Regression without Regrets: A modular approach to linear models in (quasi)-experiments.

Jake Bowers [*] Costas Panagopoulos [†]
Mark M. Fredrickson [‡]

April 6, 2013 (Version: c3b976a)

Abstract

The design of a randomized study guarantees not only clear and "interpretable comparisons"(Kinder and Palfrey, 1993, page 7) but valid statistical tests even in the absence of large samples or known data generating processes for outcomes (Fisher, 1935, Chap 2). Yet, while design alone yields valid tests the tests could lack power: a valid but wide confidence interval may be more useful than a misleadingly narrow confidence interval, but still shed little light on the theory motivating the study. After a brief demonstration of Fisher's statistical framework (to fix ideas about the validity of tests and to distinguish it from frameworks which estimate average treatment effects), we show a method by which a researcher may use substantive background knowledge about outcomes in order to increase the power of her statistical tests. Combining substance and design in this particular way enables valid *and* powerful tests. We build on methods developed by Hansen and Bowers (2009) and with a long history in survey sampling and combine them with techniques of machine learning/data mining/statistical learning

[*]Assistant Professor, Dept of Political Science, University of Illinois @ Urbana-Champaign (jwbowers@illinois.edu). *Acknowledgements:* Thanks to Chris Achen, Joe Bowers, Dan Carpenter, Wendy Tam Cho, Tommy Engstrom, Don Green, Ben Hansen, Jim Kuklinski, and Cara Wong. Part of this work was funded by NSF Grants SES-0753168 and SES-0753164 (collaborative grants with Ben Hansen). Thanks are also due to participants in seminars, talks, and workshops at the Center for Political Studies at the University of Michigan, the Experiments in Governance and Politics network at Columbia University, the American Sociological Association Methodology Section, the Institute for Government and Public Administration at the University of Illinois at Urbana-Champaign, and the Department of Political Science at the University of Southern California. **Early Draft. Please do not cite or circulate without permission.**

[†]Assistant Professor, Dept of Political Science, Fordham University

[‡]PhD student, Department of Political Science, University of Illinois at Urbana-Champaign.

in order to maximize power without compromising the integrity of the resulting statistical inference.

> "It is apparent, therefore, that the random choice of the objects to be treated in different ways would be a complete guarantee of the validity of the test of significance, if these treatments were the last in time of the stages in the physical history of the objects which might affect their experimental reaction."
>
> (Fisher, 1935, page 20)

> "Though the test of significance remains valid, it may be that without special precautions even a definite sensory discrimination would have little chance of scoring a significant success."
>
> (Fisher, 1935, page 25)

This paper proposes a modular approach for assessing the causal effects of interventions in comparative studies. If one can isolate the task of drawing statistical inferences about causal effects from other auxiliary tasks, one can make more robust and transparent claims about relationships between an intervention and an outcome and enhance the precision of statistical inferences. The ability to isolate statistical inference about causal effects from other tasks arises from the use of design-based statistical inference (either Fisher's permutation-based randomization inference or Neyman's sampling-based randomization inference). The enhanced precision arises from the use of predictive models to represent past scholarly knowledge about the outcome.

Further, this modular approach enables the use of machine/statistical learning techniques for prediction, model and variable selection; and the validity of the inferences is not contingent on getting any given model correct, having a correct likelihood function, or the like.

We think of the workflow of analysis of a study of interventions organized by a series of questions:

Is the comparison clear? Given a dataset of units i with an intervention, $Z_i \in \{0, 1\}$ measured for each unit, and background variables, or covariates, $\mathbf{X} = \{X_{i,1}, \ldots, X_{i,K}\}$, available, we would prefer that comparisons of observed outcomes, Y_i, reflect Z_i and not $\mathbf{X}$. Random assignment itself makes this case, in expectation. And other features of design also enhance arguments about unconfoundedness. So, it is common to try to organize data to ensure that like is compared with like (where "like" is defined by $\mathbf{X}$). In this stage of the analysis, one might also decide that certain units are not comparable — that they provide no useful information about the effect of Z_i (for example, they may be overly extreme with no counterpart in the other group as defined by Z_i).

This module in the analytic workflow (often called the "design" phase of the analysis) raises another question: By what standard should we judge whether a comparison is clear? Because we do not assess treatment effects or inspect the outcome as we consider this problem, we are free to propose a standard and to assess the results of different design choices.

What is the scientific question? What are the potential outcomes in the study? What is the quantity of interest? In this paper, we use the

attributable effect, where each treated unit may have its own heterogeneous additive treatment effect. However, other studies may have other targets of causal inference (the difference of average potential outcomes, or ATE, is a common such target). Specification of the target of causal inference tends to imply a procedure that links such targets, stated in terms of unobserved potential outcomes, to observed treatment assignment, Z_i, and observed outcomes, Y_i. So, as we will explain later, for the very skewed outcomes in the London Bombing study, we ask whether the potential outcome after the bombing, $y_{Z=1,i}$ might be related to the potential outcome before the bombing by an individual-specific, heterogeneous, additive effect, such that $y_{Z=1,i} = y_{Z=0,i} + \tau_i$. The average treatment effect target would involve no specific statement about processes at the individual level, and instead focus on a comparison of $\bar{y}_{Z=1,i}$ and $\bar{y}_{Z=0,i}$.

What is the standard by which we should judge a useful causal model? We are free to engage with this question about standards because we separate this stage of the analysis from the others.

How can we use what we know about the outcome to improve precision?

A noisy outcome can mask treatment effects. If we can remove from the outcome variation unrelated to the treatment, then we can enhance the precision in our assessments of causal models. Previous work has shown that linear models can play a role as noise-removers without requiring that they play a role as treatment-effect assessors (Hansen and Bowers, 2009; Rosenbaum, 2002*a*) but that work raises a new question: how should we specify our linear model? That is, given a list of plausible prognostic covariates, we might wonder which particular combination of terms is best at predicting the outcome under control. So, we would like a way to select a model and/or variables in a principled fashion but also in a way that continues to enhance precision without impugning the validity of the statistical inference.

Here, we demonstrate the use of the adaptive lasso (Zou, 2006) but many other algorithms are possible and useful. The key in this part of the workflow is to only use the data from the control group — such that when outcome modeling is happening no treatment effect can be viewed.

How would we know when we have selected a useful model specification? Finding an excellent predictive model does not guarantee precise statistical inference about causal models. In the extreme case, one could have a model which removes nearly all of the variation from the outcome, leaving no variation left for the treatment itself [this could happen if some of $\mathbf{X}$ is correlated in the sample with Z_i even after balancing/matching has been done]. So, in this paper we advocate power analysis constrained by Type I error evaluation. Thus, we can assess the operating characteristics of whatever estimation and/or testing procedure chosen to reflect

on the causal model.

We will show that, up until this stage, no treatment effects are actually estimated. We can organize our data to bolster arguments against confounding, choose outcome models, and assess the power of alternative statistical procedures, all without actually estimating an effect.

What is the effect of treatment? Having chosen a powerful combination of approaches, and being satisfied with the operating characteristics of the method, one may then assess the causal model. The statistical inferences here do not require correct models because they occur separate from the modeling stages.

To what standard should we hold hold our statistical inferences about causal effects (i.e. confidence intervals or hypothesis tests)? That is, how would we know when we have a "useful" confidence interval versus one that may mislead? In this paper we use simple standards that refer to the repeated operation of these procedures: when we know the true value, a useful confidence interval should contain it frequently (and, in fact, a 95% confidence interval should contain this value at least 95% of repeated administrations of the procedure). As is the case with all of the modules of this analysis, other standards may be proposed, defended, and assessed. We start with simple operating characteristics because they are a well known and easy to explain standard.

How fragile are the statistical inferences to omitted and unmeasured confounders? In observational studies, the arguments against confounding may be made more persuasive with the help of a version of a sensitivity analysis — a formal way to address the question, "If I have failed to account for some confounder, how large would it have to be in order to render my results qualitatively different."

The paper illustrates this workflow by pursuing the question of the impact of the July 2005 London bombings on the civic activity of roughly 15,000 UK residents(Office, 2005). But we first briefly describe how statistical inference about a causal effect can be separated from the other aspects of the workflow using a small study studying the impact of a randomly assigned newspaper get-out-the-vote 2005 newspaper advertisement on the aggregate vote turnout of eight US cities (Panagopoulos, 2006).

1.1 Simple Statistical Inference in a Small Field Experiment

In the days just before the November 2005 elections, C. Panagopoulos fielded an experiment to assess the effects of non-partisan newspaper ads on turnout in low salience elections (Panagopoulos, 2006). This was, to our knowledge, the first experiment to investigate the impact of newspaper ads on turnout.

This small pilot study involved eight cities, matched into pairs based on similarity of proportion turning out to vote in the previous election and a few other covariates. One city in each pair was assigned at random to receive newspaper ads in local newspapers encouraging citizens to vote. All cities assigned to receive advertisements had advertisements run in the local newspapers. Table 1 shows all of the observations in the study with key design and outcome features. In three of the four pairs turnout after treatment was higher in the treated city than the control city.

			Turnout	
City	Pair	Treatment	Baseline	Outcome
Saginaw	1	0	17	16
Sioux City	1	1	21	22
Battle Creek	2	0	13	14
Midland	2	1	12	7
Oxford	3	0	26	23
Lowell	3	1	25	27
Yakima	4	0	48	58
Richland	4	1	41	61

Table 1: Design and outcomes in the Newspapers Experiment. The Treatment column shows treatment with the newspaper ads as 1 and lack of treatment as 0. Panagopoulos (2006) provides more detail on the design of the experiment.

Was the effect of the advertisements manifest or negligible? Manifest or negligible compared to what? Fisher's answer was, in essence, "Compared to theory." He suggested the use of a substantively motivated claim about a state of the world as an object against which the nature of our observed effect might be judged. In particular, he focused on one such claim, the hypothesis that the treatment had no effects. In other words, if we compare our observed state of the world with a theoretically motivated claim about way our manipulation ought to work in the world, we will have a way to talk about how surprised we should be upon observing extant differences between treated and control cities. In Fisher's language, an effect is manifest if a hypothesis of no effects makes the effect surprising. There are three parts to Fisher's hypothesis test: a hypothesis, a summary of an observed relationship (often called a "test statistic"), and a probability distribution. The hypothesis and design imply a distribution for the test statistic. And the probability distribution quantifies how surprising it would be to observe the test statistic if the null were true. Let us explain these three parts briefly.

A hypothesis provides a conceptual standard for assessing the 'no effects' question: 'No effects' means no effects. What does 'no effects' mean? Fisher (1935, Chap 2) suggests that a treatment has no effect when units would display the same outcome regardless of treatment condition: for

example, 16% of Saginaw citizens would vote if given advertising and Sioux City would show 22% turnout absent advertising.

A test statistic summarizes observed data. And we might summarize what we observe about the overall relationship between advertising and turnout with the difference in mean turnout between the treatment and control groups — $29.25 - 27.75 = 1.5$ — as our test statistic.

The counter-factual depends on treatment assignment. How surprising is a 1.5 percentage point difference from the point of view of strictly no effects? A phenomenon is surprising when it is not common or expected. So, what kinds of mean differences in turnout ought we to expect if the null were true? What generates variation against which surprise can be assessed? Although there are many ways to conceptualize this variation (if we observed new samples of cities during this election, observed different elections for these same cities, etc...), the one variation producing operation that we can easily formalize in this study is the act of assigning treatment to one city within each pair. As is not uncommon in political science applications, other sources of variation are more vague and thus more difficult to formalize enough to generate a probability distribution — for example, neither the one election nor the eight cities are a sample drawn with some known sampling plan or known probabilistic mechanism from some clear population.

Repeating the experiment produces a distribution of test statistics to represent the hypothesis. So, let us consider what would happen if we were to repeat this experiment, randomly assigning a different set of treatments within pairs. If, by chance, we swapped the observed assignments in all pairs, and represented the null hypothesis of no effects by leaving the outcomes unchanged regardless of treatment assignment, the difference in mean turnout would be -1.5. If we kept the same assignments, but merely switched the first pair, the test statistic would be 0. If we continued to repeatedly reassign treatment differently, calculating the test statistic each time, we would build up a picture of what kinds of mean differences are more or less common in the world of the null hypothesis. This picture is the distribution of the test statistic under the hypothesis of no effects. Another way to put this is to say that the distribution represents the kind of variation we would expect under the no effects hypothesis and our design and data. This "no effect"-distribution (more commonly called the "null-" or "randomization-" distribution) for the mean-difference in turnout between treated and control cities ranges from -5 to 5. About 0.38 of these differences are greater than or equal to 1.5. Convention in the social sciences would be to call the observed number surprising from the perspective of the hypothesis if, say, only 1 in 20 replications of the "no effects"-thought experiment produced values as large as or larger than 1.5. Fisher in 1935 would say that we can never know whether a treatment had an

effect but that, in this study, we can't easily rule out the idea that treatment had no effect — the observed mean difference is just not that surprising from the perspective of the null hypothesis.[1]

The nature of Fisher's statistical inference is that simple. Following Fisher, our investigation centered around converting a vague but substantively interesting question about effects into a precise hypothesis about no effects. The computational thought-experiment produced a quantitative description of the hypothesis in action. Our observed value is not very surprising from the perspective of the hypothesis of no effects.

1.2 The Promise of Fisher's Method as of 1935

Fisher's insight in 1935 allows the precise assessment of a substantively meaningful hypothesis with a probability statement justified by the research design without necessary recourse to a linear model, let alone the classical linear regression assumptions. Fisher's approach is promising at first glance because of what it does not require. The validity and meaning of the p-value reported above did not require knowledge of a sampling plan or a population from which sampling occurred. In fact, these cities are not a random sample of a well delineated population of cities. Nor did this p-value require us to turn our knowledge of the stochastic process producing aggregate turnout in US cities into a likelihood function and associated parameterization. Nor did we make asymptotic arguments. The eight cities support a valid statistical inference here with no apologies necessary for a small sample.[2]

So, it should be clear that statistical inference can exist outside the context of a linear model. Linear models, however, are useful. How can we take advantage of the benefits of the linear model (and perhaps even the burgeoning literature on data mining and machine learning) while keeping statistical inference separate from all of the many choices required of data model building? We use a quasi-experiment to engage with this question.

[1] All computations reported in this paper (like the p-value of 0.38 reported above) will be available for reproduction and exploration by interested readers from `http://jakebowers.org`. Some of the functions come from a pre-release version of the `RItools` (Bowers, Fredrickson and Hansen, 2009) package for R. This paper is written in the mixture of R and LaTeX known as Sweave (Leisch, 2002, 2005). Those interested in learning more may download the source code of the paper and apply themselves to adapting it for their own purposes.

[2] We should note here that research design combined with asymptotic arguments and a commitment to a particular scientific quantity (the "average treatment effect") can also justify tests of "weak hypothesis of no effects." We do not engage with that approach here first because we have a tiny sample of cities and second because it is well discussed in many extant sources [cites — Gerber and Green textbook, etc..].

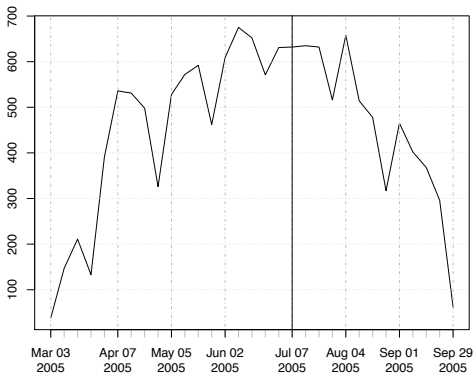

Figure 1: Number of respondents interviewed in the 2005 Home Office Citizenship Survey by week. The vertical line marks July 7, the date of the subway and bus bombings in London.

1.3 The London Bombings of July 2005

During the morning commute on July 7th, 2005, four suicide bombers placed explosives on the London subway system and one bus killing 52 people and injuring over 700 (BBC News, 2008). The bombers claimed to be soldiers in a battle between "the West" and "Islam." In general, we know that terrorist attacks aim to disrupt the targeted civil society. The question we pursue here is whether this political violence influenced the civic engagement of ordinary people in Britain. The data come from the 2005 Home Office Citizenship Survey, which was an in-person survey of roughly 14,000 British residents in England and Wales. The survey was conducted in respondents' homes over the period of March 8 to September 30. 8103 people were interviewed in the weeks proceeding the bombings, and 5975 people were interviewed after the bombings. Figure 1 shows the number of subjects interviewed per-week before and after the attacks of July 7.

Terrorism attacks seek to disrupt civil society. If these attacks were successful, respondents interviewed after the bombing should report fewer volunteer activities and a willingness to assist other citizens. Yet the attacks may impel citizens to rally and may provide a very salient reason for public action. If the bombings increased social cohesion, respondents interviewed after the bombings should report more civic participation.

To address these competing hypotheses, we consider two questions asked on the 2005 Home Office Citizenship survey:

- "Approximately how many hours have you spent helping [previously listed groups, clubs or organizations] in the past 4 weeks?"
- "Approximately how many hours have you spent doing [previously listed volunteer activities] in the past 4 weeks?"

Figure 2 shows the distribution of the sum of the answers to these two questions compared before and after the bombing of July 7, 2005. While the distributions before-vs-after are similar at the lower end (with the median subject reporting zero hours and zero pounds donated), after the bombing, subjects in the upper end of the distribution reported volunteering more hours.

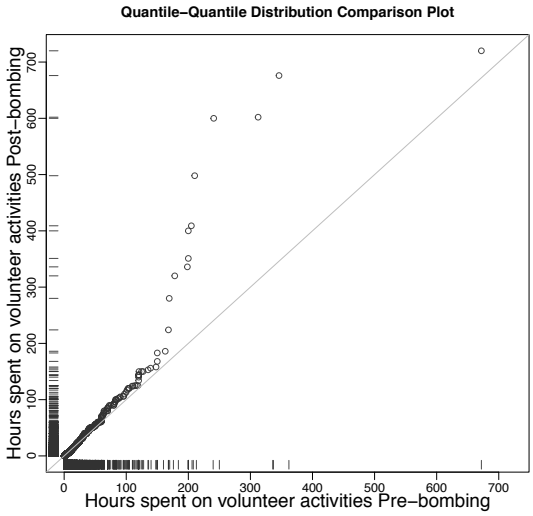

Figure 2: Quantile-Quantile plot comparing pre- and post-bombing respondents on number of hours reported spent volunteering for groups, organizations, or clubs plus number of house spent on volunteer activities in the last 4 weeks.

Did the bombings on July 7, 2005 cause an effect on the short term volunteering of British residents? Our aim in this paper is not to answer this question conclusively, but to use it to illustrate the benefits and process of breaking the task of doing statistical inference about causal effects into semi-independent modules.

2 Design: Clarify Comparisons

If we compare the civic activity of survey respondents before versus after the bombing, will we capture the effect of the bombing? The danger, of course, is that subjects who would donate more to charity after a bombing could also be more likely to be interviewed after the bombing (or vice-versa). If this were the case, any differences in the behavior of pre-vs-post bombing respondents may tell us more about differences in type of person than about counter-factual response to the bombing. We can think of these possible sources of imbalance as either the result of the mechanism by which the survey was carried out or the result of individual characteristics of respondents themselves.

In the first case, the survey was designed to be a nationally representative probability sample. The order of interviews within the survey, however, was not random. Figure 3 shows the counts of subjects interviewed before and

after the London bombings, grouped by the 10 administrative regions used for sampling. While the overall survey was designed to a random sample of the population, the implementation of survey varied greatly in which regions were most heavily interviewed before and after the bombing. For example, while about half of the Londoners were interviewed before the bombing, most of the respondents in Wales were interviewed before the bombing. Naive estimates of the bombing's effects by compared volunteering before before-vs-after may simply be estimates of the effect of living in London instead of Wales.

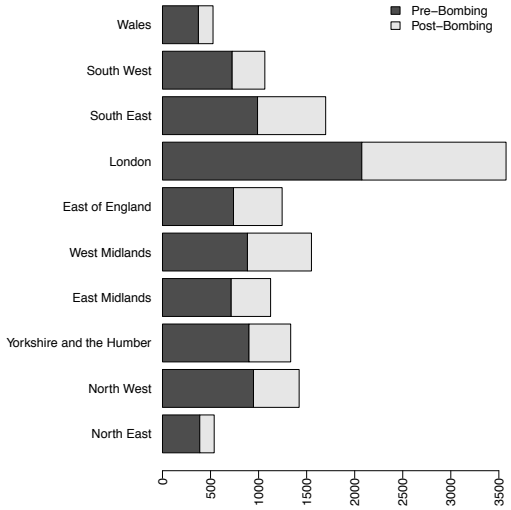

Figure 3: The counts of respondents interviewed before and after the London bombings, grouped by the governmental regions used for sampling and organization of the fieldwork for the 2005 Home Office survey across the UK. While the overall survey was designed to be a representative sample of the population, the order of the interviews during the sampling period was not uniform.

Individual characteristics may also influence which citizens are willing to answer a survey. If the bombing increased civic participation among those predisposed to participate, then such people might have been easier to reach and interview before the bombing compared to after the bombing. If people (or the survey interviewers) systematically tried to include people in the survey on the basis of their predicted or past civic activity, then pre-vs-post comparisons will tell us more about the kinds of people who wanted to be interviewed at a particular moment rather than about an effect of the bombings. As with the regional comparisons, we can assess whether the kinds of people who were interviewed before the survey differed from those interviewed after the survey in terms of variables that might confound pre-vs-post differences. The leftmost box plot in Figure 4 plots the standardized mean differences across many covariates for pre and post bombing respondents. With many differences in excess of 0.05sd (positive or negative), this plot suggests that the kinds of people interviewed before the bombing differed from the kinds of people interviewed after the bombing. However, the range of this plot suggests

that, the people interviewed before-vs-after the bombing were quite similar in general — no differences were more extreme than about .15sds. If we assessed differences between respondents using another cut point (say, differences between people living in London versus Wales, or differences between Muslims and Christians) we would see differences well in excess of 2sds.

Is a difference of .15sds large? Recall Fisher's thought-experiment. We do not know how to answer questions about "large" without also asking, "compared to what?" What standard ought to govern answers to this question? In this paper we use what we consider to be a minimal standard: the equivalent randomized experiment. That is, we could ask, "If there were no systematic relationship between this covariate and the timing of the bombing such as the relationship that we would expect from a randomized experiment, how strange would it be to see a difference of .15sds?" If it would be very strange to see a difference of .15sd from the perspective of a randomized experiment, then we suspect that comparisons pre-vs-post bombing will have no where near the clarity of comparison offered by a randomized experiment. If differences of .15sd would be typical of a randomized experiment then we have less clarity of comparison than a randomized experiment, but we know that the confidence interval for our outcome comparisons will be large compared to the potential confounding offered by this one covariate Hansen (2008); Bowers (2011). Assessing balance in this way in a non-randomized study raises other questions — for example, our representation of the randomized experiment that serves as our minimal standard could ignore some other covariate which drives the timing of interviews. So, we would not accept the null both because it is not sensible to accept null hypotheses in general, but also because we know the null cannot be true. We will engage with these questions both by assessing balance using many covariates and also using a formal sensitivity analysis.

Now, although we could answer this question about one covariate using a randomization-based hypothesis test just as we assessed this hypothesis in the Newspapers data, here we have chosen to inspect balance on 140 covariates. In practice, of course, even excellent randomization does not guarantee that all variables will be perfectly balanced. And it would be reasonable to see some p-values that are small merely through chance: here we would expect to see 7 p-values less than .05, and 1.4 less than .01. Hansen and Bowers (2008) proposed to assess the imbalance across all covariates and their correlations using a test statistic summarizing the entire set of mean differences assessed: they represent the overall balance of the sample with a statistic, d^2, which will be small when the data when the data are incongruent with the hypothesis of no systematic relation and large when data could easily have emerged when no systematic relationship exists. Figure 4 also shows the p-value for a χ^2 based on the d^2 statistic.

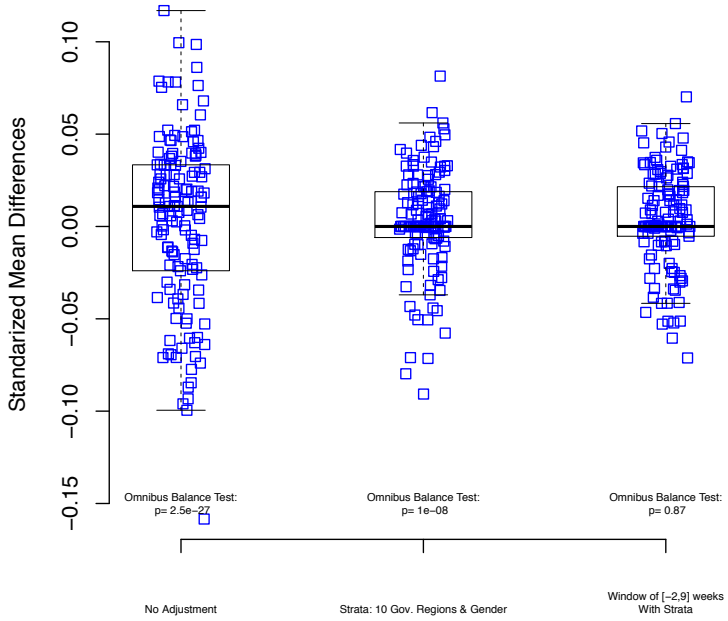

Figure 4: Balance assessments for 140 covariate terms before versus after the London 2005 bombings. Left panel shows difference in proportions or means for binary and continuous covariates respectively. The p-value for the d^2 omnibus balance test is reported in the x-axis labels. The comparison is between unadjusted assessments, assessments conditional on gender and governmental region of the country (making 20 strata), and assessments conditional on those strata within a window of 2 weeks before the bombing and 9 after the bombing.

This statistical test also suggests a manner to enhance the clarity of our comparisons. If we can organize our data such that we achieve a large p-value for d^2, then our design will only protect us from confusing treatment effects with observed covariate-differences. Given fact that the bombing occurred without warning, and that our covariates do not look so imbalanced in substantive terms (even if the p-value on the omnibus test is very small because of the large sample size), we proceed here in the simplest possible way to create a research design: by simply grouping our respondents into homogeneous groups. If the within-group differences are small, we have the analog of a block-randomized experiment. While we could use a technique such as propensity score matching to achieve this goal, we will use a slightly less complicated procedure. We begin by stratifying by region (because the survey fieldwork was organized by region) and gender (because we know that civic engagement and willingness to be interviewed differ between men and women in places like the USA and UK). The second column of Figure 4 shows the distribution of differences after stratifying this way. While the spread of

imbalances decreases, the omnibus test of balance still indicates that our data differ from what we would expect from an equivalently blocked randomized experiment. In order to further constrain our data, we restrict attention to a window of weeks around the time of the bombing. While the original survey was conducted over a 30 period, we searched over all small windows of time to find that period that had the highest p-value on our d^2 test, still stratifying on region and gender. A window of 2 weeks before the bombing to 9 after the bombing provided a dataset that was well balanced on the measured covariates, at the cost of sample size. While the original survey included 14078 respondents, after shrinking the window 6451 subjects remain.

Although other post-stratifications might enable better or equivalent balance, we proceed with this 20 strata design because it follows from the field operations of the survey, because we doubt that individuals waited until after the bombing to be interviewed (because, the bombing was unexpected), and because we can address questions about fragility of inferences to confounds during the sensitivity analysis process.

2.1 Module 1: Enhance clarity of comparisons

Because we separated the examination of treatment effects from statistical adjustment of the data, we were free to search for a design that satisfied a standard for interpretable or clear comparisons. We used the same testing framework that we described in our analysis of the Newspapers study, but this time the testing was focused on the question of congruence between different methods of organizing our data and the minimal standard of the equivalently randomized experiment. In the end, because the bombings happened in a manner that was unexpected by the Home Office survey field staff, we had to do relatively little to produce a research design where outcome comparisons would not overly reflect covariate differences conditional on strata.

3 Formalizing Science: Causal Models and Causal Effects

Until now we have engaged with questions about "effects" or differences by focusing attention on the thought-experiment of no effects or no differences. Yet, theoretically interesting questions tend to include concerns about "no effects" but also questions about "effects" themselves. Here we formalize the ideas behind the "no effects" hypothesis and use this notation to structure thinking about simple hypothesis about effects themselves.

3.1 Writing Causal Models

The Neyman-Rubin causal model Rubin (1974) asserts that outcomes are fixed quantities realized with respect to a treatment. This heuristic for thinking

about causal relations allows a formalization of hypotheses about such relations, and also allows us to link questions about unobserved quantities and comparisons (i.e. hypotheses) with observed data in order to assess evidence against those hypotheses.

Indexing these fixed potential outcomes for a single subject i and a treatment vector $\mathbf{Z}$ assigning all other units in the study to either treatment, $Z_i = 1$, or control $Z_i = 0$, we write $y_{\mathbf{Z},i}$ as the outcome we would observe for i if the realized treatment were the vector $\mathbf{Z}$. If we are willing to disallow interference between units, the previous notation can be reduced to $y_{Z_i,i}$ (Bowers, Fredrickson and Panagopoulos, 2012), a convenience we will use for the rest of this paper. A *causal model* is a function that relates potential outcomes. The simplest causal model states that all potential outcomes (for unit i) are equivalent:

$$y_{z,i} = y_{z',i} \, \forall z, z' \in Z$$

In the case of our bombings example, there are two levels of Z, which we label "control" and "treated", respectively. And the simplicity of the example allows us to collapse the previous model into a simpler statement: $y_{1,i} = y_{0,i} \equiv y_{1i} = y_{0i}$. In the case of the bombings example, we label respondents interviewed prior to the bombing as control subjects and post-bombing interviewees as treated subjects.

Causal models also allow us to define *causal effects*. A causal effect summarizes the differences in potential outcomes for different levels of Z. The model represented in the previous equation states that all potential outcomes are equivalent and thus implies that Z has no effect on the outcomes. For this reason, this model is frequently referred to as the "model of no effects". To include causal effects in the model, we add parameters. For example, a model stating that each unit may receive an idiosyncratic boost from treatment would be written as:

$$y_{1i} = y_{0i} + \tau_i \, \tau_i \geq 0 \forall i$$

Although one could assess hypotheses about all possible τ_i the process would be very time consuming, and perhaps yield little of substantive use: knowing that it is implausible that $\tau_1 = 1, \tau_2 = 0, \tau_3 = 0, ..., \tau_N = 1$ but plausible that $\tau_1 = 0, \tau_2 = 0, \tau_3 = 1, ..., \tau_N = 0$ tends not to address a scientific question about overall effects. Rosenbaum (2001, 2002*b*) proposes a function of the τ_i as an object of scientific interest when outcomes are binary, and he writes $A = \sum_{i=1}^{n} Z_i \tau_i$ as the effect attributable to the treatment on the treated, or the "attributable effect." In this paper we extend the analysis of attributable effects to the case with a count outcome (so that τ_i can be any non-negative integer rather than restricted to 0 and 1). Although there are many ways for all of the possible τ_i to add up to a given A a hypothesis about A that receives a p-value of, say, .05, indicates that no hypothesis about the set of τ_i adding up to A would have a p-value greater than .05. Thus, a confidence

interval for A not only tells us something substantively interesting: how many more hours of volunteering that would be plausibly caused by the bombings in our sample (or, if we dichotomize the outcome, how many people volunteered at all after the bombing who would not have volunteered before the bombing).

3.2 Randomization-based Statistical Inference for Causal Effects

In the previous section, we wrote models relating fixed potential outcomes as if we knew these outcomes for all units. Of course, for any practical application, this is not the case. We observe one, and only one, outcome per unit. In the case of the bombings in London, we observed the pre-bombing outcomes only for respondents interviewed prior to the bombing and the post-bombing outcome only for respondents interviewed after the event. To know τ_i with certainty, we would need to observe both y_{0i} and y_{1i}, but due to the fundamental problem of causal inference we can only observe one of these outcomes per unit (Holland, 1986). We can, however, make *inferences* to the population of unobserved potential outcomes, which allow us to investigate hypotheses about causal parameters. In this section we demonstrate how these inferences can proceed directly from the randomization of treatment (Z).

In the previous section, we defined two example models: a model of no effects and a model of varying additive effects (where τ_i can vary by unit). For any model, we can derive testable hypotheses to apply to data related to the causal effect parameter. These hypotheses take the form of a transformation of observed data y_{zi} into the data that would been observed if the hypothesized parameters were true and all units received $z_i = 0$, which we label $\tilde{y}_{0i}$. When τ_i is allowed to vary for each unit, our hypothesis is a vector $\boldsymbol{\tau}_0$, and $\tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\tau}_0^T \mathbf{z}$. Observe that for the hypotheses of $\tau_0 = 0$ and $\boldsymbol{\tau}_0 = \mathbf{0}$, the transformation of the observed data is the same as that implied by the model of no effects. That is, our hypothesis comparing partially observed potential outcomes and our understanding of the design of the study (that generates and provides structure for $\mathbf{z}$) combine to imply particular adjustments to our observed values if we are to entertain the hypothesis seriously. Note that for all control units $y_{0i} = \tilde{y}_{0i}$.[3]

To evaluate a hypothesis, we entertain the question, "what is the probability of seeing data as or more extreme than the observed if the hypothesis were true?" Using a suitable test statistic $T(\tilde{\mathbf{y}}, \mathbf{z})$, a function that maps data and a treatment assignment to a real value, every possible $\mathbf{z}$ could be evaluated, generating the distribution of the test statistic that describes the null hypothesis. The p-value of the hypothesis is then:

[3] If units are allowed to interfere — the treatment to i changes the outcome of j — this relationship does not automatically hold. Bowers, Fredrickson and Panagopoulos (2012) extend the basic methodology in this section to include cases of interference.

$$\Pr(t > T(\tilde{\mathbf{y}}, \mathbf{Z})) = \sum_{i=1}^{|\Omega|} \mathrm{I}\left(T(\tilde{y}, \mathbf{Z}_i)\right) \Pr(\mathbf{Z}_i)$$

Where $t = T(\tilde{y}, \mathbf{z})$ is the observed value of the test statistic and Ω is the sample space of the treatment vector $\mathbf{Z}$. While the sample space can be very large, convenient large sample approximations exist for many sampling plans and test statistics. Confidence intervals are defined as the set of hypotheses that are not rejected for a given α level.

While the previous approach describes how to test models at the unit level, such as the model $\tilde{\mathbf{y}} = \mathbf{y} - \tau_0 \mathbf{z}$, models that consider aggregations merit additional consideration. In the case of the London bombings, our model allows the treatment effect to vary by unit, and we wish to test hypotheses over the sum of the treatment effects, $A = \sum_{i=1}^{n} Z_i \tau_i$. In theory, we could develop a confidence interval for hypothesized A_0 with the following algorithm: For a given hypothesized A_0 (say, $A_0 = 1$), we list all k of the ways that a vector of $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$ can be summed to that A_0.[4] For each possible vector, generate a p-value using the previously described method, labeling it p_k. We define the p-value of A_0 as $\Pr(A_0) = \max(p_1, p_2, \ldots, p_k)$. Therefore an A_0 with a large p-value indicates the data were not unlikely for at least one vector of A_0. As before, by repeating this procedure for many hypotheses A_0, we can generate a confidence set using the hypotheses not rejected at a given level.

An advantage of this approach is that the target of statistical inference is defined in terms of a counterfactual causal relation. It is also worth noting here that we never required any statements about the distribution of the outcome in any of our hypothesis tests so far. This fact becomes particularly useful in this case: respondents frequently listed zero hours volunteered and the distribution also had very long tails. With such data, a scholar using a conventional linear model based approach would be forced to choose between various zero-inflated count-distributions. Randomization inference, on the other hand, side steps these issues by focusing on the random assignment to treatment or control, instead of the distributional qualities of the outcome. With the selection of a powerful test statistic T, these methods can work as well for zero-inflated or skewed outcomes as they do for bell shaped outcomes. More importantly, in a randomized experiment, the validity of the tests is the same across the different distributions, even if the power may differ.

An obvious disadvantage of this algorithm is that in a dataset of thousands of respondents, the number of atomic hypotheses associated with each composite hypothesis is enormous. Luckily, Hansen and Bowers (2009) show that, in large samples, with binary outcomes that are not too skewed, statistical inference about A can proceed building on the survey sampling literature: a confidence interval about A is, in essence, a confidence interval about a finite population total where the finite population is the experimental pool. Since

[4] Constraining τ_i to a non-negative integer allows the use of a *partition*, the set of which will be finite, if large, for any given A_0.

a binary outcome is a kind of count (a count with only two values), we can directly apply those results to the problem of totals of counts. We explain this reasoning here.

We begin by relating A to observed outcomes:

$$A = \sum_{i=1}^{N} Z_i \tau_i \tag{1}$$

$$= \sum_{i=1}^{N} Z_i (y_{i,1} - y_{i,0}) \tag{2}$$

write C as the set of control units

$$= \sum_{i \not\in} y_{i,1} - \sum_{i \not\in} y_{i,0} \tag{3}$$

recall $Y_i = Z_i y_{i,1} + (1 - Z_i) y_{i,0}$ so

$$= \sum_{i \not\in} Y_i - \sum_{i \not\in} y_{i,0} \tag{4}$$

So we can write $A = \sum_{i \not\in} Y_i - \sum_{i \not\in} y_{i,0}$ and note that this is the same as $A = \sum_i Y_i - \sum_i y_{i,0}$ because $\sum_{i \in C} Y_i - y_{i,0} = 0$

The attributable effect, then, combines the fixed total, $t_U = \sum_i Y_i$ (the total number of hours volunteered by the respondents in the experimental pool or universe), with the partially observed quantity, $t_C = \sum_i y_{i,0}$ the total number of hours the survey respondents would have volunteered had they been interviewed before the bombing. Inference about A then is equivalent to inference about the potential total in the control group $t_C = \sum_i y_{i,0}$

We know from the survey sampling literature that an unbiased estimator of t_C is $\hat{t}_C = N\bar{Y}_C$ Lohr (1999) and in large samples under regularity conditions allowing a central limit theorem to operate an approximate confidence interval would be: $\text{CI}(\hat{t}_C) = \hat{t}_C \pm z_{\alpha/2}\text{SE}(\hat{t}_c)$. Returning to A, using the decomposition above, we can write a confidence interval as: $t_U - \widehat{\text{CI}}(\hat{t}_C)$. This conceptualization of A was used by Hansen and Bowers (2009) to create an large-sample approximate confidence interval for binary outcomes without the tedious algorithm described above: by thinking about the problem as a survey sampling problem where a confidence interval for a finite population total is desired.

3.3 Module 2: Specify the question.

Although many possible substantively useful questions exist, we here focus on questions about heterogeneous additive effects of the bombing on those exposed to the bombing. The attributable effect is a useful conceptual model

in this case both because it allows us to represent the values of the outcome (which is integer valued with many zeros and long tails) and also because it can be re-conceptualized and simplified with reference to the survey sampling literature. If, however, we found that using the large-sample approaches developed for statistical inference about finite population totals did not work well (meaning, say, confidence intervals created on this basis contained the truth less than $100 \cdot \alpha\%$ of the time), we could return to the core algorithm if needed.

To what standard should we hold causal models? Clarke and Primo (2012) echo a long literature in embracing simple models when they suggest that a good model is "useful" rather than "correct." Although this paper is not the place to discuss the philosophy of science, we suspect that questions about the adequacy of a causal model are best engaged with reference to the particular theory and substance motivating a study. Other work (Bowers, Fredrickson and Panagopoulos, 2012) and (Rosenbaum, 2009, Chap 2) suggests that assessing multiple causal models may be more productive than defending any one model as most useful. In this paper, we stick with the single model of additive heterogeneous causal effects because it appears to address the kinds of substantive questions that arise about interventions like the London bombings: how much activity did we see after the bombings that we would not have seen in the absence of the bombings?

4 Improving precision through modeling

Political scientists know a lot about civic activity. For example, we strongly suspect that those people with more education will report comparatively more civic activity than those people with less education whether or not these people are exposed to a bombing. Even if education is balanced between the treated and control groups (conditional on design) we still might want to in some way remove education-related variation from the outcome in order to enhance the precision of our tests. There are a few ways to convey the intuition behind this idea: one way is to imagine two distributions of the outcome, one each each treatment group. If we are asking about whether the two centers of the distributions differ, we will have more power to address that question the tighter the distribution around the respective centers. Another intuition might be familiar to those who have studied the classic linear regression model. In the canonical iid version of that model recall that the standard error of the coefficients is determined by $\sqrt{\text{diag}(\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})}$. The residual variability of the outcome influences this expression multiplicatively: $\sigma^2 = \mathbf{e}^T\mathbf{e}/(n-k)$ where $n-k$ are the degrees of freedom of the model and $\mathbf{e}^T\mathbf{e}$ is the sum of squared residuals. So, a model which allows large residuals will also have large standard errors. Although we are not here using the standard errors that arise from the linear model for statistical inference, nor are we comparing two distributions, the intuitions are the same: a comparison of highly variable quantities allows for less precise statements than a comparison of comparatively less noisy quantities.

So, the question for us at this point in the workflow is whether we know something about the outcome that we can use to reduce noise. We have long known that one can shrink the sizes of confidence intervals using such information to "adjust" our test statistics for covariates (often known as "covariance adjustment").[5] Yet, neither Fisher nor Neyman made it clear how one might use the additional information in randomization-based hypothesis testing. And, the extensive literature on adjustment of randomized experiments is contentious because of fears that exploration of adjustment strategies will led to misleading characterizations of treatment effects: estimating many different linear regression models in the process of adjustment may yield a treatment effect that appears particularly large, but which, in fact, is merely an artifact of the kinds of linear interpolation and extrapolation of a given specification. Moreover, statistical inference for treatment effects after such data snooping is also suspect: one hundred tests with pre-specified rejection level $\alpha = .05$ will falsely reject the true null of no effect five times. These problems of data snooping are well known. Current best practice in the world of clinical trials involves a public declaration in advance of the experiment of an adjustment strategy using one of the online trial registries and also to report unadjusted results.[6]

At first glance it appears that any attempt to adjust experimental data without advance public registration will run afoul of such criticisms. While we do not dispute that registration is a gold standard, here we use the fact that we can separate our assessments of causal effects from our adjustment process to enable a search over adjustment specification without risking the problems of data snooping. We follow the developments in Hansen and Bowers (2009) which in turn relies on some of the basic theory of survey sampling to build a method for covariance adjustment. In this paper we add the step in which we use a machine learning procedure to search for the model specification, and we contribute to the machine learning literature in a small way by using Type I error rate and power of the test as our criteria for choice of tuning parameters rather than cross-validation and some information criteria or error rate estimate.

Recall that we were able to express our basic counter-factual quantity, $A = \sum_i Z_i \tau_i$ using observed data combined with partially observed data: $A = \sum_i Y_i - \sum_i y_{i,0}$. Inference about A thus can focus on inference about $\sum_i y_{i,0}$ or t_C – the total number of hours volunteered if all respondents in the survey had been interviewed before the bombing. If we consider the survey respondents as a finite population and those interviewed before the bombing as random sample from this population, we can estimate $\hat{t}_C$ using either an unadjusted

[5] See Cox and McCullagh (1982); Bowers (2011); Keele, McConnaughy and White (2010); Miratrix, Sekhon and Yu (2012) for only a few of the overviews of such adjustment.

[6] http://www.consort-statement.org/ The Experiments in Governance and Politics group is currently working to translate such procedures to the social sciences. (http://e-gap.org/resources/standards-project-registration/). See also [cite to recent political analysis issue].

method — $\hat{t}_C = N\hat{Y}_{i\in C}$ — or if we have some covariates observed for the whole population, we can take advantage of them and use what is commonly known as the "regression estimator of the finite population total" such that $\hat{t}_C = \sum_{i\in U}\hat{Y}_i + \sum_{i\in C}(Y_i - \hat{Y}_i)$ (Lohr, 1999, Chap ?4?) (See also (Särndal and Swensson, 2003, Chap ?)) where we estimate (1) $\hat{\boldsymbol{\beta}}$ from a linear model fitting $Y_{i\in C}$ as a function of the control group covariates and then (2) extrapolate from the control group to the entire study using the covariates observed for both groups $\hat{Y}_i = \mathbf{X}\hat{\boldsymbol{\beta}}$. Notice that because we subtract $\hat{Y}_i$ for $i \in C$ in the second term of the expression for $\hat{t}_C$ the estimator of the total outcome in the control amounts to using the total of the observed control outcomes plus the extrapolated control outcome for the treated observations. In stratified designs, $\hat{t}_C = \sum_{i\in U}\hat{Y}_i + \sum_{i\in C}(Y_i - \hat{Y}_i)/\pi_i$ where we weight the second term by the sampling probability, π_i which is the ratio of controls to treated in a given stratum.

4.1 Model Selection

Although this is a well established way to use auxiliary information to improve estimates of finite population totals from sample information, any attempt to use covariates must engage with a number of questions: Which covariates ought to be included? Which function of covariates ought to be fit? Which fitting procedure (least squares? least absolute deviations? outlier resistant least squares?)? Luckily, the rough procedure of fitting on the control group only allows analysts to use the data to answer such questions without calculating treatment effects: covariance adjustment in the context of randomization inference separates adjustment from assessment of treatment effects.

In this paper we use one of the more well-established machine learning techniques to answer some of the questions raised in the previous paragraph. That is, the general answer to the those questions is that we want the function and set of covariates which enable the most powerful tests of our causal model. The best fitting prediction model of outcomes in the control group is one candidate for creating most powerful tests. And a host of procedures for selecting predictive models exists — most notably the penalized linear model based approaches inspired by Tibshirani (1996)'s lasso penalized least squares model. In this paper we use a penalized regression model that is a superset of the lasso and ridge models know as the elastic-net model Zou and Hastie (2005) in an adaptive framework Zou (2006). We see a $\hat{\boldsymbol{\beta}}$ which minimizes the least squares criterion plus the elastic-net penalty:

$$\hat{\boldsymbol{\beta}}(\text{Elastic-Net}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha)|\beta_j|)/w_j \quad (5)$$

where, the w_j is a weight on each element β_j of $\boldsymbol{\beta}$ that arises from a prior

elastic-net fit.[7] This criterion involves both the lasso or L_1 penalty as the $|\beta_j|$ (i.e. models with larger coefficients in absolute value will be less preferable than models with smaller coefficients (or coefficients of 0)) and the ridge or L_2 penalty as β_j^2 (i.e. models with larger coefficients will be increasingly less preferred although the penalty for models with small coefficients is less severe than the lasso penalty would be). The parameter α weights the two different penalties. A large body of evidence suggests that models like this (of which there are now many varieties) make more accurate extrapolations (i.e. would predict how treated subjects would act in the control condition) than models without such penalties.[8] Figure 5 shows an example of how tuning parameter choice amounts to model choice. Here, we specified a simple model whereby hours spent volunteering in the control group would depend on immigrant status, household size, age, years lived in current home, household income, and gender. We also set $\alpha = .5$ to equally weight the two types of penalties. And we set $w_j = 1$ for all coefficients. As the penalty parameter λ goes from approx 0.0113873314966577 or $log(\lambda) = -4.4752538137255$ to approx 3.02457520686365 or $log(\lambda) = 1.10677065413885$, the sizes of the coefficients predicting volunteering among control group respondents decrease to zero. For the largest λ only whether the respondent was female and age remain in the model. And the coefficients for household size and home tenure (or length of time lived in current home) are set to zero very quickly.

[7] Zou (2006) shows that, under certain regularity conditions, by adding the weights from either a preliminary OLS or ridge-penalized regression to a lasso penalized model, the correct model will be chosen.

[8] See Efron's latest book on more theory about why penalized models seem to work so well compared to unpenalized models [cite].

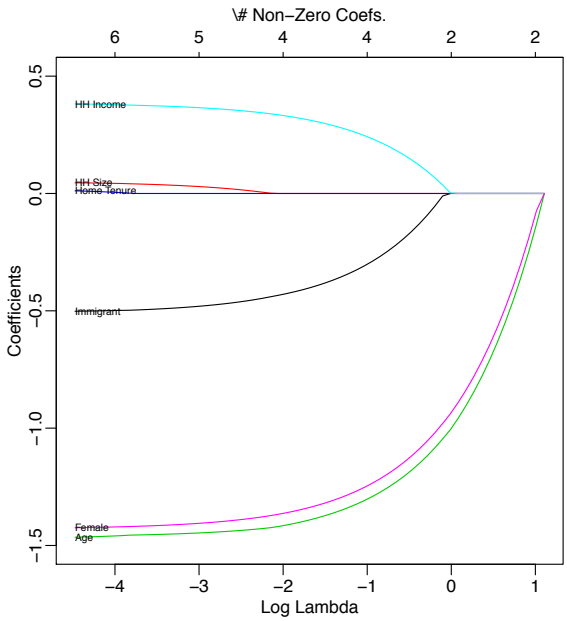

Figure 5: Example of the influence of choice of λ on model choice given $\alpha = .5$. As the penalty parameter λ goes from approx 0.0113873314966577 or $log(\lambda) = -4.4752538137255$ to approx 3.02457520686365 or $log(\lambda) = 1.10677065413885$, the sizes of the coefficients predicting volunteering among control group respondents decrease to zero.

Yet, merely choosing a model that fits the control group outcome well is not enough: a model which perfectly fit the control group could either also perfectly fit the treatment group, which would lead to invalid statistical tests (as noted by (Hansen and Bowers, 2009, §3.3)) or imagining the rejection rate of a statistical test with a process that has no variation. More likely, an excellent fit in the control group would extrapolate very poorly — leading to poor power in casual effect assessment.[9] Although the literature on machine learning advises analysts to choose tuning parameters using cross-validation and guided by mean-squared error or a penalized version of mse (like the AIC or BIC), in this paper we focus attention directly on the operating characteristics of the statistical inference procedure itself. That is, we choose a set of tuning parameters (λ and α) that maximizes the power of the test while maintaining the the size of the test at .05 when the level of the test, α, is .05. One way to represent this optimization problem is as a lexical optimization problem: maximize power subject to the constraint that the size of the test as at or less than the level of the test. In other words, we would like confidence intervals for A which contain the truth $100\alpha\%$ of the time (i.e. a confidence interval with correct coverage), which are increasingly likely to exclude values of A as they diverge from the truth, and which, in general, are as narrow as possible

[9]See (Hastie, Tibshirani and Friedman, 2009, Chap 7) for a lucid discussion of the problems of overfitting for prediction and for attempts to characterize error of extrapolation for a chosen model.

subject to the previous two constraints.

A confidence interval for A under a Normal approximation would be $A \pm z_{\alpha/2}\sqrt{\text{Var}(\hat{A})}$. Because we write $A = t_U - t_C$ and $t_U = \sum_{i=1}^{N} Y_i$ is fixed across randomizations, the estimator of A is simply $\hat{A} = t_U - \hat{t}_C$ and $var(\hat{A}) = var(\hat{t}_C)$. The variance of the regression estimator of the finite population total, $\hat{t}_C = \sum_{i \in U} \hat{Y}_i + \sum_{i \in C}(Y_i - \hat{Y}_i)/\pi_i$ is well know from the survey sampling literature. If $\boldsymbol{\beta}$ were fixed (i.e. not re-estimated for each randomization), then the first term in $\hat{t}_C$ is fixed and the variance of $\hat{t}_C$ would be equal to the variance of $\sum_{i \in C}(Y_i - \hat{Y}_i) = \sum_{i \in C} e_i$ or the sum of the residuals from the control-group restricted model fitting process. Thus, for a single stratum, we can write $var(\hat{t}_C) = n^2(1 - m/n)(s^2/m)$ where $s^2 = \sum_{i \in C}(e_i - \bar{e})^2/(m-1)$ and m is the number of controls, n is the total population size. When we have multiple independent strata, we can calculate $var(\hat{t}_C)$ for each strata and sum the results across strata. Hansen and Bowers (2009) showed that when $\boldsymbol{\beta}$ is estimated (i.e. differs from randomization to randomization) the variance of A is closely approximated by the expressions derived from holding $\boldsymbol{\beta}$ fixed.

In this paper we engage with the approximation problem in two ways. First, we calculate an adjusted version of this variance, $var(\hat{t}_C) = n^2(1 - m/n)((s^2 + \text{adj})/m)$ where the adjustment represents an estimate of the error of extrapolation from the control group (or training set) to the whole population which varies across randomizations and models. To represent error of extrapolation (which ought to inflate the variation of predictions beyond than expected from randomization itself) (Hastie, Tibshirani and Friedman, 2009, §7.26) suggest an analytic calculation akin to Akaike's Information Criterion, $\text{adj} = 2 * (d/m) * \sigma_l^2$, where d is the "effective model size" or an estimate of degrees of freedom (here the number of non-zero coefficients), and σ_l^2 is the mean-squared error from a low bias model (i.e. a very saturated and unpenalized model). Although we could have used a bootstrap procedure directly or followed other advice from Hastie, Tibshirani and Friedman (2009) and used cross-validation, we decided to use the analytic approaches here and then to assess the operating characteristics of the resulting intervals before adding another computationally intensive step to our workflow. That is, we first addressed the approximation by an approximate adjustment known to characterize the kinds of additional variation that might arise from different fits on the control group, and second because the assessment of the coverage rates and width of the confidence intervals was a part of the optimization process, we merely inspected the results from the chosen model to see if, in fact, the approximation worked. If the our intervals exclude the truth too often, then we suspect that at least one of our approximations is not working, and we would be free to use other strategies such as the bootstrap analogues to these random sampling based inference techniques or direct permutation based randomization-inference (as we will show below, some strategies are very simple and just involve transforming the outcome).

4.1.1 Summary

In summary, we suggest that we can use covariates to increase the precision by which we assess causal effects. Here we began with a causal model where each subject has his or her own additive causal effect, but we focus attention on the sum of these effects. In large samples with outcomes that are not terribly skewed and where randomization ensures that covariates relate to controls more or less as they would to treated units, then we can calculate confidence intervals very quickly by relying on theory that suggests that (1) that the randomization distribution under the null hypothesis will be well approximated by a Normal distribution Hansen and Bowers (2009) and (2) that covariance adjustment where $\boldsymbol{\beta}$ varies will approximate covariance adjustment where $\boldsymbol{\beta}$ is fixed across randomizations. Moreover, because we include inspection of the operating characteristics of our intervals as a part of our model search process, we will be directly assessing the performance of these approximations along the way.

4.2 Simulation Studies of Known Models

Consider the following two outcomes shown in Figure 6. We created these outcomes using the Home Office data on age of respondents, household income, conditional on the strata that we chose for the design of the study. Specifically we generated the outcomes using the following steps:

1. Generate stratum specific means of the "hours helped others" outcome (where the 20 strata, recall, are UK administrative units and gender of the respondent).

2. The simulated outcome is the stratum specific mean outcome plus a linear function of respondent's age and household income and an error term. The Normal outcome has an error term drawn from a normal distribution with sd set to the sd of the "hours helped others" variable. Specifically, $y_i = \beta_{0j} + \beta_1 \text{Age}_i + \beta_2 \text{Income}_i + e_i$ where $e_i \sim N(0, \sigma^2)$ for the Normal outcome and $e_i \sim \pi_i \text{Geom}(.7) + (1 - \pi_i)\text{Geom}(.07)$ where $\pi_i \sim \text{Bernoulli}(.5)$ for the Skewed and Zero-Inflated outcome (which also collapses all negative values to zero). The linear model specified $\beta_1 = 16$ and $\beta_2 = -12$.

These simulated outcomes are useful because they are created using observed data, relate to observed covariates, and have the same number of observations as the real data, yet allow us a test of our procedures: if, for example, our procedure fails to control Type I error rates on the Normal-generated outcome, we would imagine that we have a coding error.

We assessed the realized size of these tests (their performance when the hypothesis is true) and their performance when the hypotheses false, in three different ways. First, we re-assigned treatment 10000 times, each time calculating

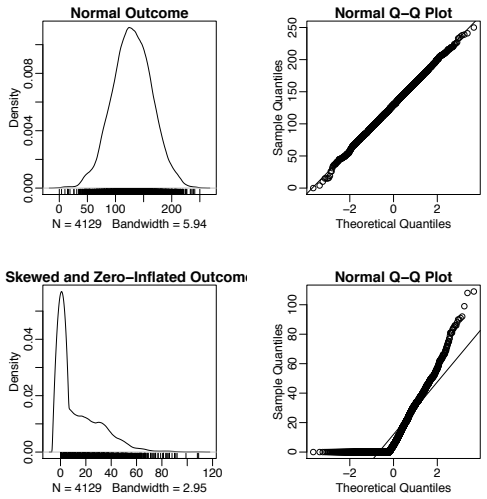

Figure 6: Distributions of simulated outcomes. Each outcome is non-negative integer and is a known linear additive function of respondent age and household income and twenty strata specific effects (where strata are defined by administrative unit and gender of the respondent). The R^2 of the known model for the Normal outcome and Skewed outcome are 0.37 and 0.58 respectively.

the confidence interval for A as described above, using the Normal approximation that arises from thinking about this problem as producing a confidence interval for the estimate of a total in a finite population. Figure **??** shows this result using black lines. The true A here is 0 (by virtue of breaking the relationship with outcomes by re-assigning the treatment). We see that rarely are p-values for $A_0 = 0$ less than .05 — i.e. they are less than .05 no more than 5% of the time (within simulation error of $2*\sqrt{p(1-p)/B} = 2*\sqrt{.05(1-.05)/10000}$ shown by the dashed lines). This is one sign of a well operating testing procedure. Another sign is that a well operating test should cast doubt on hypotheses as they diverge more and more from the truth. We see this also: at some point, hypotheses about A differ from 0 (the constructed truth) so much that they are so incongruent with the data that our procedure would always deem them implausible.

We also assessed the size and power of the tests when we apply the true model (the red lines in Figure **??**). We see that the size is still at or below .05 when $\alpha = .05$. And we also see that confidence intervals based on the true model will be more likely to reject hypotheses that are not true (where truth for this simulation study is $A = 0$) than the unadjusted confidence intervals.

However, we rarely know the true model by which covariates predict outcomes. To assess the performance of our method under more realistic conditions, we started the model selection algorithm with a model containing 81 terms — including Age and Income as multi-term natural cubic regression spline bases rather than as the true linear functions. Could we produce confidence intervals as tight as those arising from the true model even when we did

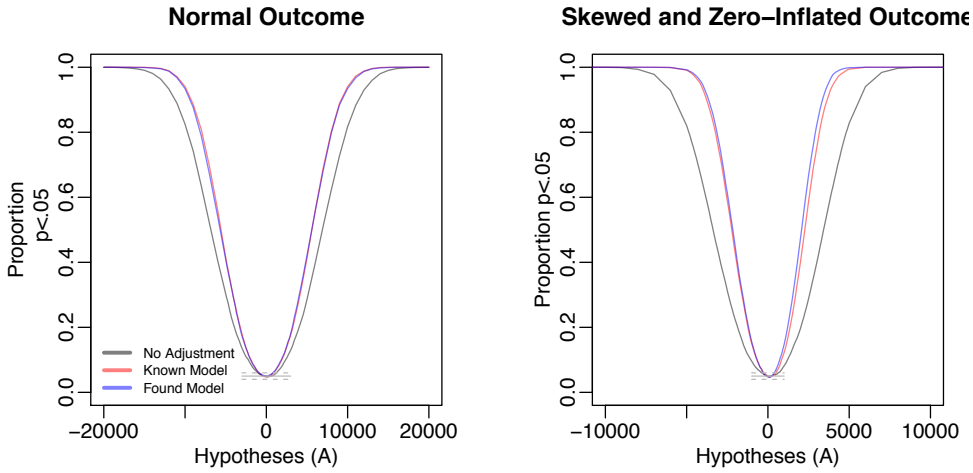

Figure 7: Power curves for $\alpha = .05$ for assessments of A_0. When $A_0 = 0$, no more than 5% of p-values are greater than .05 (gray horizontal line) (within ± 2 standard errors of simulation (horizontal gray dashed lines). Black lines show the power curve for the unadjusted tests, red lines show that power improves if one knows the true model relating covariates to outcomes, blue lines show the power curves arising from model selection starting with a model containing 81 terms.

not include the true model as a subset of the test model? The blue lines in Figure **??** show that the answer is yes.[10] We also see, again, that the coverage of our interval is correct.

4.2.1 Key Features and Details of the Algorithm

In simulated data we showed that our model selection procedure delivers a model that operates the same way that the true model operates. We used lexical optimization using a genetic optimizer (R's `genoud` function) and an adaptive elastic net procedure with four parameters (the elastic net mixing parameter for each of the two stages of the adaptive elastic net and the penalty parameters for each of those stages) (R's `glmnet` package). We should note that (1) we made no specific assumptions about the outcomes here [the integer valued zero-inflated and skewed-outcome, for example, is manifestly not-normal] and (2) that our large sample approximations allowed the computing to go quickly but are not central in any conceptual way to our statistical inference.

Other choices of machine learner may be faster or more convenient. The key is to have a relatively small number of tuning parameters for optimization: for each combination of values of tuning parameters, we assess the Type I error rate and power against some alternative hypothesis and the best set of tuning parameters maximizes power while keeping the coverage of the confidence in-

[10] We have not yet assessed the power of this procedure when we omit Age and Income entirely. We suspect it will be between the unadjusted model and the true model.

terval correct. It may not be necessary to use a formal optimizer, in fact, and a grid search may also be effective if the number of parameters is low or the fitting procedure itself is relatively inexpensive in computing time. Thus, one should not read this paper as advocating our particular adaptive elastic net machine learner plus lexical genetic optimization. Rather, this paper should make one ask what kinds of machine learners and/or search procedures would be best given the different applications that political scientists confront.

4.3 Real Outcomes

The procedure also works using real outcomes, however we were surprised at the lack of predictive power provided by the covariates.[11]

Figure 8 reminds us of the distribution of the outcome. We see that it is even more zero-inflated and skewed than the fake outcomes. Even though we have 4129 respondents in the study, it is possible that the Normal approximation may not successfully guarantee appropriate operating characteristics. This is another reason to execute the kinds of simulation studies that we used to assess the model selection algorithm above.

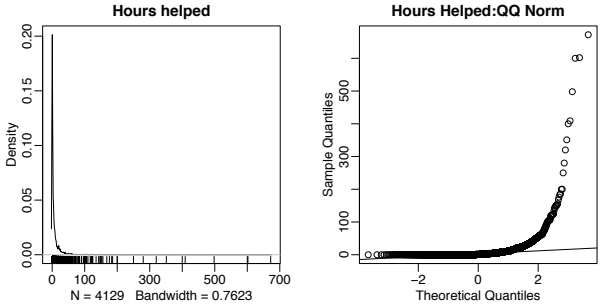

Figure 8: Distribution of hours spent helping others.

Figure 9 compares the coverage of the confidence intervals between unadjusted and the selected covariance adjusted confidence intervals for raw outcome (left panel) and two transformations. We see that the Normal approximation is not working well for the raw model. In this case the 95% confidence interval contains the true value of $A = 0$ (true for the purposes of this assessment of operating characteristics) 8.08% of the 10000 simulated experiments. Although most analysts with 4129 observations would imagine that a limit theorem would work in their favor, we see here that the outcome is just too skewed and zero-inflated. Analysts have long known that the easiest way to handle failures of Normal approximations is by transformation [cites to Tukey]. Since in this paper we are marrying techniques developed by Fisher and Neyman and survey samplers in the 1930s–1960s to machine learning of 2013, we

[11] We are working on expanding the list of covariates, including interactions, etc. . . because we basically have trouble believing that covariates like education should have such little power to predict civic activity.

tried transformation as our first attempt to handle this problem — knowing that we could resort to non-parametric methods like the bootstrap if these transformations failed.

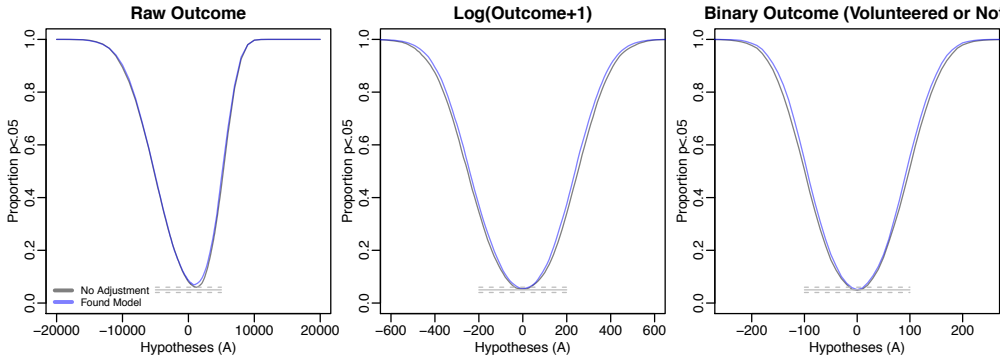

Figure 9: Real Outcomes: Power curves for $\alpha = .05$ for assessments of A_0. When $A_0 = 0$, no more than 5% of p-values are greater than .05 (gray horizontal line) (within ± 2 standard errors of simulation (horizontal gray dashed lines). Black lines show the power curve for the unadjusted tests, blue lines show the power curves arising from model selection using a set of 81 terms.

The middle and right panel in Figure 9 show the results of applying our algorithm to these transformations. We see that the approximations enable the intervals to have correct coverage (by the fact that the bottom of the curves is at the hypothesis of $A = 0$ and the value of the curve at that point is .05 or less (and within the standard error of simulation). We also see that the covariance-adjusted confidence intervals would be slightly narrower than the unadjusted intervals. This result surprised us. Yet, upon further inspection the result does not impugn the method. The list of 81 — including education, income, religion and religious activity, gender, spline bases of age and income among others — does not in fact strongly predict volunteering among people interviewed before the bombing. The R^2 from the models predicting the outcomes are 0.078,0.16,0.13 for the raw outcome, the logged outcome, and the binary outcome respectively. That is, although our procedure has the potential to substantially improve power when our list of covariates predicts the outcome, one should not expect large increases in power when we do not have strong relationships between covariates and outcomes.

4.3.1 Summary

So far we have limited our search for a covariance adjustment model to a relatively small number of terms — excluding, for example, interactions between terms. Most contemporary machine learning algorithms are designed to handle

situations where there are many more terms than observations (more columns than rows in $\mathbf{X}$). So, it is possible that when we expand our model matrix we will see an increase in power here. It is also possible that the responses to the civic voluntarism questions in the UK Home Office 2005 survey are not well predicted by any of the covariates (or any combination of the covariates) in that survey [although we would be surprised if this were the case]. In such a case, we would have lost some time to the search for a good covariance adjustment model, but, because we, as yet, would not have assessed any claims about treatment effects directly (using observed treatment and observed outcome together), we would not have raised any concerns about the validity or interpretability of our resulting statistical inferences.

5 Sensitivity Analysis

[This is a very drafty section and as such displays raw R code. We are basically writing about what we'd like to do. Overall, we'd like a flexible algorithmic approach to sensitivity analysis where the size of the strata and test statistic is not fixed.]

We have (1) adjusted our comparison in a way that (a) does not require us to look at treatment effects and cherry pick our favorite adjustment and (b) can relate to some reasonable standard by which we might judge an adjustment method (not the only standard, but a standard nonetheless); (2) choose a method of precision enhancement using machine learning techniques, also without ever assessing treatment effects, and (3) assessed hypotheses (via a confidence interval) about an effect (often called "causal" because it is an unobserved comparison of potential outcomes).

But, in the end, we did not randomize. Instead we re-organized our data and then proceeded as if we had randomized.

As-if-randomized story: For strata s, $\text{prob}(Z_{is} = 1) = m_s/n_s$ for all i in s where m_s is the number of treated units and n_s the number of control units in stratum s. All statistical inference in this mode requires this story: the story of a blocked randomized experiment.

General question: How strong would unobserved confounds have to be in order to change the substantive interpretation of the study? (Rosenbaum 2002, Chap 4) OR What if $\text{prob}(Z_{is} = 1) = \pi(m_s/n_s)$ (say, $\pi = 2$)?

To entertain this idea would be to change the distribution describing our hypotheses. So, our p-values and confidence intervals would change.

The Hope: Our confidence interval would not change by much! Or that π would have to be unrealistically large before our results would change!

New Questions: What are reasonable π to consider? Hosman, Hansen and Holland (2010) suggest that we calibrate our "what if" analysis using existing covariates. We might say that the unobserved variable driving the selection effect would have to be as powerful as the most predictive other variable in the data (for example). Rosenbaum tends to pick a range of π

(which he expresses in terms of odds rather than probabilities) and reports at which point the qualitative results of the study change (i.e. when the confidence interval covers zero; when the hypothesis test fails to reject the null of no effects).

Here is a brief demonstration of the reasoning behind Rosenbaum's approach to sensitivity analysis.

The as-if-randomized probability of treatment is:

```
load("../build/ho05.noNA.s.rda")
p <- with(ho05.noNA.s, tapply(z, strat3, mean))
p  ## for pairs p=.5
```

```
             North East               North West Yorkshire and the Humber
                 0.7458                            0.7242                   0.7261
          East Midlands            West Midlands          East of England
                 0.6778                            0.7191                   0.7688
                 London               South East               South West
                 0.7056                            0.6944                   0.6733
                  Wales
                 0.6803
```

But now, we presume that, in all sets, s, the actually treated were twice as likely to be treated. Recall that when $p_{is} = m_s/n_s$ for all $i \in s$ then odds=1 (i.e. everyone in the set has equal probability of being treated). Now, double the odds does not mean double the probability (since probability is bounded between 0 and 1). In fact, odds=p/(1-p). If all of the sets were pairs, then p=.5 for all sets. And we could just say 2=p/(1-p) and solve for p. Here we get p=2/3.

This next is some code to repeat the putative randomization while entertaining this kind of unequal probability of assignment:

```
## Very drafty code here!!

odds <- p/(1 - p)  ## calculate the odds of treatment
newodds <- 2 * odds  ## new doubled odds to reflect the What If thought experiment
newp <- newodds/(1 + newodds)  ## convert odds back to probabilities
newpt <- unsplit(newp, ho05.noNA.s$strat3)  ## add to data
newptc <- with(ho05.noNA.s, ifelse(z == 1, newpt, 1 - newpt))  ## control units get 1-newpt

strata.shuffle.sens <- function() {
    ## randomize z within strata with probabilities governed by the code above
    ## each unit has its own probability of receiving treatment assignment
    ## that depends on whether it is a treated or control unit and its stratum
    unsplit(sapply(split(ho05.noNA.s[, c("z", "newptc")], ho05.noNA.s$strat3),
        function(dat) {
            with(dat, sample(z, prob = newptc))
        }), ho05.noNA.s$strat3)
}

## testing to make sure always has fixed number of treated and controls in
## each set
```

```
## apply(sapply(split.data.frame(junk,ho05.noNA.s$strat3),colSums),2,unique)
```

Now, to depart from Rosenbaum and Hosman, Hansen and Holland (2010) we refer to our power analysis to alert us to the sensitivity of our results to changes in the process assigning individuals to be interviewed before versus after the bombing. That is, `strata.shuffle.sens` function can be plugged into the power analyses that we have been using to assess the operating characteristics of our confidence intervals but now centering the intervals on the center of the treatment effect interval (i.e. on $\hat{A}$). Say our interval did not include $A = 0$ and had high power against that hypothesis. Now imagine an unobserved confound that makes treatment assignment twice as likely among the treated. If the power to reject $A = 0$ becomes much lower (i.e. imagine that the new power curve flattens) then we might say that our interval is sensitive to confounds that double the odds of treatment. If, however, the power curves do not diverge, then we might say that our interval is not sensitive to such a potential confound.

[Note: We have not executed this analysis yet because it is dependent on us evaluating the confidence interval for an actual treatment effect. We also would love comments about this idea for a new form of sensitivity analysis based on power analysis since we have not seen such a proposal elsewhere in the literature.]

6 Discussion and Conclusion

The standard approach to assessing treatment effects in randomized (or as-if randomized) comparative studies is to simply regress outcomes on treatment assignment indicators. A growing body of work is developing ways to produce confidence intervals for the average treatment effect while also taking advantage of background information to enhance precision (Freedman, 2008; Green, 2009; Schochet, 2009; Lin, 2011; Miratrix, Sekhon and Yu, 2012). In this paper we have engaged with one of the questions raised by both the standard practice and the latest methodological literature on the ATE: how should we choose a covariance adjustment specification? Our answer is that we should target exactly the characteristics of the statistical inference that we desire to improve: namely the precision of the inferences. And along the way we discovered that one could do this without also prejudicing the validity of the tests themselves — i.e. their ability to react correctly and in a controlled manner to the truth.

We did this work by taking advantage first and foremost of the ability to separate statistical inference from other analysis work that is often delegated to the linear model (i.e. we took interval creation and improving a design/balance assessment away from the linear model). And secondly we took advantage of the growing literature on machine learning. Because we choose tuning parameters for machine learners with size and power of tests as

objective functions, it is easy to inspect the operating characteristics of our tests for our final, chosen model. And this in turn helps us allay worries about approximations that are convenient and speedy, but which are not in any fundamental way, required. In our particular application, we were able to rely on the approximations throughout, although, if we had pursued the newspapers study that we used to demonstrate the modularity of Fisher's randomization inference, we would have certainly have needed to use other approaches (such as directly assessing hypotheses about collections of τ_i leading to A).

We followed our model selection and statistical inference with a sensitivity analysis — and this sensitivity analysis is novel for continuing to focus on the power curves rather than the confidence intervals themselves.

Although advance registration of covariance adjustment strategies is ideal for ensuring transparency and the validity of statistical inferences, such strategies in simple form may leave information on the table. We speculate that if an analyst registered a study, declaring a long list of covariates to be entered as ingredients into an automatic process, such as the one we develop here, both validity and power might be achieved without loss of transparency [since the algorithms and code are all open source and thus scrutiny and reproduction would be easy by any suitably skilled critic]. Finally, we used the London Bombings study to highlight the use of these ideas when we do not register a strategy before data collection, but when data has already been collected; and when, in fact, treatment has not been randomized. Further, we used this study to show that even very strange outcomes (which might otherwise force analysts to worry about varieties of zero-inflated count models and their associated asymptotic properties and functional forms) can be handled in this framework because we make no assumptions about the nature of the outcome in general — and weak assumptions about the outcome when we want to take advantage of speedy Normal approximations.

Finally and most generally, by breaking the process of drawing statistical inferences about causal effects into parts, we can ask questions about standards for each of the tasks that we set ourselves. The adequacy of a linear model aiming to "control for", to enhance precision via covariance adjustment, and to produce confidence intervals for treatment effects, is difficult to debate as a whole. For example, answers to questions about the adequacy of the randomization mechanism (or generated design of a quasi-experiment like the London bombings example) ought to face different standards than answers to questions about confidence intervals about treatment effects. Modules enable standards. The nature of the modules and standards can be debated, but such debate can be focused and grounds for decisions can be agreed upon even in advance of seeing results. Thus, social scientists can focus more on articulating the grounds for persuasive social scientific evidence (i.e. on the aspects of design and comparison which ought to persuade) and less on memorizing abstract assumptions about model fitting.

References

BBC News. 2008. "Special Reports: London Explosions.".
URL: *http://news.bbc.co.uk/2/shared/spl/hi/uk/05/london_blasts/what_happened/html/*

Bowers, Jake. 2011. Making Effects Manifest in Randomized Experiments. In *Cambridge Handbook of Experimental Political Science*, ed. James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupia. New York, NY: Cambridge University Press chapter 32.

Bowers, Jake, Mark Fredrickson and Ben Hansen. 2009. *RItools: Randomization Inference Tools*. R package version 0.1-6.
URL: *http://www.jakebowers.org/RItools.html*

Bowers, Jake, Mark M. Fredrickson and Costas Panagopoulos. 2012. "Reasoning about Interefernce Between Units: A General Framework." *Political Analysis* 21(1):97 – 124.

Clarke, Kevin and David Primo. 2012. *A Model Discipline*. Oxford University Press.

Cox, DR and P. McCullagh. 1982. "Some aspects of analysis of covariance (with discussion)." *Biometrics* 38:541–561.

Fisher, R.A. 1935. *The design of experiments. 1935*. Edinburgh: Oliver and Boyd.

Freedman, David A. 2008. "On regression adjustments to experimental data." *Advances in Applied Mathematics* 40(2):180–193.

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

Hansen, B.B. 2008. "Comment: The essential role of balance tests in propensity-matched observational studies." *Statistics in Medicine* 27(12).

Hansen, B.B. and J. Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23:219.
URL: *doi:10.1214/08-STS254*

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(487):873—885.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd ed. Springer-Verlag.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

Hosman, Carrie A, Ben B. Hansen and Paul W. Holland. 2010. "The sensitivity of linear regression coefficients's confidence limits to the omission of a confounder." *The Annals of Applied Statistics* 4(2):849–870.
URL: *http://arxiv.org/abs/0905.3463*

Keele, Luke, Corrine McConnaughy and Ismail White. 2010. "Adjusting Experimental Data: Models versus Design." Unpublished manuscript.

Kinder, D.R. and T.R. Palfrey. 1993. "On behalf of an experimental political science." *Experimental foundations of political science* pp. 1–39.

Leisch, Friedrich. 2002. Dynamic generation of statistical reports using literate data analysis. In *Compstat 2002 - Proceedings in Computational Statistics*, ed. W. Haerdle and B. Roenz. Heidelberg, Germany: Physika Verlag pp. 575–580.

Leisch, Friedrich. 2005. *Sweave User Manual.*
URL: *http://www.ci.tuwien.ac.at/ leisch/Sweave*

Lin, Winston. 2011. "Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique." Unpublished manuscript.

Lohr, S. 1999. *Sampling: Design and Analysis.* Brooks/Cole.

Miratrix, L.W., J.S. Sekhon and B. Yu. 2012. "Adjusting treatment effect estimates by post-stratification in randomized experiments." *JR Stat. Soc. Ser. B. Stat. Methodol. To appear* .

Office, UK Home. 2005. "Home Office Citizenship Survey, 2005." UK Data Archive Study Number 5367.

Panagopoulos, Costas. 2006. "The Impact of Newspaper Advertising on Voter Turnout: Evidence from a Field Experiment." Paper presented at the MPSA 2006.

Rosenbaum, Paul. 2009. "Design of Observational Studies." Unpublished book manuscript.

Rosenbaum, Paul R. 2001. "Effects Attributable to Treatment: Inference in Experiments and Observational Studies with a Discrete Pivot." *Biometrika* 88(1):219–231.

Rosenbaum, Paul R. 2002*a*. "Covariance adjustment in randomized experiments and observational studies." *Statistical Science* 17(3):286–327.

Rosenbaum, PR. 2002*b*. "Attributing effects to treatment in matched observational studies." *Journal of the American Statistical Association* 97(457):183–192.

Rubin, D.B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66(5):688–701.

Särndal, C.E. and B. Swensson. 2003. *Model Assisted Survey Sampling*. Springer.

Schochet, Peter. 2009. "Is regression adjustment supported by the Neyman model for causal inference." *Journal of Statistical Planning and Inference* .

Tibshirani, R. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Zou, H. 2006. "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association* 101(476):1418–1429.

Zou, H. and T. Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.