18

Treatment effects in before-after data[1]

Andrew Gelman[2]

18.1 Default statistical models of treatment effects

The default analyses for experiments and observational studies assume constant treatment effects. The usual modeling or Bayesian approach with ignorable treatment assignment starts with a constant treatment effect; for example, $y_i = \beta_0 + \beta_1 T_i + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \epsilon_i$, where T_i is the treatment variable (most simply, an indicator that equals 1 for treated units and 0 for controls). In Fisher's classical test, the null hypothesis is that treatment effects are zero for all units. More generally, this approach can be inverted to obtain confidence intervals for a constant treatment effect. Neyman (1923) allowed the possibility for varying effects (see Rubin, 1990) but only as a goal toward estimating or testing hypotheses about average treatment effects.

Before-after designs have been much discussed in the statistical literature (see Brogan and Kutner, 1980, Laird, 1983, Crager, 1987, Stanek, 1988, Stein, 1989, Singer and Andrade, 1997, and Yang and Tisatis, 2001). It is recognized that treatment effects can vary with pre-treatment covariates ($x_2, x_3, \ldots$ in the above model), and that these interactions can be substantively important (see

[1]To appear in *Applied Bayesian Modeling and Causal Inference from an Incomplete Data Perspective*, ed. A. Gelman and X. L. Meng. London: Wiley.

[2]Department of Statistics, Columbia University, New York.

We thank Gary King, Iain Pardoe, Don Rubin, Hal Stern, and Alan Zaslavsky for helpful conversations and the National Science Foundation for financial support.

Dehejia, 2004). We argue here that interaction between treatment and covariates is a general phenomenon that can be seen as deriving from an underlying variance components model. We posit fundamental variation among experimental (or observational) units that is not fully captured in pre-treatment predictors and manifests itself in experimental or observational outcomes.

18.2 Before-after correlation is typically larger for controls than for treated units

Our point is not merely that treatment effects vary—in practice, everything varies—but that they vary in systematic, predictable ways. We begin by reviewing a ubiquitous pattern in experiments and observational studies with before-after data: the correlation between "before" and "after" measurements is commonly higher for controls than in the treatment group.

An observational study of legislative redistricting

Figure 18.1 gives an example from our research on the effects of redistricting on the partisan bias of electoral systems (Gelman and King, 1994). The symbols in the graph represent state legislatures in election years (e.g., California in 1974), with the estimated "partisan bias" (a measure of the fairness of the electoral system) of the legislature in that year plotted vs. the estimated partisan bias in the previous election. The small dots in the graph represent "control" cases in which there was no redistricting, and the larger symbols correspond to "treated" cases, or redistrictings. The treatment has three levels—corresponding to redistrictings controlled by Democrats, Republicans, or both parties—but here we consider all treatments together. Elections come every two years and redistricting typically happens every ten years, so most of the data points are controls. The correlation between before and after measurements is much larger for controls than treated cases. (The regression lines for the three levels of treatment are constrained to be parallel and equally spaced because there were not enough data points to accurately estimate separate slopes or separate effects for the two parties.)

From the usual standpoint of estimating treatment effects, the interaction between treatment and x (estimated partisan bias in previous election) in Figure 18.1 is dramatic—and, in fact, we had not thought to include an interaction in our model until it jumped out at us from the graph. Stepping back a bit, however, the different slopes for the two groups should be no surprise at all. In the control cases with no redistricting, the state legislature changes very little, and so the partisan bias will probably change very little from the previous election. In contrast, when the legislative districts are redrawn, larger and more unpredictable changes occur.

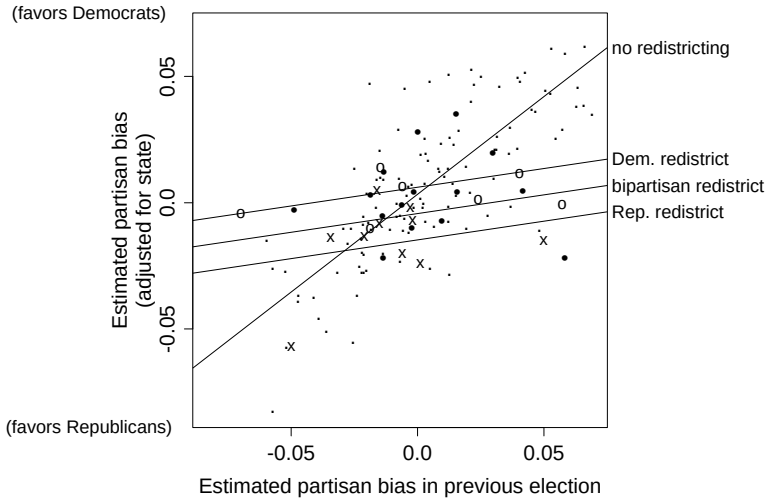

Figure 18.1: Effect of redistricting on partisan bias. Each symbol represents a state and election year, with dots indicating control cases (years with no redistricting) and the other symbols corresponding to different types of redistricting. As indicated by the fitted regression lines, the "before" value is much more predictive of the "after" value for the control cases than for the treated (redistricting) cases. In contrast to the minor differences between Democratic, bipartisan and Republican redistricting, the dominant effect of the treatment is to bring the expected value of partisan bias toward 0, and this effect would not be discovered with a model that assumed parallel regression lines for treated and control cases. From Gelman and King (1994).

In fact, in this example, the interaction effect of redistricting—that it tends to reduce partisan bias—is larger than the original object of this study, which was the partisan advantage of redistricting (the slight difference between the lines for Democratic, bipartisan, and Republican treatment lines in Figure 18.1). It was crucial to model the variation in the treatment effects to see this effect.

An experiment with pre-test and post-test data

Figure 18.2 summarizes before-after correlations from an educational experiment performed on a set of elementary-school classes.[3] In each of four grades,

[3]The treatment in this experiment was exposure to a new educational television show called "The Electric Company." The experiment was conducted around 1970 and used as

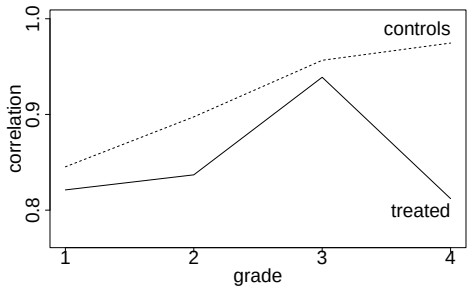

Figure 18.2: Correlation of pre-test and post-test scores for an educational experiment, for control and treated classrooms in each of four grades. Correlations are higher in the control groups, which is consistent with models of varying treatment effects.

the classes were randomized into treated and control groups, with pre-test and post-tests taken for each class. Figure 18.2 shows the correlation between before and after measurements, computed separately among the control and treated classes. At each grade level, the correlation is higher for the controls.

As in our previous example, the pattern of correlations makes sense: the pre-test is a particularly effective predictor of post-test scores for the control classes, where no intervention has been imposed (except for a year of schooling). In the treatment group, it is reasonable to expect the intervention to have different effects in different classrooms, thus attenuating the correlation of before and after measurements.

Congressional elections with incumbents and open seats

We give one more example of before-after correlations, in an observational study of the effect of incumbency in elections in the U.S. House of Representatives.[4] The units in this example are Congressional districts, the before and after measurements are the Democratic Party's share of the vote in two successive elections, and the "treatment" is incumbency. For simplicity, we separately analyze in each year the seats held by Democrats and by Republicans.

In the context of our discussion here, the "control" districts are those where the incumbents are running for reelection, and the "treated" districts are the open seats, where the incumbent party is running a new candidate. We use this labeling because the races with incumbents represent less change from the previous election, whereas running a new candidate can be viewed

an example in Don Rubin's course at Harvard University in 1985.

[4]See Gelman and King (1990) and Gelman and Huang (2004) for details.

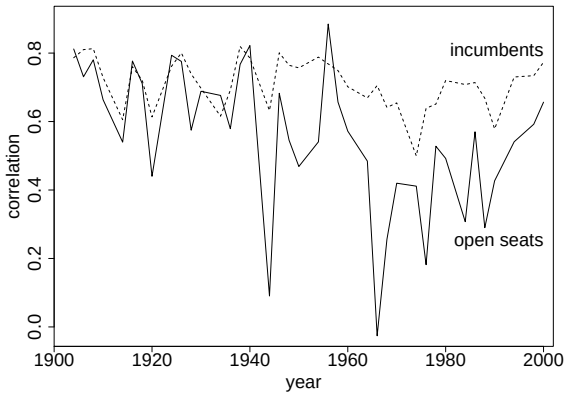

Figure 18.3: Correlations of party vote share in each pair of successive Congressional elections in the past century, computed separately for the incumbents running for reelection (the "control group") and open seats (the "treatment group"). Correlations are consistently higher in the control group, which makes sense since there is less change between before and after in these districts. In the early part of the century, when correlations in the two groups were about the same, the effect of incumbency was very small.

as an intervention. The effect of incumbency in a given district is then the negative of the treatment effect as defined here.

Figure 18.3 shows the correlations between the Democratic vote shares in each pair of two successive elections, computed separately for controls (incumbents running) and treated districts (open seats).[5] As in our previous examples, the before-after correlation is much higher in the control group. Again, this picture is consistent with the idea that there is little change among the controls, whereas a varying treatment effect reduces the predictive importance of past data.

A careful look at Figure 18.3 reveals that the before-after correlations within the two groups did not diverge until the second half of the century. A separate analysis (not shown here) estimates the average advantage of incumbency in Congressional elections to be near zero for the first half of the century, then increasing dramatically through the 1950s and 1960s to its current high level. Thus, as the treatment effect increased, its variation increased also. (The jaggedness of the solid line in Figure 18.3 can largely be explained

[5]We exclude uncontested elections and years ending in "2," when district lines are redrawn. Within each group (incumbents running and open seats), we compute correlations separately for the Democratic and Republican-held seats; Figure 18.3 presents the averages of the within-party correlations for each pair of election years.

as sampling variability given the small number of open seats, especially in recent decades.)

18.3 A class of models for varying treatment effects

When only "after" data are available in an experiment, it is not possible to see the consequences of varying treatment effects, and the classical t interval gives appropriate superpopulation inference for average treatment effects (see Gelman et al., 2003, Section 7.5). In contrast, treatment effects that vary as a function of "before" data can be modeled and estimated in a number of ways.

Plots such as Figure 18.1 suggest regression models with treatment effects interacted with pre-treatment covariates. We would like to think more generally of treatments that can have varying effects, both additive and subtractive. For example, suppose we label the "before" and "after" measurements for unit j as $y_{jt}, t = 0, 1$, and fit the two-error-term model,

$$\begin{aligned} \text{before:} \quad y_{j0} &= (X\beta)_{j0} + \alpha_j + \gamma_{j0} + \epsilon_{j0} \\ \text{after:} \quad y_{j1} &= T_j\theta + (X\beta)_{j1} + \alpha_j + \gamma_{j1} + \epsilon_{j1}, \end{aligned} \tag{18.1}$$

where T represents the indicator for treatment (which in this setup occurs between the "before" and "after" measurements) and θ is the average treatment effect—the usual object of inference in an observational study. The matrix X represents other linear predictors in the regression model (e.g., demographic variables for a model of individuals, or district-level characteristics for a model of election outcomes), and the unit-level term α_j represents persistent variation among units not explained by the predictors. The error terms $\epsilon_{j0}, \epsilon_{j1}$ are the usual independent observation-level errors.

The terms γ_{j0}, γ_{j1} take model (18.1) beyond the usual longitudinal or panel-data hierarchical regression framework, and our key innovation is in linking this variance component with the treatment, so that it is affected differently by the treatment and controls. Various models are possible here, all of which allow treatment effects to vary by unit and have the byproduct that before-after correlation is higher for controls than treated units. We list some possibilities here.

Replacement treatment error. Suppose that under the control condition, γ_j is unchanged (that is, $\gamma_{j1} \equiv \gamma_{j0}$), but under the treatment, γ_{j0} and γ_{j1} are independent draws from have the same probability distribution. In this model, the treatment has the effect of replacing a random error component. This could make sense if the control corresponded to staying with a particular regimen and the treatment corresponded to switching to a new approach. For

example, in the redistricting example in Figure 18.1, the treatment replaces an old districting plan with a new one.

Additive treatment error. Suppose that $\gamma_{j0} \equiv 0$ for all units, and $\gamma_{j1} = 0$ for controls but is drawn from a distribution for treated units. In this model, the treatment adds a source of variability that was not present before. This could happen if the treatment is a new, active intervention (for example, the educational TV program in Figure 18.2).

Subtractive treatment error. For a different model, suppose that γ_{j0} comes from some probability distribution, and under the control condition, $\gamma_{j1} \equiv \gamma_{j0}$), but under the treatment, $\gamma_{j1} \equiv 0$. In this model, the treatment subtracts a source of variability. This could apply to a setting in which an active intervention has already been applied to the "before" measurements, and the control and treatment conditions correspond to staying with or dropping the intervention. For example, in the incumbency example in Figure 18.3, the "treatment" corresponds to an open seat—the disappearance of an incumbent (see Gelman and Huang, 2004).

More formally using the potential-outcome notation of Rubin (1974), the error terms γ_{j1} could be written as γ_{Tj1}, where $T = 0$ or 1 corresponds to the control and treatment conditions. In any case, these models, or more general distributions on these error terms, capture the idea that the treatment *changes* the affected units as well as having some average additive effect. Similar models are used in animal breeding to model genetic variation and treatment effects (see Lynch and Walsh, 1988), and Sargent and Hodges (1997) present related ideas for hierarchical models of complex regression interactions. We would also like to formulate a class of models in which treatments with larger main effects naturally have larger variation, as this is another property that often seems to hold in practice.

18.4 Discussion

It has been argued that statistical models should be adapted individually to applied problems (see, for example, chapter 27 in this volume). However, in practice default procedures and models are used in a wide variety of settings. This is not merely for convenience (or because certain models are easier to access in statistical software packages such as SPSS) but because default models often work. Methods such as t-intervals, the analysis of variance, and least-squares regression have been effective in all sorts of problems (see, for example, Snedecor and Cochran, 1989), and much of the methodological research of the past few decades has resulted in extensions of these and other approaches. Our current toolbox of default methods includes t models for

robust regression and multivariate imputation (generalized from the normal; see Liu, 1995), wavelet decompositions (generalized from Fourier analysis; see chapter 31 in this volume), generalized linear models (McCullagh and Nelder, 1989), splines and locally weighted regressions (Wahba, 1979, and Cleveland, 1979), and model averaging for regressions and density estimates (Hoeting et al., 1999, and Richardson and Green, 1997). All these methods have been demonstrated for specific examples but are intended to be flexible generalizations of previous default approaches.

In this chapter, we have tried to motivate an expansion of the default model of experiments and observational studies to allow for treatment effects to vary among units. This variation can sometimes be expressed as interactions with pre-treatment measurements but more generally can be understood as effects on unobserved unit-level variance components of the sort that are used in instrumental variables and principal stratification (see chapters 8 and 9 in this volume). Our models are still under development and we hope they will reach "default" stage sometime in the not-so-distant future, as a small part of a general applied framework for causal inference deriving ultimately from the potential-outcome perspective of Rubin (1974).