), although whether we worry about the evident differences on percent black or median income of the cities may depend somewhat on the extent to which we fear that such factors matter for our outcome – or whether these observed imbalances suggest more substantively worrisome imbalances in variables that we do not observe. Does this mean the experiment is broken? Not necessarily.[4]

---

[4] It would be strange to see such imbalances in a study run with thirty-two observations rather than an eight-observation study expanded to thirty-two observations artificially as done here. These imbalances, however, dramatize and clarify benefits and dangers of post-treatment adjustment as explained throughout this chapter.

Table 32.1: One-by-One Balance Tests for Covariates Adjusted for Blocking in the Blocked Thirty-Two-City Study

x	$\bar{x}_{\text{Ctrl}}$	$\bar{x}_{\text{Trt}}$	d_{blocks}	$\sqrt{\text{Var}(d_{\text{blocks}})}$	Std.d_{blocks}	z	p
Baseline outcome	27.1	25.8	−1.2	0.9	−0.1	−1.4	.2
Percent black	17.3	2.6	−14.7	4.2	−1.2	−3.5	.0
Median HH income	30.6	46.3	15.7	3.2	2.6	4.9	.0
Number of candidates	13.2	10.3	−2.9	2.0	−0.5	−1.5	.1

Notes: Two-sided p values are reported in the p column referring z to a std. normal distribution, approximating the distribution of the mean difference under null of no effects. An omnibus balance test casts doubt on the null hypothesis of balance on linear combinations of these covariates at $p = .00003$. Test statistics (d_{blocks}) are generalizations of Equations (1) and (4) developed in Hansen and Bowers (2008) and implemented in Bowers et al. (2009). Statistical inference (z and p values) is randomization based but uses large-sample normal approximations for convenience. Other difference-of-means tests without the large-sample approximation, and other tests such as Kolmogorov-Smirnov and Wilcoxon rank sum tests, provide the same qualitative interpretations. For example, the p values on the tests of balance for baseline outcome (row 1) ranged from $p = .16$ and $p = .17$ for the simulated and asymptotic difference of means tests, respectively, to $p = .33$ and $p = .72$ for exact and simulated Wilcoxon rank sum and Kolmogorov-Smirnov tests.

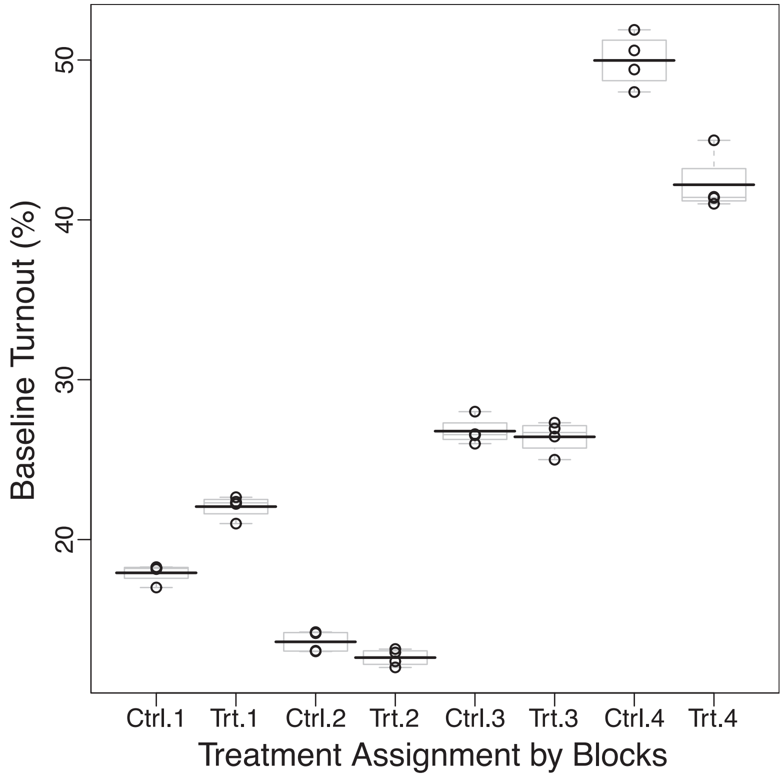

Figure 32.2. Graphical Assessment of Balance on Distributions of Baseline Turnout for the Thirty-Two-City Experiment Data
Notes: Open circles are observed values. Means are long, thick black lines. Boxplots in gray.

Table 32.2: One-by-One Balance Tests for Covariates in the Blocked Eight-City Study

x	$\bar{x}_{\text{Ctrl}}$	$\bar{x}_{\text{Trt}}$	d_{blocks}	$\sqrt{\text{Var}(d_{\text{blocks}})}$	Std.d_{blocks}	z	p
Baseline outcome	26.0	24.8	−1.2	2.0	−0.1	−0.6	.5
Percent black	16.8	2.4	−14.4	11.0	−1.1	−1.3	.2
Median HH income	29.2	44.5	15.4	8.1	2.5	1.9	.1
Number of candidates	13.0	10.0	−3.0	5.1	−0.5	−0.6	.6

Notes: An omnibus balance test casts doubt on the null hypothesis of balance on linear combinations of these covariates at $p = .41$. Test statistics (d_{blocks}) are generalizations of Equations (1) and (4) developed in Hansen and Bowers (2008) and implemented in Bowers et al. (2009). Statistical inference (z and p values) is randomization based, but uses large-sample normal approximations for convenience.

Understanding p Values in Balance Tests

Say we did not observe thirty-two observations in sets of four but only eight observations in pairs. What would our balance tests report then? Table 32.2 shows the results for such tests, analogous to those shown in Table 32.1.

Notice that our p values now quantify less doubt about the null. Is there something wrong with randomization-based tests if, by reducing the size of our experiment, we would change our judgment about the operation of the randomization procedure? The answer here is no.[5]

The p values reported from balance tests used here summarize the extent to which random imbalance is worrisome. With a sample size of eight, the confidence interval for our treatment effect will be large – taking the pairing into account, a ninety-five percent interval will be on the order of ±3.5 (as roughly calculated on the baseline outcomes in Section 2). For example, both Tables 32.1 and 32.2 show the block-adjusted mean difference in percent black between treated and control groups to be roughly 14.5 percentage points. In the thirty-two-city example, this difference cast great doubt against the null of balance, whereas in the eight-city example this difference casts less doubt. Now, evaluating the difference between controls and treatment on actual outcome in the eight-city case gives $d_{\text{pairs}} = 1.5$ and, under the null of no effects, gives $\sqrt{\text{Var}(d_{\text{pairs}})} = 2.6$ – and inverting this test leads to an approximate eighty-eight percent confidence interval of roughly $[-7,7]$: the width of the confidence interval itself is about fourteen points of turnout. Even if percent black were a perfect predictor of turnout (which it is not, with a pair adjusted linear relationship of 0.07 in the eight-city case), the p value of .2 indicates that the relationship with treatment assignment is weak enough, and the confidence intervals on the treatment effect itself would be wide enough, to make any random imbalance from percent black a small concern. That is, the p values reported in Table 32.2 tell us that random imbalances of the sizes seen here will be small relative to the size of the confidence interval calculated on the treatment effect. With a large sample, a small random imbalance is proportionately more worrisome because it is large relative to the standard error of the estimated treatment effect. Given the large confidence intervals on a treatment effect estimated on eight units, the random imbalances shown here are less worrisome – and the p values encode this worry just as they encode the plausibility of the null.

Thus, even though our eyes suggested that we worry about the random imbalance on baseline turnout we saw in Figure 32.2, that amount of imbalance on baseline outcome is to be expected in both the thirty-two and eight-city cases – it is an amount of bias that would have little effect on the treatment effect were we to adjust for it. Of course, the omnibus test for imbalance on all four covariates simultaneously reported in Table 32.1 does cast doubt on the null of balance – and the tests using the *d* statistics in Table 32.1 suggest that the problem is with percent black and median household income rather than baseline outcomes or number of candidates.[6]

---

[5] Echoing Senn (1994), among others, in the clinical trials world, Imai, King, and Stuart (2008) provide some arguments against hypothesis tests for balance in observational studies. Hansen (2008) answers these criticisms with a formal account of the intuition provided here, 1) pointing out that randomization-based tests do not suffer the problems highlighted by those authors, and 2) highlighting the general role that p values play in randomization-based inference.

4. Covariates Allow Adjustment for Random Imbalance

A well-randomized experiment aiming to explain something about political participation showing manifest imbalance on education poses a quandry. If the analyst decides to adjust, then he or she may fall under suspicion: even given a true treatment effect of zero, one adjustment out of many tried will provide a p value casting doubt on the null of no effect merely through chance. Without adjustment, we know how to interpret p values as expressions of doubt about a given hypothesis: low p values cast more doubt than high p values. Adjustment in and of itself does not invalidate this interpretation: a p value is still a p value. Concerns center rather on 1) whether an "adjusted treatment effect" is substantively meaningful and how it relates to different types of units experiencing the treatment in different ways – that is, the concerns center on the meaning of "adjustment" in the context of the adjustment method (a linear model or a post-stratification); and 2) whether a specification search was conducted with only the largest adjusted treatment effect reported, representing a particularly rare or strange configuration of types of units. Such worries do not arise in the absence of adjustment. Yet, if the analyst declines to adjust, then he or she knows that part of the treatment effect in his or her political participation study is due to differences in education, thereby muddying the interpretation of his or her study.

One may answer such concerns by announcing in advance the variables for which random imbalance would be particularly worrisome and also provide a proposed adjustment and assessment plan a priori. Also, if one could separate adjustment from estimation of treatment effects, one may also avoid the problem of data snooping. For example, Bowers and Panagopoulos (2009) propose a power analysis–based method of choosing covariance adjustment specifications that can be executed independently of treatment effect estimation, and it is well known that one may poststratify and/or match without ever inspecting outcomes. Post-stratification may also relieve worries about whether comparisons adjusted using linear models are artifacts of the functional form (Gelman and Hill 2007, ch. 9).

There are two broad categories of statistical adjustment for random imbalance: adjustment by stratification and adjustment using models of outcomes. In both cases, adjustment amounts to choice of weights, and adjustment may be executed entirely to enhance precision even if there is no appreciable evidence of random imbalance. Notice that the "unadjusted" estimate may already be an appropriately weighted combination of block-specific treatment effects – and that to fail to weight (or "adjust") for block-specific probabilities of treatment assignment will confound estimates of average treatment effects (if the probabilities of assignment differ across blocks) and decrease precision (if the variation in the outcomes is much more homogeneous within blocks than across blocks).

---

[6] If we had twenty covariates and rejected the null of balance with $p < .05$, we would expect to falsely see evidence of imbalance in one of twenty covariates. Thus, Hansen and Bowers (2008) urge the use of an omnibus test – a test that assessess balance across all linear combinations of the covariates in the table. Yet, the variable-by-variable display is useful in the same way that graphs such as Figure 32.2 are useful – in suggesting (not proving) the sources of imbalance.

Post-Stratification Enables Adjustment but Must Respect Blocking and Design

Say, for example, that within blocks defined by town, the treated group on average contained too many men (and that although gender was important in the study, the researcher either could not or forgot to block on it within existing blocks). An obvious method of preventing "male" from unduly confusing estimates of treatment effects is to only compare men to men, within block. Analysis then proceeds using the new set of blocks (which represent both the pre-treatment blocks and the new post-treatment strata within them) as before.

One may also use modern algorithmic matching techniques to construct strata. Keele, McConnaughy, and White (2008) argue in favor of matching over linear models for covariance adjustment and show simulations suggesting that such post-stratification can increase precision. Notice that matching to adjust experiments is different from matching in observational studies: matching here must be done without replacement in order to respect the assignment process of the experiment itself, and matching must be full. That is, although common practice in matching in observational studies is to exclude certain observations as unmatchable or perhaps to reuse certain excellent control units, in a randomized experiment every observation must be retained and matched only once. This limits the precision enhancing features of matching (at least in theory) because homogeneity will be bounded first by the blocking structure before random assignment and then again by requiring that all observations be matched.

Figure 32.3 shows confidence intervals resulting from a variety of post-stratification adjustments made to the thirty-two-city turnout data. In this particular experiment, the within-set homogeneity increase resulting from post-stratification did not outweigh the decrease in degrees of freedom occurring from the need to account for strata: the shortest confidence interval was for the unadjusted data (shown at the bottom of the plot).

Did the post-stratification help with the balance problems with the census variables evident in Table 32.1? Table 32.3 shows the strata-adjusted mean differences and p value for balance tests now adjusting for the post-stratification in addition to the blocking. Balance on baseline turnout and number of candidates does improve somewhat with the matching, but balance on percent black and median household income does not appreciably improve. Notice a benefit of post-stratification here: the postadjustment balance test shows us that we have two covariates that we could not balance.

Discussion of the advantages of blocking in Section 2 is, in essence, about how to analyze blocked (pre- or poststratified) experimental data. The rest of the chapter is devoted to understanding what it is that we mean by "covariance adjusted treatment effects."

Linear Models Enable Adjustment but May Mislead the Unwary

> *Even with carefully designed experiments there may be a need in the analysis to make some adjustment for bias. In some situations where randomization has been used, there may be some suggestion from that data that either by accident effective balance of important features has not been achieved or that possibly the implementation of the randomization has been ineffective. (Cox and Reid 2000, 29)*

Cox and Reid's discussion in their section 2.3 entitled "Retrospective Adjustment for Bias" echoes Cox (1958, 51–52) and Fisher (1925). What they call "bias," I think might more accurately be called "random imbalance."

Although Fisher developed the analysis of covariance using an asymptotic F test that approximated the randomized-based results, others have since noted that the standard sampling-based infinite population or normal model theory of linear models does not justify their use in randomized experiments. For example, in discussing regression standard errors, Cox and McCullagh (1982) note, "It

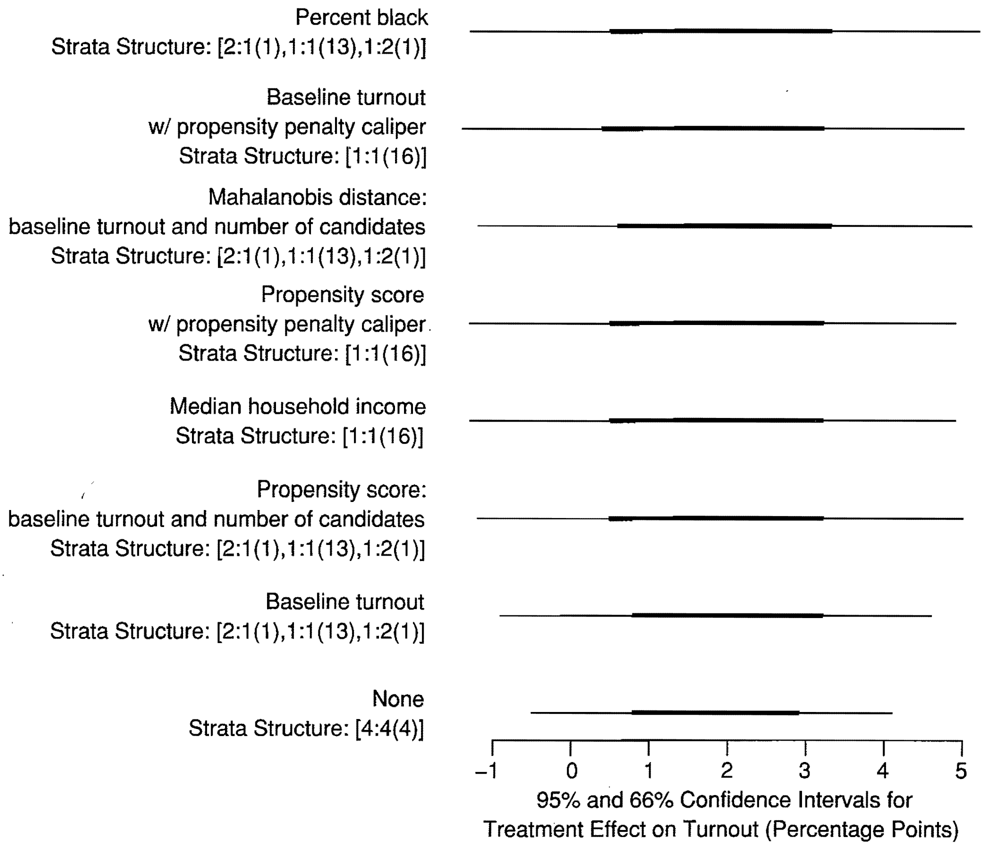

Figure 32.3. Post-Stratification Adjusted Confidence Intervals for the Difference in Turnout between Treated and Control Cities in the Thirty-Two-City Turnout Experiment

Notes: Each line represents the interval for the treatment effect after different post-stratification adjustments have been applied within the existing four blocks of eight cities. Thin lines show the ninety-five percent intervals. Thick lines show the sixty-six percent intervals. The propensity score was calculated from a logistic regression of treatment assignment on baseline turnout, number of candidates running for office, percent black, and median household income (both from the 2000 Census), and block indicators. All post-stratification and interval estimation was done with full, optimal matching using the `optmatch` (Hansen and Fredrickson 2010) and `RItools` (Bowers et al. 2009) packages for R. Numbers below the post-stratification label show the structure of the stratification: for example, without any post-stratification the experiment had four blocks, each with four treated and four control cities. The matching on absolute distance on the propensity score with a propensity caliper penalty produced sixteen pairs of treated and control cities (1:1(16)). The match on absolute distance on baseline turnout produced one set with two treated and one control (2:1(1)), thirteen paired sets (1:1(13)), and one set with one treated and two controls (1:2(1)). Because no observation could be excluded, calipers implied penalties for the matching algorithm rather than excluding observations from the matching (Rosenbaum 2010, ch. 8).

is known... that 'exact' second-order properties of analysis of covariance for precision improvement do not follow from randomization theory" (547). In this section, I provide an overview of a randomization-based method for covariance adjustment that can use normal approximations in the same way as those used for balance assessment and placebo tests. This method avoids most of the criticisms by Freedman (2008a, 2008b, 2008c), and thus suggests itself as useful in circumstances where the linear model as an approximation may cause concern or perhaps as a check on such approximations.

Table 32.3: One-by-One Balance Tests Adjusted for Covariates, by Post-Stratification in the Blocked Thirty-Two-City Study

			Post Hoc Full Matching on:			
	Blocks		*Baseline Turnout*		*Propensity Score*	
x	d_{strata}	p	d_{strata}	p	d_{strata}	p
Baseline outcome	−1.2	.2	−1.2	.3	−1.2	.3
Percent black	−14.7	.0	−14.7	.0	−14.7	.0
Median HH income	15.7	.0	15.8	.0	15.7	.0
Number of candidates	−2.9	.1	−2.6	.3	−2.9	.3

Notes: Two-sided p values assess evidence against the null of no effects. An omnibus balance test casts doubt on the null hypothesis of balance on linear combinations of these covariates adjusted for blocks, and two kinds of post hoc full matching within blocks at $p = .00003$, .003, and .004. Strata adjusted mean differences (d_{strata}) are generalizations of Equation (1) developed in Hansen and Bowers (2008) and implemented in the `RItools` package for R (Bowers et al. 2009). Statistical inference (p values) is randomization based, but uses large-sample normal approximations for convenience. Post hoc stratification results from optimal, full matching (Hansen 2004) on either absolute distance on baseline turnout or absolute distance on a propensity score with propensity caliper penalty shown in Figure 32.3.

First, though, let us get clear on what it means to "adjust for" random imbalance on a covariate.

What Does It Mean to Say That an Estimate has been "Adjusted" for the "Covariance" of Other Variables?

Let us look first at how covariance adjustment might work in the absence of blocking by looking only at the first block of eight units in the thirty-two-city dataset. Figure 32.4 is inspired by similar figures in Cox (1958, ch. 4) with dark gray showing treated units and black showing control units. The unadjusted difference of means is 6.6 (the vertical distance between the open squares that are not vertically aligned on the gray vertical line). The thick diagonal lines are the predictions from a linear regression of the outcome on an indicator of treatment and baseline outcome. The adjusted difference of means is the vertical difference between the regression lines, here, five. If there had been no relationship between baseline outcomes and post-treatment outcomes, the regression lines would have been flat and the vertical distances between those lines would have been the same as the unadjusted difference of means (the thin dark gray and black horizontal dashed lines). As ought to be clear here, parallel effects is a required assumption for covariance adjustment to be meaningful. In this case, a regression allowing different slopes and intercepts between the treatment groups shows the treatment slope of 0.25 and the control group slope of 0.23; thus, the assumption is warranted.

What about with blocked data? Figure 32.5 shows the covariance adjustment for three different covariates: 1) baseline outcomes (on the left), 2) percent black (in the middle), and 3) median household income (in \$1,000s, on the right). In each case, the data are *aligned* within each block by subtracting the block mean from the observed outcome (i.e., block centered). A linear regression of the block mean centered outcome on the block mean centered covariate plus the treatment indicator is equivalent to a linear regression of the observed outcome on the covariate plus treatment indicator plus indicator variables recording membership in blocks (i.e., "fixed effects" for blocks).

In the leftmost plot, we see that adjustment for baseline outcomes in addition to the blocking structure does little to change the treatment effect. In fact, these blocks were chosen with reference to baseline outcomes; thus, adjusting for blocks is roughly

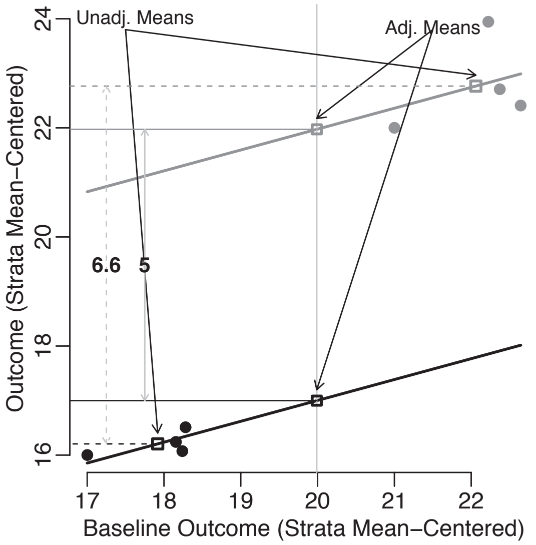

Figure 32.4. Covariance Adjustment in a Simple Random Experiment
Notes: Dark gray and black circles show treated and control units, respectively. The unadjusted difference of means is 6.6. The thick diagonal lines are $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 y_{i,t-1}$ with $Z_i = \{0, 1\}$ and the adjusted average treatment effect is $(\hat{Y}_i | Z_i = 1, y_{i,t-1} = \bar{y}_{t_1}) - (\hat{Y}_i | Z_i = 0, y_{i,t-1} = \bar{y}_{t_1}) = 5$.

equivalent to adjusting for baseline outcomes (only it does not require what is clearly a dubious parallel lines assumption). If the parallel lines assumption held in this case, however, then we might talk meaningfully about an adjusted effect. The average treatment effect in the first block is 6.6, but the treated units were more likely to participate, on average, than the control units even at baseline (a mean difference of 4.1). Some of the 6.6 points of turnout may well be due to baseline differences (no more than 4.1, we assume, and probably less so, because it would only matter to the extent that baseline turnout is also related by chance in a given sample to treatment assignment). In this case, the block-specific relationship between baseline outcomes and treatment assignment is vanishingly small (difference of means is 1.9), so only about two points of the average treatment effect is due to the baseline treatment effect (where "due to" is in a very specific linear smoothed conditional means sense). The adjusted effect is actually closer to 5 than 4.6 because the intuitions provided here with differences of means are not identical to what is happening with an analysis of covariance (although they are close and provide helpful intuition in this case).

The middle and right-hand plots show two instances in which the intuitions using differences of means become more difficult to believe. In strata 1, 2, and 4, every control unit has a higher percent black than every treated unit. The unadjusted average treatment effect is 1.8, but after adjustment for percent black the effect is 1.2. The middle plot, however, shows that the assumption of parallel effects is hard to sustain and that there is little overlap in the distributions of percent black between the treatment and control groups. In this case, however, the adjustment makes little difference in the qualitative interpretation of the treatment effect.

The right-hand figure is a more extreme case of the middle figure. This time there is no overlap at all between the distributions of median income between the controls (in black, and all on the left side of the plot) and the treated units (in dark gray, and all on the right side of the plot). The adjustment causes the treatment effect to change sign: from 1.8 to −2.1 percentage points of turnout. Is −2.1 a better estimate of the treatment effect than 1.8? Clearly, median income has a strong relationship with outcomes and also, via random imbalance, with treatment assignment (recall the test casting doubt on the null of balance in Table 32.1).

What is the problem with covariance adjustment in this way? First, as noted previously, the assumption of parallel lines is not correct. Second, we begin to notice another problem not mentioned in textbooks such as Cox and Reid (2000) or Cox (1958) – random assignment will, in large samples, ensure balance in the distributions of covariates but will not ensure such balance in small samples. This means that the distributions of the covariates, in theory, ought to be quite similar between the two groups. However, the theory does not exclude the possibility of random imbalance on one of many covariates, and it is well known that random imbalance can and does appear in practice. Adjustments for such imbalance can be done in such a

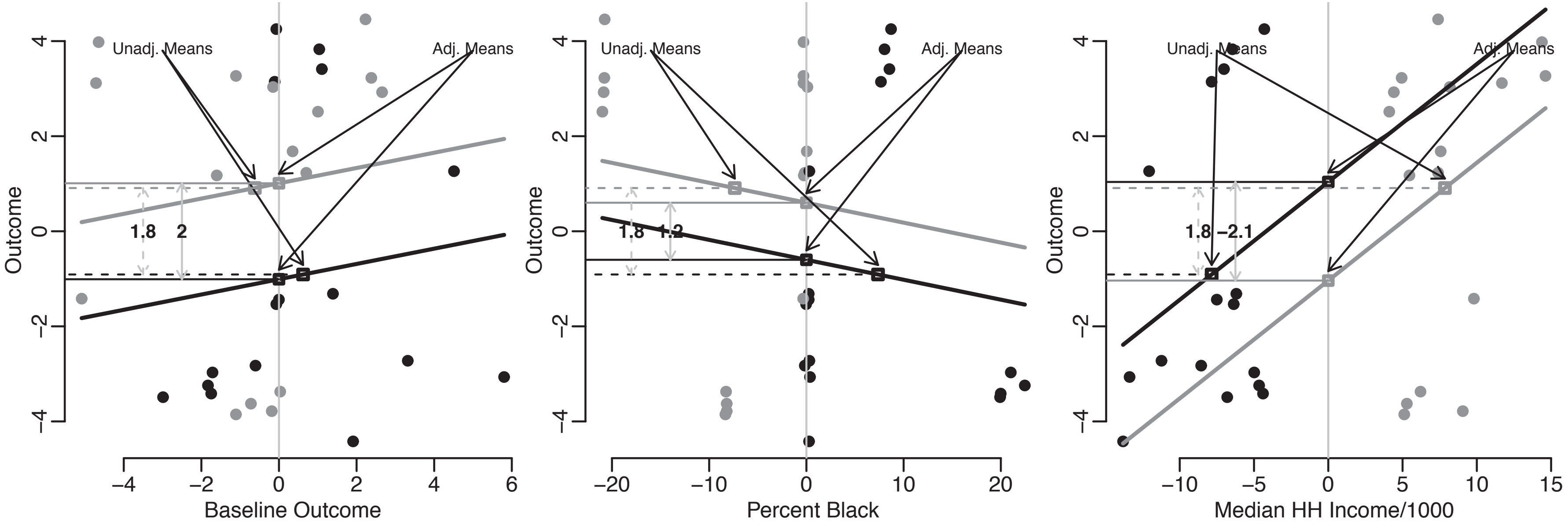

Figure 32.5. Covariance Adjustment in a Blocked Random Experiment
Notes: Dark gray and black circles show treated and control units, respectively. All variables are block mean centered.

way that adjusted mean differences are still meaningful representations of the treatment effect (as shown by the adjustment for baseline outcomes in the left plot of Figure 32.5). But, as the distributions of covariates become more unbalanced, the covariance adjustment can mislead. It is hard to claim, based on these data, that adjusting for median household income is a meaningful operation – one just cannot imagine (in these data) finding groups of treated and control units with the same median household income.[7]

Thus, we can see where some criticisms of covariance adjustment might easily come from: covariance adjustment done with multiple regression without additional diagnostics poses a real problem for diagnosing whether the imbalance is so severe as to provide no "common support" in the distributions of the covariates. In such cases, one really cannot "adjust for" the imbalance and must be resigned to the fact that treatment versus control comparisons, in the data, reflect something other than treatment assignment even if they would not in a larger sample or across repeated experiments. Of course, as sample size grows, given a finite set of covariates and a treatment with finite variance (i.e., a treatment that only has a few values and does not gain values as sample size grows), we would expect the problem of common support to disappear in randomized studies. Luckily, in many studies, one can assess such problems before treatment is administered.

Randomization Alone Can Justify Statistical Inference Covariance-Adjusted Quantities

A predominant use for covariance adjustment is not to ameliorate random imbalance but to enhance statistical precision. To the extent that a covariate predicts outcomes, one may use it to reduce the noise in the outcome unrelated to treatment assignment and thus help make treatment effects manifest. Covariance adjustment (whether for precision or for random imbalance) means linear regression. In theory, counterfactual statistical inference using the linear regression model for covariance adjustment estimator is biased (Freedman 2008a, 2008b, 2008c); however, in practice, it is often an excellent approximation (Green 2009; Schochet 2009). What should we do when we worry about the approximation: for example, when the experiment is small, there is great heterogeneity in effects and/or variance of effects across blocks, or great heterogeneity or discreteness in the outcome (such that the central limit theorem takes longer to kick in than one would prefer)? Rosenbaum (2002a) presents a simple argument that builds on the basics of Fisher's randomization-based inference. Here, I provide some brief intuition to guide study of that paper. This method of randomization-justified covariance adjustment does not rely on the linear model for statistical inference, but does "adjust" using the linear model.

Say an outcome is measured with noise caused in part by covariates. When we randomly assign treatment, we are attempting to isolate the part of the variation in the outcome due to the treatment from that due to other factors. Say we are still interested in the difference in mean outcomes between treatment and control groups as assigned. The standard deviations of those means may be large (making the treatment effect hard to detect) or small (making the treatment effect more easily manifest). If part of the noise in the outcome is due to covariates, then the residual from regressing the outcome on the covariates represents a less noisy version of the outcome – the outcome without noise from linear relationships with covariates. This residual e_{ib} (for unit i in block b) is measured in units of the outcome (i.e., "percent turning out to vote" in our running fake data example). The potential outcomes to treatment and control for units i in blocks b, y_{Tib} and y_{Cib}, are fixed, and Y_{ib} is random by virtue of its relationship with random assignment Z because $Y_{ib} = Z_{ib}y_{Tib} + (1 - Z_{ib})y_{Cib}$. A null hypothesis tells us what function of Y_i and Z_i would recover y_{Ci}: that is, if the null hypothesis is correct, then removing the effect (say, τ_{ib}) from the treated units $Y_{ib, Z=1}$ would tell us how the treated units would behave under control. Under a constant, additive model of effects, $y_{Tib} = y_{Cib} + \tau$ and so $Y_{ib} - Z_{ib}\tau_{ib} = y_{Cib}$[8]. Thus, considering our null hypothesis for the sake of argument, $H_0 : \tau_0 = \tau_{ib}$, regressing $(Y_{ib} - Z_{ib}\tau_{ib})$ on x_i is regressing a fixed quantity (i.e., y_{Cib}) on another fixed quantity, x_{ib}, and so the residuals from that regression are a fixed quantity.[9] One may substitute e for x in Equations (1) and (4). Fisher's style of inference begins with a test of a null hypothesis and inverts the hypothesis for a confidence interval: thus, the method allows us to infer consistent estimates of the causal effect by testing a sequence of causal hypotheses τ_0. Loosely speaking, the point estimate is the causal effect hypothesized by the best-supported hypothesis tested.[10] Note that this is a method of hypothesis testing, not of estimation. It would be quite incorrect to interpret the difference of means of residuals as an estimate of a treatment effect because the residuals already have specific causal hypotheses built into them as just described.

---

[7] This problem also occurred in Figure 32.4, but was not mentioned in order not to detract from the pedagogical task of describing the mechanism of covariance adjustment.

Figure 32.6 shows that the Rosenbaum style covariance adjustment in these data is well approximated by the direct regression–style covariance adjustment in the unadjusted case or when the adjustment is made for baseline turnout and number of candidates – and the version adjusted for baseline turnout and number of candidates (just) excludes zero from its ninety-five percent confidence interval. The two approaches differ when the difference of means is adjusted for the census variables. The most notable difference here is for median household income, where the direct adjustment method is based entirely on the linear extrapolation between the groups, whereas the Rosenbaum approach correctly captures the sense in which there is no reasonable way to adjust the treatment effect for this covariate. Because adjustment for percent black also requires much linear extrapolation, the randomization-based confidence interval again reflects the fact that the design itself has little information about what it means to adjust for this variable.

The advantages of this style of covariance adjustment are 1) that it sidesteps Freedman's critiques of covariance adjustment for experiments;[11] 2) although we used large-sample normal approximations to evaluate

---

[8] If this model is incorrect, then randomization-based inferences will be conservative, but the coverage of the confidence intervals will still be correct as noted independently by Gadbury (2001) and Robins (2002, Section 2.1). Other substantively meaningful models of effects are available (Rosenbaum 2002a, section 6; 2002b, ch. 5; 2010, ch. 2). For example, as Rosenbaum (2002c, 323) notes, if the treatment effect varies by a binary covariate x coding 0 for group 1 and 1 for group 2 (such that the parallel lines assumption is incorrect), then we would specify the potential responses to control as $Y_{ib} - Z_{ib}(\tau_{x=1}x_{ib} + \tau_{x=0}(1 - x_{ib}))$ for treatment effects that differ by group. I use the constant additive effects model in this chapter to map most closely onto the causal quantities implied by the choice of a linear regression model for covariance adjustment: indeed, for this very reason, I show how both styles of covariance adjustment can produce identical quantities in Figure 32.6. Interested readers might find the discussion in Rosenbaum (2002c, section 3–6) useful for thinking about the equivalence of estimating an average treatment effect and testing a sequence of hypotheses about individual causal effects.

[9] For a single covariate x and a regression fit $(Y_{ib} - Z_{ib}T_{ib}) = \hat{\beta}_0 + \hat{\beta}_1 x_{ib}, e_{ib} = (Y_{ib} - Z_{ib}\tau_{ib}) - (\hat{\beta}_0 + \hat{\beta}_1 X_{ib})$. The residual is written e, not $\hat{e}$, because the regression fit is not an estimate of an unknown quantity, but merely calculating a function of fixed features of the existing data.

[10] See discussion of the Hodges-Lehmann point estimate in Rosenbaum (2002a; 2002b, ch. 2) for more formal discussion of randomization-justified point estimates of causal effects. In the context of a large, cluster randomized field experiment with binary outcomes and nonrandom noncompliance, Hansen and Bowers (2009) show how one may approximate the results of testing sequences of hypotheses with simple calculations of means and associated randomization-based variances, including a method for randomization-based covariance adjustment. Because the Hansen and Bowers method approximates the results of testing sequences of hypotheses with simple means and variances, their method requires an asymptotic justification. Their article and related reproduction materials (Bowers, Hansen, and Fredrickson 2008) also provide tools for assessing the asymptotic justification.

[11] In particular, Freedman (2008b, 189) notes that the Fisher-style covariance adjustment is valid. "If $T_i = C_i$ for all i (the "strict null hypothesis"), then $\beta \equiv 0$ and adjustment will help – unless $\overline{\alpha Z} = 0$, i.e., the remaining variation (in C_i) is orthogonal to the covariate." Another method, elaborated in Hansen and Bowers (2009), also does not directly model the relationship between treatment and outcomes and so similarly avoids this critique.

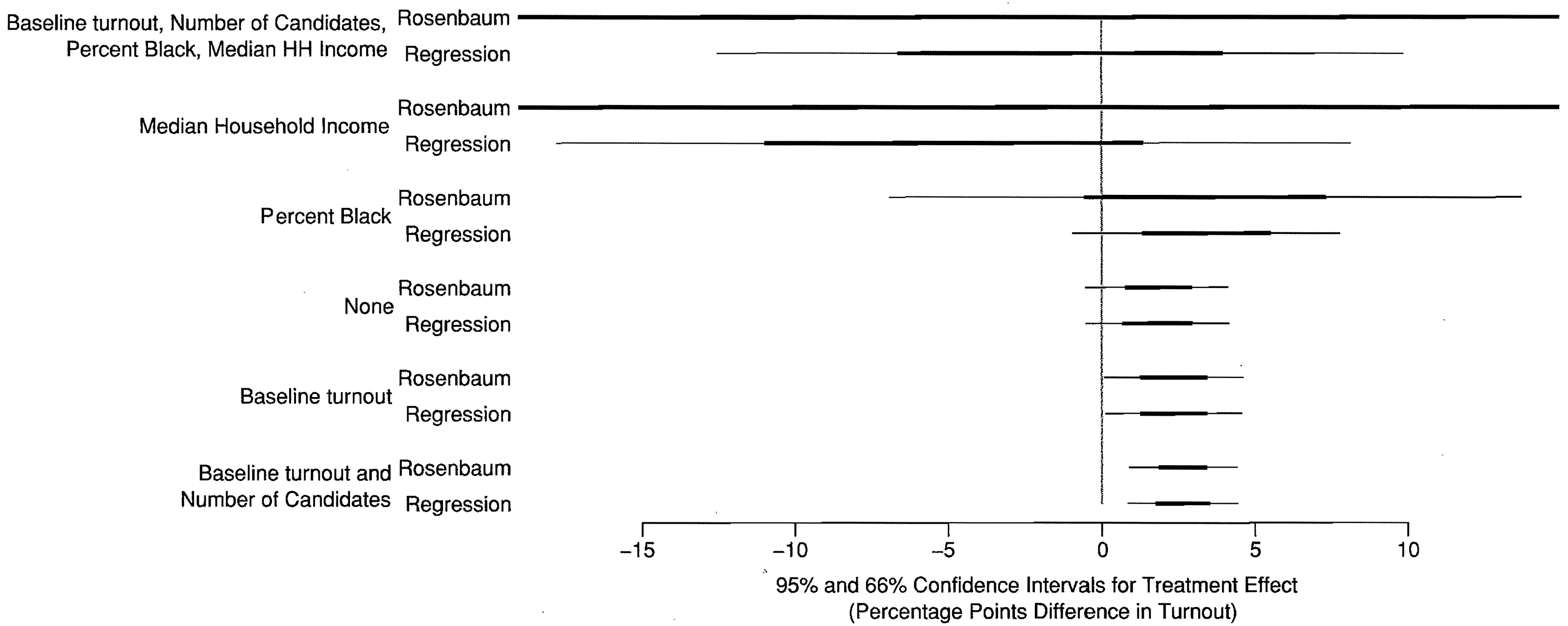

Figure 32.6. Covariance Adjusted Confidence Intervals for the Difference in Turnout between Treated and Control Cities in the Thirty-Two-City Turnout Experiment Data

Notes: "Regression"-style adjustment regressed turnout on treatment indicator, block indicators, and covariates and referred to the standard iid + *t* based reference distribution for confidence intervals. "Rosenbaum"-style adjustment regressed turnout on block indicators and covariates, and then used the residuals as the basis for tests of the null hypotheses implied by these confidence intervals. The reference distributions for the Rosenbaum-style are large-sample approximations to the randomization distribution implied by the design of the experiment using the `RItools` (Bowers et al. 2009) package for R. Thin lines show the ninety-five percent intervals. Thick lines show the sixty-six percent intervals. The randomization-based confidence intervals for outcomes after adjustment by functions including median household income are essentially infinite.

the differences of means here, neither differences of means nor large-sample normal approximations are necessary (and the large-sample approximations are checkable)[12]; and 3) unmodeled heteroskedasticity or incorrect functional form does not invalidate the statistical inference based on this method as it does for the standard approach. For example, the parallel lines assumption is no longer relevant for this approach because the only job of the linear model is to reduce the variance in the outcome. Finally, this particular data example allows us to notice another side benefit of the statistical property of correct coverage: when there is no (or little) information available in the design, the randomization-based approach will not overstate the certainty of conclusions in the same way as the model-based approach.[13]

Best Practices for Regression-Based Adjustment

Adjusted treatment effect estimates always invite suspicion of data snooping or modeling artifacts. None of the techniques discussed here entirely prevents such criticism. Of course, the easy way to avoid such criticism is to announce in advance what kinds of random imbalance are most worrisome and determine a plan for adjustment (including a plan for assessing the assumptions of the adjustment method chosen). Covariance adjustment using the standard linear regression model requires that one believe the assumptions of that model. For example, this model as implemented in most statistical software requires a correct model of the relationship between outcomes and covariates among treatment and control units (i.e., a correct functional form), that the heteroskedasticity induced by the experimental manipulation is slight, and that the sample size is large enough to overcome the problems highlighted by Freedman (2008a, 2008b, 2008c). As with any use of linear regression, one may assess many of these assumptions. If one or more of these assumptions appear tenuous, however, then this chapter shows that one may still use the linear model for adjustment, but do so in a way that avoids the need to make such commitments. Readers interested in the Rosenbaum (2002a) style of covariance adjustment should closely study that paper. The code contained in the reproduction archive for this chapter may also help advance understanding of that method.

5. The More You Know, the More You Know

A randomized study that allows "the phenomenon under test to manifest itself" provides particularly clear information and thus enhances theory assessment, theory creation, and policy implementation. Thus, researchers should attend to those elements of the design and analysis that would increase the precision of their results. This chapter points to only a small part of the enormous body of methodological work on the design of experiments.

Random assignment has three main scientific aims: 1) it is designed to balance distributions of covariates (observed and unobserved) such that, across repeated randomizations, assignment and covariates should be independent; 2) it is designed to allow assessment of the uncertainties of estimated treatment effects without requiring populations or models of outcomes (or linearity assumptions); and 3) it is a method of manipulating putatively causal variables in a way that is impersonal and thus enhances the credibility of causal claims. Political scientists have recently become excited about experiments primarily for the first and third aims, but they have ignored the second aim. This chapter discusses some of the benefits (and pitfalls) of the use of covariates in randomized experiments while maintaining a focus on randomization as the basis for inference.

---

[12] Rosenbaum (2002a) focuses on normal approximations to the Wilcox rank sum statistic as his preferred summary of treatment effects (and the normal approximations there are not necessary either, but merely convenient and often correct in large enough samples with continuous outcomes).

[13] The disadvantages of this mode are not on display here (although they will be obvious to those who use the code contained in this chapter for their own work). First, remember that this approach produces confidence intervals by testing sequences of hypotheses. It does not "estimate" causal effects as would a standard regression estimator, but rather assesses the plausibility of causal effects using tests. Of course, such assessments of plausibility are also implicit in confidence intervals for standard regression estimators. However, the mechanics of the two methods of covariance adjustment are quite different. The randomization-based approach as implemented here builds a confidence interval by direct inversion of hypothesis tests: in this case, we tested hypotheses about the treatment effect from $\tau_0 = -20$ to $\tau_0 = 20$ by 0.1. This can be computationally burdensome if the number of hypotheses to test is large or if we eschew large-sample approximations. Second, we did not work to justify our choice of mean difference (rather than rank or other summary of observed outcomes and treatment assignment). The standard linear regression estimator requires attention to mean differences as the quantity of interest, whereas any test statistic may be used in the randomization-based method of adjustment shown here.

Although randomization allows statistical inference in experiments to match the causal inference, covariate imbalance can and does occur in experiments. Balance tests are designed to detect worrisome imbalances. One ought to worry about random imbalances when they are 1) large enough (and relevant enough to the outcome) that they should make large changes in estimates of treatment effects, and 2) large relative to their baseline values such that interpretation of the treatment effect could be confused.

Small studies provide little information to help detect either treatment effects or imbalance. The null randomization distribution for a balance test in a small study ought to have larger variance than said distribution in a large study. The same observed imbalance will cast more doubt on the null of balance in a large study than it will in a small study: the observed value will be farther into the tail of the distribution characterizing the hypothesis for the large study than it will in the small study. The same relationship between small and large studies holds when the test focuses on the treatment effect itself. Thus, a p value larger than some acceptance threshold for the null hypothesis of balance tells us that the imbalance is not big enough to cause detectable changes in assessments of treatment effects. A p value smaller than some acceptance threshold tells us that the imbalance is big enough to cause detectable changes when we gauge the effects of treatment.

Given random imbalance, what should one do? Adjustment can help, but adjustment can also hurt. This chapter showed (using a fake dataset built to follow a real dataset) a case in which adjustment can help and seems meaningful and a case in which adjustment does not seem meaningful, as well as an intermediate case. One point to take away from these demonstrations is that some imbalance can be so severe that real adjustment is impossible. Just as is the case in observational studies, merely using a linear model without inspecting the data can easily lead an experimenter to mislead him- or herself – and problems could multiply when more than one covariate is adjusted for at a time. Rosenbaum's (2002a) proposal for a randomization-based use of linear regression models is attractive in that covariance-adjusted confidence intervals for the treatment effect do not depend on a correct functional form for the regression model. In this paper, all of the adjustment for median household income depended on a functional form assumption, so the randomization-based confidence interval was essentially infinite (signaling that the design of the study had no information available for such adjustment) while the model-based regression confidence interval, although much wider than the unadjusted interval, was bounded. Modern matching techniques may also help with this problem (Keele et al. 2008). In this chapter, precision was not enhanced by matching within blocks, but matchings including median household income did not radically change confidence intervals for the treatment effect, and balance tests before and after matching readily indicated that the matchings did not balance median household income.

This chapter did not engage with some of the other circumstances in which covariate information is important for randomized studies. In particular, if outcomes are missing, then prognostic covariates become ever more important in experimental studies given their ability to help analysts build models of missingness and models of outcomes (Barnard et al. 2003; Horiuchi, Imai, and Taniguchi 2007). Thus, I understate the value of collecting more information about one's units. The trade-offs between collecting more information about units versus including more units in a study ought to be understood from the perspectives long articulated in the many textbooks on experimental design: simple random assignment of units to two treatments (treatment vs. control) can be a particularly inefficient research design.

References

Barnard, John, Constantino E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. 2003. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association* 98 (462): 299–324.

Bowers, Jake, Mark Fredrickson, and Ben Hansen. 2009. *RItools: Randomization Inference Tools*. R package version 0.1-6 ed.

Bowers, Jake, Ben B. Hansen, and Mark Fredrickson. 2008. Reproduction achive for: 'Attributing Effects to a Cluster Randomized Get-Out-the-Vote Campaign'. Technical Report, University of Illinois and University of Michigan.

Bowers, Jake, and Costas Panagopoulos. 2009. '"Probability of What?': A Randomization-Based Method for Hypothesis Tests and Confidence Intervals about Treatment Effects." Unpublished paper, University of Illinois, Urbana-Champaign. Retrieved January 24, 2011 from http://www.jakebowers.org/PAPERS/BowPanEgap2009.pdf.

Brady, Henry E. 2008. "Causation and Explanation in Social Science." In *Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press, chap. 10, 217–70.

Cochran, William G., and Gertrude M. Cox. 1957. *Experimental Designs*. New York: Wiley.

Cox, David R., and Peter McCullagh. 1982. "Some Aspects of Analysis of Covariance (with Discussion)." *Biometrics* 38: 541–61.

Cox, David R., and Nancy Reid. 2000. *The Theory of the Design of Experiments*. Boca Raton, FL: Chapman & Hall/CRC Press.

Cox, David R. 1958. *The Planning of Experiments*. New York: Wiley.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Freedman, David A. 2008a. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics* 2: 176–96.

Freedman, David A. 2008b. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40: 180–93.

Freedman, David A. 2008c. "Randomization Does Not Justify Logistic Regression." *Statistical Science* 23: 237–49.

Gadbury, Gary L. 2001. "Randomization Inference and Bias of Standard Errors." *American Statistician* 55: 310–13.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multi-Level/Hierarchical Models*. New York: Cambridge University Press.

Green, Donald P. 2009. "Regression Adjustments to Experimental Data: Do David Freedman's Concerns Apply to Political Science?" Unpublished paper, Yale University.

Hansen, Ben, and Mark Fredrickson. 2010. *Optmatch: Functions for Optimal Matching, Including Full Matching*. R package version 0.6-1 ed.

Hansen, Ben B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99 (467): 609–18.

Hansen, Ben B. 2008. "Comment: The Essential Role of Balance Tests in Propensity-Matched Observational Studies." *Statistics in Medicine* 27: 2050–4.

Hansen, Ben B., and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23: 219–36.

Hansen, Ben B., and Jake Bowers. 2009. "Attributing Effects to a Cluster Randomized Get-Out-the-Vote Campaign." *Journal of the American Statistical Association* 104: 873–85.

Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. "Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment." *American Journal of Political Science* 51: 669–87.

Imai, Kosuke, Gary King, and Clayton Nall. 2009. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." *Statistical Science* 24: 29–72.

Imai, Kosuke, Gary King, and Elizabeth Stuart. 2008. "Misunderstandings among

Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 111: 1–22.

Keele, Luke J., Corrine McConnaughy, and Ismail K. White. 2008. "Adjusting Experimental Data." Unpublished Manuscript, The Ohio State University.

Leisch, Friedrich. 2002. "Dynamic Generation of Statistical Reports Using Literate Data Analysis." In *Compstat 2002 – Proceedings in Computational Statistics*, eds. Wolfgang Haerdle and Bernd Roenz. Heidelberg, Germany: Physika Verlag.

Leisch, Friedrich. 2005. *Sweave User Manual.* Retrieved from http://www.stat.uni-muenchen.de/~leisch/Sweave/Sweave-manual.pdf.

Neyman, Jerzy. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 (1923)." *Statistical Science* 5: 463–80.

Nickerson, David W. 2005. "Scalable Protocols Offer Efficient Design for Field Experiments." *Political Analysis* 13: 233–52.

Panagopoulos, Costas. 2006. "The Impact of Newspaper Advertising on Voter Turnout: Evidence from a Field Experiment." Unpublished Manuscript, Fordhan University.

Robins, James M. 2002. "Covariance Adjustment in Randomized Experiments and Observational Studies: Comment." *Statistical Science* 17: 309–21.

Rosenbaum, Paul R. 2002a. "Covariance Adjustment in Randomized Experiments and Observational Studies." *Statistical Science* 17: 286–327.

Rosenbaum, Paul R. 2002b. *Observational Studies.* 2nd ed. New York: Springer-Verlag.

Rosenbaum, Paul R. 2002c. "Rejoinder." *Statistical Science* 17: 321–27.

Rosenbaum, Paul R. 2010. *Design of Observational Studies.* New York: Springer.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66: 688–701.

Rubin, Donald B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5: 472–80.

Schochet, Peter. 2009. "Is Regression Adjustment Supported by the Neyman Model for Causal Inference?" *Journal of Statistical Planning and Inference* 140: 246–59.

Sekhon, Jasjeet S. 2008. "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods." In *Oxford Handbook of Political Methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. New York: Oxford University Press, chap. 11, 271–99.

Senn, Stephen J. 1994. "Testing for Baseline Balance in Clinical Trials." *Statistics in Medicine* 13: 1715–26.

Cambridge Handbook of Experimental Political Science

Edited by

JAMES N. DRUCKMAN

Northwestern University

DONALD P. GREEN

Yale University

JAMES H. KUKLINSKI

University of Illinois at Urbana–Champaign

ARTHUR LUPIA

University of Michigan

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521174558

© Cambridge University Press 2011

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2011

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data

Cambridge Handbook of Experimental Political Science / [edited by] James N. Druckman,
Donald P. Green, James H. Kuklinski, Arthur Lupia.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-19212-5 (hardback) – ISBN 978-0-521-17455-8 (paperback)
1. Political science – Methodology. 2. Political science – Research.
3. Political science – Experiments. I. Druckman, James N., 1971– II. Title.
JA71.C325 2011
320.072 – dc22 2010044869

ISBN 978-0-521-19212-5 Hardback
ISBN 978-0-521-17455-8 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLS for external or third-party Internet Web sites referred to in this publication and does not guarantee that any content on such Web sites is, or will remain, accurate or appropriate.

Contents

List of Tables *page* viii
List of Figures ix
Contributors xi
Acknowledgments xiii

INTRODUCTION

1 Experimentation in Political Science 3
James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia

PART I: DESIGNING EXPERIMENTS

2 Experiments: An Introduction to Core Concepts 15
James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia

3 Internal and External Validity 27
Rose McDermott

4 Students as Experimental Participants: A Defense of the "Narrow Data Base" 41
James N. Druckman and Cindy D. Kam

5 Economics versus Psychology Experiments: Stylization, Incentives, and Deception 58
Eric S. Dickson

PART II: THE DEVELOPMENT OF EXPERIMENTS IN POLITICAL SCIENCE

6 Laboratory Experiments in Political Science 73
Shanto Iyengar

7 Experiments and Game Theory's Value to Political Science 89
John H. Aldrich and Arthur Lupia

8 The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation 102
Paul M. Sniderman

9 Field Experiments in Political Science 115
Alan S. Gerber

PART III: DECISION MAKING

10 Attitude Change Experiments in Political Science 141
Allyson L. Holbrook

11 Conscious and Unconscious Information Processing with Implications for Experimental Political Science 155
Milton Lodge, Charles Taber, and Brad Verhulst

12 Political Knowledge 171
Cheryl Boudreau and Arthur Lupia

PART IV: VOTE CHOICE, CANDIDATE EVALUATIONS, AND TURNOUT

13 Candidate Impressions and Evaluations 187
Kathleen M. McGraw

14 Media and Politics 201
Thomas E. Nelson, Sarah M. Bryner, and Dustin M. Carnahan

15 Candidate Advertisements 214
Shana Kushner Gadarian and Richard R. Lau

16 Voter Mobilization 228
Melissa R. Michelson and David W. Nickerson

PART V: INTERPERSONAL RELATIONS

17 Trust and Social Exchange 243
Rick K. Wilson and Catherine C. Eckel

18 An Experimental Approach to Citizen Deliberation 258
Christopher F. Karpowitz and Tali Mendelberg

19 Social Networks and Political Context 273
David W. Nickerson

PART VI: IDENTITY, ETHNICITY, AND POLITICS

20 Candidate Gender and Experimental Political Science 289
Kathleen Dolan and Kira Sanbonmatsu

21 Racial Identity and Experimental Methodology 299
Darren Davis

22 The Determinants and Political Consequences of Prejudice 306
Vincent L. Hutchings and Spencer Piston

23 Politics from the Perspective of Minority Populations 320
Dennis Chong and Jane Junn

PART VII: INSTITUTIONS AND BEHAVIOR

24 Experimental Contributions to Collective Action Theory 339
Eric Coleman and Elinor Ostrom

25 Legislative Voting and Cycling 353
Gary Miller

26 Electoral Systems and Strategic Voting (Laboratory Election Experiments) 369
Rebecca B. Morton and Kenneth C. Williams

27 Experimental Research on Democracy and Development 384
Ana L. De La O and Leonard Wantchekon

PART VIII: ELITE BARGAINING

28 Coalition Experiments 399
Daniel Diermeier

29 Negotiation and Mediation 413
Daniel Druckman

30 The Experiment and Foreign Policy Decision Making 430
Margaret G. Hermann and Binnur Ozkececi-Taner

PART IX: ADVANCED EXPERIMENTAL METHODS

31 Treatment Effects 445
Brian J. Gaines and James H. Kuklinski

32 Making Effects Manifest in Randomized Experiments 459
Jake Bowers

33 Design and Analysis of Experiments in Multilevel Populations 481
Betsy Sinclair

34 Analyzing the Downstream Effects of Randomized Experiments 494
Rachel Milstein Sondheimer

35 Mediation Analysis Is Harder than It Looks 508
John G. Bullock and Shang E. Ha

AFTERWORD

36 Campbell's Ghost 525
Donald R. Kinder

Name Index 531
Subject Index 548