Machine Learning and Causal Inference: A Modular Approach to Assessing the Effects of the London Bombings of 2005

Jake Bowers [*] Mark M. Fredrickson [†] Ben B. Hansen [‡]
Costas Panagopoulos [§]

August 30, 2015

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, 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 combine modern methods of machine learning with Fisher's conceptual framework and survey sampling based design-based statistical inference originating with Neyman in order to maximize power without compromising the integrity of the resulting statistical inference. We apply our ideas in the context of a natural experiment created by the London subway bombings of 2005.

[*] Associate Professor, Dept of Political Science and Statistics, University of Illinois @ Urbana-Champaign (`jwbowers@illinois.edu`). *Acknowledgements:*Thanks to Chris Achen, Joe Bowers, Dan Carpenter, Wendy Tam Cho, Tommy Engstrom, Don Green, Danny Hidalgo, 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. Version: fa5ac68

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

[‡] Associate Professor, Dept of Statistics, University of Michigan

[§] Associate Professor, Dept of Political Science, Fordham University

> "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)

Baseline information on experimental subjects promises to increase the precision of statistical inferences about causal effects. At least since Fisher (1935) and Cox (1958) scholars have known that information measured pre-intervention, also known as covariates, can help shrink confidence intervals and p-values. Yet, concerns about multiple testing and test validity have limited the ability of covariate adjustment to fulfill its promise. This paper proposes a method that enables the precision gains of covariance adjustment without multiple testing while maintaining the validity of randomization-based statistical inference.

A covariate adjusted confidence interval will tend to be smaller than the unadjusted version. The common practice to produce covariance adjusted confidence intervals using a covariate x_i in an experiment with a binary treatment $Z_i \in \{0 = \text{control}, 1 = \text{treated}\}$ and an outcome Y_i involves using the coefficient $\hat{\beta}_1$ from a linear model like $Y_i = \beta_0 + \beta_1 Z_i + \beta_2 x_i$.[1] Without X_i, a confidence interval for the $\hat{\beta}_1$ coefficient is the confidence interval for the difference of means of Y_i between the treated and control groups. With x_i in the model, this confidence interval still refers to the difference of means of Y_i but Y_i with variation due to x_i removed. Since x_i is a covariate (i.e. cannot have been influenced by the experimental treatment), in a large experiment, the difference of means between treated and control groups should not change, but the confidence interval should shrink because the extraneous variation in Y_i unrelated to the experiment (but related to x_i) should have diminished by the adjustment process.[2] If, however, analyst tries many different adjustment specifications (say, adding and dropping different covariates, or trying different functional forms), then the reported confidence interval may no longer be valid: an interval claimed to falsely exclude the truth no more than 5% of the time, may do so more often. Thus, skeptical readers may view the report of a statistically significant result from an experiment that is only seen after covariance adjustment with suspicion: one might ask, for example, in how many previous analyses was the result insignificant?[3]

To counter such concerns, one may declare a covariance adjustment strategy in ad-

[1] In this paper a lowercase letter is a fixed quantity and uppercase for random quantities (for example, x_i is unaffected by the randomized experimental treatment, Z_i, and so is fixed).

[2] To see how covariance adjustment removes extraneous variation from Y_i recall that we can estimate $\hat{\beta}_1$ in three steps: (1) produce the residuals from a regression of Y_i on x_i, $e_{i,Y|x}$ (these x-residualized versions of Y represent Y_i with no linear relationship with x_i), (2) produce residuals from a regression of Z_i on x_i, $e_{i,Z|x}$ (these x-residualized versions of Z represent a Z_i with no linear relationship with x_i — which, in a large experiment, should be basically the same as Z_i itself given the randomization of Z_i), and (3) regress $e_{i,Y|x}$ on $e_{i,Z|x}$ to produce $\hat{\beta}_1$. Since the x-residualized version of Y_i will have lower variance than Y_i itself, standard error on $\hat{\beta}_1$ will be lower than it would be in the unadjusted case.

[3] This process of trying out different covariance adjustment strategies is not a process of wrongdoing, by the way, as Gelman and Loken (2013) wisely note, the problem involves tailoring the adjustment to the data in some non-pre-specified manner.

vance: the error rate of statistical tests will be controlled if the analyst follows the pre-determined plans.[4] Yet, even if an analysis plan is announced in advance (to counter concerns about multiple testing), one may ask whether the precision enhancement promise of a given covariance adjustment strategy will be fulfilled: a claim made before inspecting covariate to outcome relationships may leave statistical power on the table. Even if x_i and Y_i relate linearly and strongly in past experiments, they may relate non-linearly, or not at all, in any given experiment. In fact study dependent relationships between outcomes and background information will determine, in part, the width of the confidence intervals that one will produce with or without adjustment. So, then, how can one a produce a study- or data-dependent mode of covariance adjustment without running the risks of what some have called ‘p-hacking’ or ‘fishing’?

This paper proposes such a method. Rather than declare a specific covariance adjustment model specification before seeing the data, we demonstrate a method that requires only pre-specification of the ingredients (i.e. the names of the covariates) and the algorithm for covariance specification search. Recent advances in procedures for ’machine learning’ or ’statistical learning’ allow very flexible approaches to model specification search. We take advantage of those advances, but because we adopt a modularized approach to the analysis of experimental data here, we can take advantage of the promise of covariance adjustment using machine learning techniques while avoiding multiple testing, and which has the same validity of statistical testing as any randomization inference based analysis of experimental data. Basically, we propose to delegate the choice of covariance adjustment strategy to a machine learner, but to separate covariance adjustment from confidence interval production so that no multiple testing occurs. We define the targets of our causal inference at the individual level following Fisher (1935) but speed the process of randomization-based statistical inference with an approach inspired by Neyman (1923 [1990]) and developed in Hansen and Bowers (2009).[5] We demonstrate this modularized approach with a natural experiment of the effect of terrorist attacks on social capital constructed by comparing the responses of roughly 1200 2005 UK Home Office Survey respondents before versus 3300 after the July 2005 London bombings. The responses in this survey are counts but our statistical inferences do not depend on assumptions about outcome distributions. At the end of the paper, we show that our approach never makes statistical inferences less precise, and that gains of around

[4] Humphreys, de la Sierra and Van der Windt (2013) articulate the fishing problem and advocate pre-analysis plans. Rosenbaum (2008) and Hansen and Sales (2015) provide some examples of how pre-specified plans for hypothesis testing control error-rates even when multiple hypothesis tests are conducted.

[5] Our approach is explicitly modular — allowing us to separate the task of choosing a covariance adjustment model from the task of statistical inference about causal effects. In this way, we are inspired by Rosenbaum (2002*b*) although our approach is more like the Peters-Belson approach to covariance adjustment (Peters, 1941; Belson, 1956). Our use of machine learning links us with Bloniarz et al. (2015) and Van der Laan and Rose (2011) although we differ from those two different approaches to using machine learning in causal inference by our Fisherian/individual level targets of causal inference and our statistical inference procedure.

50% in precision are possible in our application and simulations based on the UK Home Office Survey+2005 Bombings Data.

1.1 Why worry?

Imagine an experimenter that reports a statistically significant results for a treatment effect using a linear regression model including the outcome, the treatment, and age. The experimenter argues, correctly, that age is independent of treatment assessment (in expectation) and, since age predicts the outcome, including age in the model will merely help the treatment "score a significant success" when it might otherwise be a valid but noisy (i.e. not statistically significant) result. The risks run by this experimenter include: (1) **The Suspicion of Snooping.** Some could wonder whether age was chosen after a hunt. Did the experimenter try 100 different covariance adjustment strategies, each time inspecting the p-value on the treatment effect, and only stopped hunting when he found a $p < 0.05$? If so, we cannot interpret the p-value as a clear measure of information against the null of no effects. And post-hoc adjustments for multiple testing require knowledge of the amount of and pattern of multiple testing (which may not be available if the hunt was not disciplined) (2) **Concern and Controversy about Test Validity** if the experiment did not randomly assign treatment and control to two equal sized groups or the experiment is small, some might point out that conventional covariance adjustment (i.e. adding a covariate like a control variable to a linear model) produces biased estimates of treatment effects (and biased statistical tests) (Freedman, 2008; Lin et al., 2013). Others might argue that these biases should be very small and likely not worrisome (Green, 2009; Schochet, 2010). Which perspective would be true for a given experiment would depend on the details of the experiment and would require extra work; (3) **Power left on the table** Even if the analyst declared a covariance adjustment strategy in advance, followed Lin et al. (2013)'s advice in regards the specification of the covariance adjustment in the linear model and the asymptotically valid standard error and showed simulation evidence that their study was close to asymptopia, we might wonder whether the given confidence intervals could have been shrunk more if we had known that, in the given dataset, age had a nonlinear relationship with the outcome in the control group and that shoe-size also turned out to account for some of the non-treatment-effect related variance in the response.

This paper contributes to the theory and practice of the statistical analysis of experiments with a modular approach for assessing the causal effects of interventions. 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 modern machine learning in predictive models which represent past scholarly knowledge about the outcome. The validity of the statistical inferences from this approach is not contingent on the veracity of any given model except for the model of treatment assignment of the experiment itself. This paper also contributes to the literature on design-based statistical inference with the application: we produce confidence intervals for a diverse set of count-variables without requiring any particular probability model of outcomes.

2 The London Bombings of July 2005: Count Outcomes and Natural Experiment

During the morning commute on July 7th, 2005, four suicide bombers placed explosives on the London public transport system killing 52 people and injuring over 700 (BBC News, 2008). The bombers claimed to be soldiers in a battle between "the West" and "Islam." We know that terrorist attacks aim to disrupt the targeted civil society. We demonstrate the method developed here by asking 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. Although the bombing occurred unexpectedly in the middle of the survey field work, the survey interviews did move around the country systematically such that relatively more interviews happened after the bombing in London when relatively few were happening in Wales. To bolster the claim that we are analyzing a natural experiment we first stratified the data by the ten governmental regions which had organized the survey sample and fieldwork. We then restrict attention to a window of 2 weeks before the bombing to 9 after the bombing. And, finally, we decided that we would only compare men to men and women to women (because we figured that reactions to the bombing as manifested in answers to questions about social capital might vary between those groups). Our final design compared favorably with an equivalent block-randomized experiment: The Hansen and Bowers (2008) d^2 omnibus balance test produced a p =0.87 against the null hypothesis of simultaneous balance on 140 covariate terms. In the end, and as we explain in detail in appendix Appendix A, we ended up with about 4500 subjects in the research design.

Terrorism attacks seek to disrupt civil society. If these attacks were successful, respondents interviewed after the bombing should report less social cohesion. Yet the attacks may impel citizens to rally and may provide a very salient reason for public action. The 2005 UK Home Office Survey asked a series of questions to gauge neighborhood cohesion such as "People in this neighbourhood do not share the same values? (Agree/Disagree)", community efficacy such as "If some children were spray-painting graffiti on a local building, how likely is it that people in your neighbourhood would do something about it?", and trust in institutions such as "How much do you trust the police?" To combine the responses to the different items measuring the same concept, we coded responses in the top categories as "positive social capital." Figure 7 in Appendix B shows the distribution of the positive responses along with the full text of these questions. If the bombing had not occurred, would social cohesion and trust have been different? Did the bombing have an effect on social capital?

3 Attributing Effects to Treatment for Count Outcomes

If the bombing had an effect on a person i, then the number of positive responses to questions asked before the bombing, $y_{i,Z_i=0}$ would differ from the number of positive responses to questions asked after the bombing $y_{i,Z_i=1}$, where Z_i records the timing of the survey interview (0=before the bombing and 1=after the bombing). We might write $y_{i,Z_i=1} = y_{i,Z_i=0} + \tau_i$ to represent a theoretical expectation that each person experienced a different effect, τ_i, from the bombing (and that each effect was additive (or subtractive)). For situations with one parameter, like the constant effects model $y_{i,Z_i=1} = y_{i,Z_i=0} + \tau$, Rosenbaum (2010, Chap 2) explains how confidence intervals for τ could be created. With varying τ_i, one could, in principle, produce N-dimensional confidence sets. For example, 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 $\{\tau_1 = 1, \tau_2 = 0, \tau_3 = 0, \ldots, \tau_N = 1\}$ but plausible that $\{\tau_1 = 0, \tau_2 = 0, \tau_3 = 1, \ldots, \tau_N = 0\}$ tends not to address a scientific question about overall effects. To simplify the problem Rosenbaum (2001, 2002*c*) proposes a function of the τ_i as an object of scientific interest when outcomes are binary, writing $A = \sum_{i=1}^{n} Z_i \tau_i$ 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). In theory, we could develop a confidence interval for hypothesized A_0 with the following algorithm: For a given hypothesized A_0 (say, $H_0 : A_0 = 1$), we list all k of the ways that a vector of $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)$ can be summed to equal A_0.[6] For each possible vector, generate a p-value 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. By repeating this procedure for many hypotheses A_0, we can define a confidence set as the collection of hypotheses not rejected at a given level. 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.

[6] 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.

How can we use what we observe to learn about the number of positive responses that we would not have seen were it not for the bombing without millions of hypothesis tests?[7] When the research design includes many strata or pairs, Rosenbaum (2002*a*) and Rosenbaum (2010, §2.5) develop an approach to this problem in the binary outcome case which identifies, in advance, which two hypothesis tests would provide the most and least evidence against any given null, if both tests have small p-values we reject the composite null, if either has a large p-value, we cannot reject the composite null. With a small number of strata and a binary outcome, Hansen and Bowers (2009) showed that we can think of A as a finite population total such that one can produce approximate confidence intervals for A by estimating the total $y_{i,0}$ under the idea that the observed outcome among controls is a random sample from the total. This idea extends naturally to count outcomes: given an individual level model $\tau_i \equiv y_{i,1} - y_{i,0}$ where $y \geq 0$ and $y_{i,1} \geq y_{i,0}$, we define A for the whole experimental pool in the set U as the sum of the positive differences **Generalized attributable effect:** $A \equiv \sum_{i \in U} Z_i \tau_i$.

$$A = \sum_{i \in U} Z_i \tau_i = \sum_{i=1}^{N} Z_i (y_{i,1} - y_{i,0}) \tag{1}$$

C is the set of control group observations (pre-bombing respondents).

$$= \sum_{i \notin C} y_{i,1} - \sum_{i \notin C} y_{i,0} \tag{2}$$

because $\sum_{i \in C} Y_i - y_{i,0} = 0$. we can write

$$= \sum_{i \in U} Y_i - \sum_{i \in U} y_{i,0} \equiv t_U - t_C \tag{3}$$

$$= \underset{\text{fixed and observed}}{\text{observed total overall}} - \underset{\text{unobserved, to estimate}}{\text{total outcome under control}} \tag{4}$$

The attributable effect, then, combines the fixed total, $t_U = \sum_i Y_i$ (the total number of positive responses 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 positive responses the survey respondents would have volunteered had they been interviewed before the bombing. Causal and statistical inference about A then is equivalent to inference about the partially observed total in the control group $t_C = \sum_i y_{i,0}$ because that is the only quantity that might vary with the design: selecting different control group members from the experimental pool will yield different t_C whereas the total in the experimental pool, t_U remains fixed regardless of who is chosen to be a control.

[7] We speak in terms of positive responses caused by the bombing here, but later allow for there to be fewer such responses after the bombing than before it — i.e. allowing for a negative effect of the bombing.

We know from the survey sampling literature that an unbiased estimator of t_C is $\hat{t}_C = N\bar{Y}_C$ (Lohr, 2001) and in large samples under regularity conditions allowing a central limit theorem to operate an approximate confidence interval would be: $\text{CI}(t_C) = \hat{t}_C \pm z_{\alpha/2}\text{SE}(\hat{t}_c)$. Returning to A, using the decomposition above, we can write a confidence interval as: $\text{CI}(A) = t_U - \widehat{\text{CI}}(t_C)$. This conceptualization of A was used by Hansen and Bowers (2009) to create an large-sample approximate confidence interval for binary outcomes.

Now, the same literature which established that $\hat{t}_C = N\bar{Y}_C$ is an unbiased estimator of t_C, also developed the "regression estimator of the finite population total" such that $\hat{t}_C = \sum_{i \in U} \hat{Y}_i + \sum_{i \in C}(Y_i - \hat{Y}_i)$ where we estimate (1) $\hat{\boldsymbol{\beta}}$ from a 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, for example, using a linear model, $\hat{Y}_i = \mathbf{X}\hat{\boldsymbol{\beta}}$ (See (Lohr, 2001, Chap 4) and Särndal and Swensson (2003)). 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. In the simple unstratified case for n control units and N total units in the experimental pool, the standard error of the regression adjusted estimator is: $SE(\hat{t}_C) = N\sqrt{(1 - \frac{n}{N})s_e^2/n}$ here $s_e^2 = \sum_{i \in C} e_i^2/(n-2)$ and $e_i = Y_i - \hat{Y}_i$. The standard error of the unadjusted estimator is $SE(\hat{t}_C) = N\sqrt{(1 - \frac{n}{N})s_Y^2/n}$ where $s_Y^2 = \sum_{i \in C}\left(Y_i - \bar{Y}\right)^2/(n-1)$. If our model of $Y_{i \in C}$ predicts it well, then s_e^2 will be smaller than s_Y^2 and the standard error of $\hat{t}_C$ will be smaller.[8]

4 How should we choose a covariance adjustment specification?

Covariance adjustment requires covariates, $\mathbf{X} = \{x_1, \ldots, x_p\}$ and fit $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}$ that predicts the outcome in the control group very well (such that s_e^2 is small). Although the regression estimator is a well established way to use auxiliary information to improve estimates of finite population totals from sample information and we can reconceptualize this process to enable us to learn about a control group from an experimental pool, 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

[8] The regression estimator is not unbiased, but, as Lohr (2001, §4.3) explains, in a simple regression with one single covariate, x, the bias is proportional to $-cov(\beta_1, \bar{x})$ and tends to be small in large samples. In our application, we guard against the small effects of this bias during the model selection stage.

resistant least squares?)? Luckily, the rough procedure of restricting model fitting to the control group allows analysts to use the data to answer such questions without calculating treatment effects: covariance adjustment in the context of randomization inference for causal effects separates adjustment from assessment of treatment effects.

In this paper we demonstrate the use of one of the more well-established machine learning techniques to answer some of the questions raised in the previous paragraph. Our answer to 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 a most powerful test. 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 known as the elastic-net model (Zou and Hastie, 2005). Any other producer of $\hat{Y}_i$ as conditional means of Y_i based on inputs of $\mathbf{X}$ would also work.[9] We choose the elastic-net model here because it is faster than a random forest or other more iterative methods of machine learning. It also has a form that should not be too strange for a social science audience: we choose a $\hat{\boldsymbol{\beta}}$ which minimizes the least squares criterion plus the elastic-net penalty:

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

This criterion involves both the lasso or L_1 penalty, $|\beta_j|$, (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).[10] 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.[11]

A key feature of penalized linear models that helps us in the task of choosing a covariance adjustment model is that the process of choosing a model (i.e. choosing covariates and $\boldsymbol{\beta}$) can be reduced to a choice of one or two tuning parameters. In

[9] With skewed outcome with long tails, it would be tempting to use a lasso penalized quantile regression model. This won't work here because we are estimating a sum or total using conditional means and the relationship between means and conditional quantiles is not straightforward. An alternative approach would be to use a Huberized or M-estimator for the least squares part of the model.

[10] Although we began this project using the adaptive elastic net because of its oracle properties, it provided no additional benefits over the simple version in our specific application.

[11] See Efron (2010) or Hastie et al. (2005), for example, on why penalized models seem to work so well compared to unpenalized models.

this case, λ and α together determine a model — if λ is large, then the model will have many zero coefficients (i.e. the model will exclude many covariates or covariate terms). If λ is small, then nearly every covariate will get a little weight in the final prediction of $\hat{t}_C$. Since machine learning is not yet common in political science, figure 1 shows an example of how tuning parameter choice amounts to model choice. Here, we specified a simple model whereby number of positive community efficacy responses 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. As the penalty parameter λ goes from approx 0 or $log(\lambda)$=-6.48 to approx 0.45 or $log(\lambda)$=-0.81, the sizes of the coefficients decrease to zero. For the largest λ only immigrant status remains as a strong predictor.

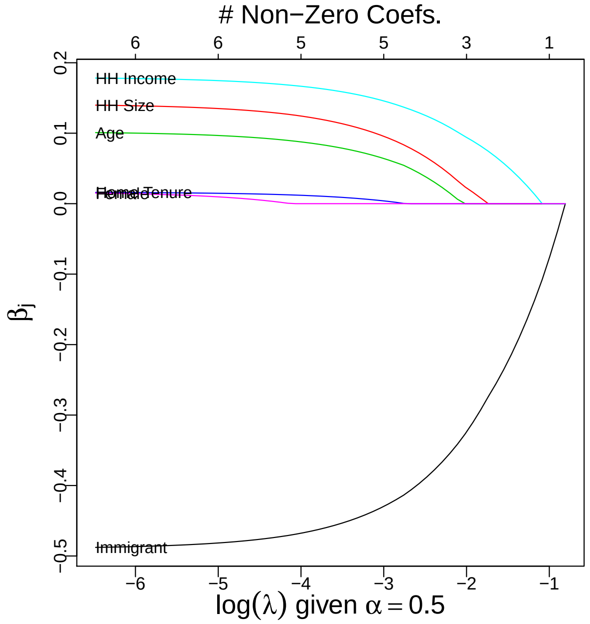

Figure 1: Example of the influence of choice of λ on model choice given $\alpha = .5$. As the penalty parameter λ goes from approx 0 or $log(\lambda)$=-6.48 to approx 0.45 or $log(\lambda)$=-0.81, the sizes of the coefficients predicting positive community efficacy responses among control group respondents decrease to zero.

Each vertical slice of Figure 1 represents one model. And, one can imagine many other such figures for different α weights. So, this is one of our proposals: we can side step the problems of covariance adjustment specification choice by delegating the problem to an algorithm. The analyst is still responsible for the ingredients — one still has to exercise some judgement about which covariates to include and how to include them (and what counts as a covariate and not a post-treatment variable). Yet, we can reconceptualize and simplify covariance adjustment specification choice as tuning parameter choice within a machine learning model, then we can take advantage of the gains in prediction from those types of models.

4.1 How do choose tuning parameters for covariance adjustment?

How should we choose among possible models now that we have simplified model choice from all models with all combinations of p covariates to only 2 tuning parameters? 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: the false rejection rate of a statistical test with a process that has no variation would be either incredibly high or low and not controlled.[12] In addition, an excellent fit in the control group would extrapolate very poorly to the treatment group, and this would lead to poor power in casual effect assessment.[13] 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 mean-squared error (like the AIC or BIC), in this paper we choose tuning parameters based on the power and error rate of the resulting confidence intervals. Although we began with the idea that we could simply plug-in the procedures common in the machine learning literature, our experiments taught us that the tuning parameters which are optimal from the perspective of cross-validated mean-squared error are not those that both enhance power and maintain a controlled false rejection rate of statistical tests. To account for the extra variability that arises from the model search, we follow Faraway (1992)'s general idea to replace t-quantiles with nonparametric bootstrap-t quantiles, $t^*_{\alpha/2}$ such that:

$$CI(t_C) = \hat{t}_C \pm t^*_{\alpha/2} SE(\hat{t}_C) eq:theci)$$

Here is the procedure that we used:

4.1.1 Ingredients:

- **strata:** a vector containing strata indicators. all operations are conditional on these fixed sets.
- **pseudodata:** A dataset the same size as the original but with only control units sampled with replacement. The point is to make a dataset with the same number of rows as the original but containing only information from the controls. We sample controls with replacement within stratum so that the stratum sizes are the same.
- **treated:** a vector with the same treatment design as the original study assigned to the pseudodata. So, a set with 25% controls in the original data will assign 25% of the members of that set in the pseudodata to control.

[12] Hansen and Bowers (2009, § 3.3)) noted that an overfit covariance adjustment model might lead to overly optimistic confidence intervals in the context of specifying covariance adjustment models by hand.

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

- **yhatmat:** A matrix with each column containing a $\hat{y}_C$ vector arising from an elastic-net fit to the pseudodata using covariates with strata weights for each unit. Each column represents $\hat{y}_C$ for different set of tuning parameters.
- **response:** a vector containing the outcomes measured for the pseudodata.

4.1.2 Choose a powerful covariance adjustment model.

1. For each column of yhatmat estimate the total hours that would be volunteered in the pseudocontrol condition: $\hat{t}_C = \sum_{i \in U} \hat{y}_{i0} + \sum_{i \in C} (y_{i0} - \hat{y}_{i0})/\pi_i$ where $\hat{y}_{i0} = f(\mathbf{X}_i, \boldsymbol{\beta})$, $\mathbf{X}$ contains the covariates and $\boldsymbol{\beta}$ the fitted coefficients from one elastic-net model fit to the controls within the pseudodata, U is the set of the whole pseudodata, C is the set of psuedocontrols within the pseudodata, and π_i is the proportion of i's in that control group in i's strata. $\widehat{SE}(\hat{t}_C)$ is defined in the next step.
2. For a given stratum, s of size n_s with m_s in the control group, $SE_s(\hat{t}_C) = n_s^2(1 - m_s/n_s)(s_C^2(\hat{y}_{i0}) + \text{AICadj})/m_s$ where the variance of the $\hat{y}_{i0}$ among the controls is calculated as $s_C^2(x) = \sum_{i \in C,s} (x_s - \bar{x_s})^2/(m_s - 1)$. The AICadj term follows Hastie, Tibshirani and Friedman (2009, §7.26) 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 of 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 in the $\boldsymbol{\beta}$ vector above), and σ_l^2 is the mean-squared error from a low bias model (i.e. a very saturated and unpenalized model). The overall approximate standard error is the simple sum across sets as we would use in a block-randomized experiment or stratified finite-population sampling plan — $\widehat{SE}(\hat{t}_C) = \sum_{s=1}^{S} SE_s(\hat{t}_{C,s})$.
3. To further penalize the model search, we multiple this standard error by the quantile of a bootstrapped distribution of t-statistics (i.e. Produce the studentized bootstrap distribution of $\hat{t}_C$.) Here is that procedure:
   (a) Draw a bootstrap sample from each stratum.
   (b) repeat B times to get B versions of the total and SE for each model.
   (c) Form a studentized statistic for each iteration, b, of the bootstrap: $z_b = (\hat{t}_{C,b} - \hat{t}_C)/\widehat{SE}(\hat{t}_{C,b})$. The distribution of the z_b is the reference distribution for the test. The quantiles of this distribution, say, $z^*_{\alpha/2}$ for the value of z at the $\alpha/2$ quantile, provides more accurate approximations to the t-distribution after model search than appealing to the t-distribution directly.

   (d) A bootstrap-t or studentized bootstrap $100(1-\alpha)$ % CI for $\hat{t}_C$ has a lower bound of $\hat{t}_C - z^*_{1-\alpha/2}\widehat{SE}(\hat{t}_C$ and an upper bound of $\hat{t}_C - z^*_{\alpha/2}\widehat{SE}(\hat{t}_C$. That is, the bootstrap procedure is only used to calculate the quantiles of the reference distribution, not the standard error (which we already know from the survey sampling theory of estimators of finite population totals).

4. Assess power. This process would yield as many CIs as models in yhatmat. We use the cdf of each bootstrap distribution to calculate power analogously to the way one would use the non-central t-distribution to calculate the power of a t-test.

4.1.3 Verification of error rate

After we choose the best model, we add a verification step that may not be strictly necessary for all applications. We can assess the overfitting/coverage/error-rate problem by then mounting another simulation study using the real data but permuting the treatment assignment (within sets).

1. Re-shuffle treatment many times (usually 1000 times), each time calculating a bootstrap-t CI following B bootstrap iterations.
2. Verify that the CI contains $A = 0$ at least 95% of the simulations.

4.1.4 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 by relying on theory that suggests that (1) that the randomization distribution under the null hypothesis will be governed by a central limit theorem 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 2. We created these outcomes using the following procedure:

1. Generate stratum specific means of one of the survey outcomes (“hours helped others in the last four weeks”), where the 20 strata, are UK administrative units crossed by 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 the standard deviation set to the standard deviation 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 $\beta_2 = -12$.

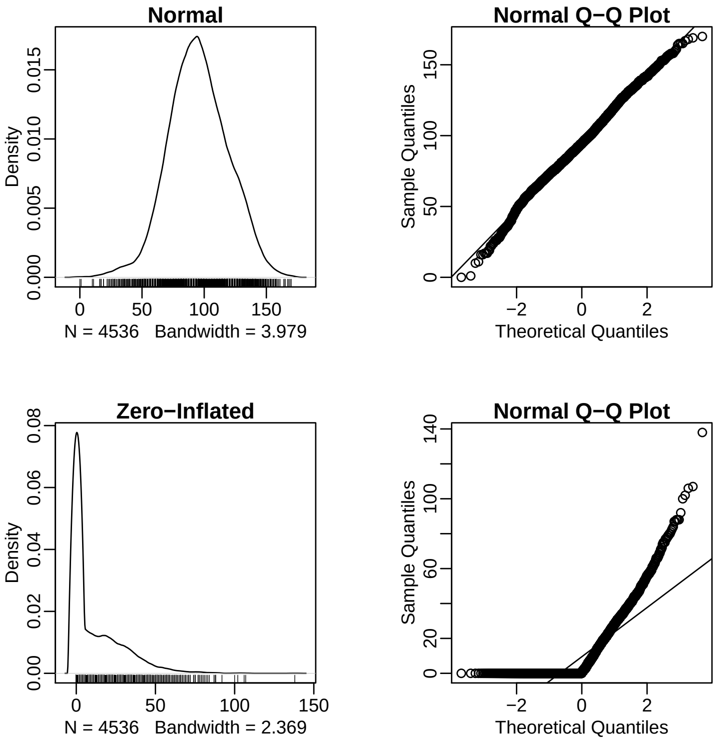

Figure 2: Distributions of simulated outcomes. Each outcome is non-negative integer and is a known linear additive function of respondent age and household income and twenty strata specific fixed effects (where strata are defined by administrative unit and gender of the respondent). The R^2 of the known model for the Normal outcome and Skewed outcome are 0.79 and 0.55 respectively.

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.

Figure 3 shows the result of following our covariance adjustment method. The thin black line shows the unadjusted power curve. The thick black line shows the improved power that would arise if we knew the true covariance adjustment model. Since, we rarely know the true model by which covariates predict outcomes we started the model selection algorithm with a model containing 146 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 not include the true model as a subset of the test model? The red lines in Figure 3 show that the answer is 'almost'.

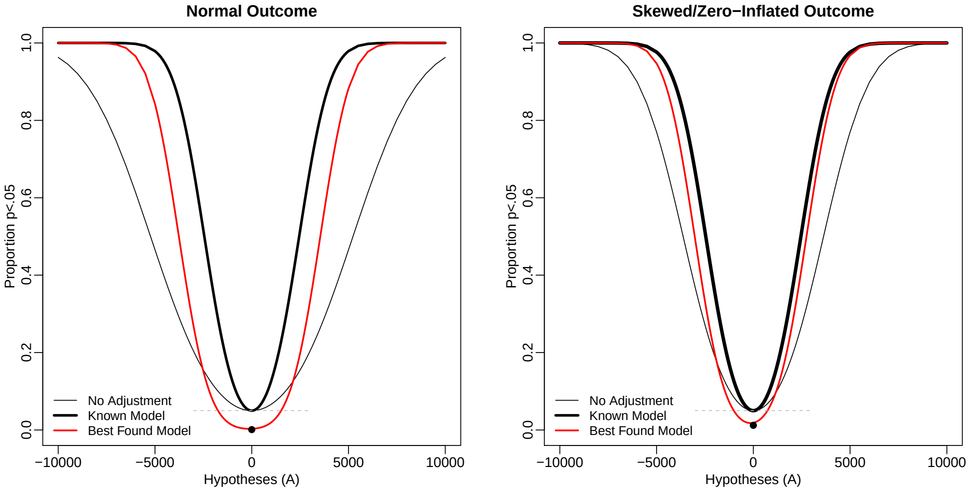

Figure 3: Power curves for $\alpha = .05$ for assessments of A_0. When $A_0 = 0$, no more than 5% of p-values are greater than .05 (gray dashed horizontal line). Thin black lines show the power curve for the unadjusted tests, the thick black lines show that power improves if one knows the true model relating covariates to outcomes, red lines show the power curves arising from model selection starting with a model containing 146 terms.

We also verified the validity of the tests based on the machine-generated covariance adjustment strategy: the proportion of p-value below 0.05 is shown with the black dots at $A = 0$. A well-operating test should reject the truth no more than 5% of the time if the pre-specified error rate is $\alpha = 0.05$. In this case, we see that our approach is conservative and thus valid.

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 in almost the same way that the true model operates. We should note that we made no specific assumptions about the outcomes here — the integer valued zero-inflated and skewed-outcome, for example, is manifestly not-normal.

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 interval correct. Thus, one should not read this paper as advocating our particular adaptive elastic net machine learner. 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

Here we show that our approach improved the power of statistical inferences using real outcomes.

Figure 4 compares the coverage of the confidence intervals between unadjusted and the selected covariance adjusted confidence intervals using as input a set of 2634 covariate terms. In each case, the covariance adjusted model has higher power for alternatives distant from the truth, and is conservative for tests of and near the truth (set to be $A = 0$ for the purpose of covariance adjustment finding).

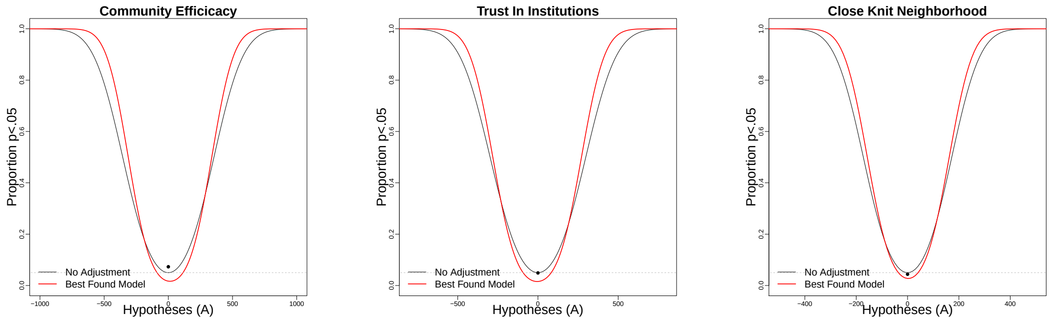

Figure 4: Real Outcomes: Power curves for $\alpha = .05$ for assessments of A_0. When $A_0 = 0$, no more than 5% of p-values are greater than .05 (gray horizontal line) or within simulation error of .05. Black lines show the power curve for the unadjusted tests, red lines show the power curves arising from choosing the most powerful model using a set of 2634 terms using the elastic net fitting procedure on the control group only. Black dots at $A = 0$ show simulation based false rejection rates for the chosen covariance adjustment model.

We see that the covariance adjusted confidence intervals have more power to exclude false hypotheses than the unadjusted tests — although this difference is slight. That is, this plot suggests that we would gain some power, but it ought not be as large as the power that we gained using the simulated data.

4.3.1 Confidence Intervals

Table 1 shows the confidence intervals for the attributable effect that arise from applying this method to the 2005 UK Home Office simulated and real outcomes.

	A	A per 100 Treated	Adj Width/Raw Width
Community Efficacy	(-400,420)	(-12,13)	1.1
Neighborhood Cohesion	(-220,-44)	(-6.7,-1.3)	0.51
Trust Institutions	(-450,-96)	(-13,-2.9)	0.58
Sim. Normal Outcome	(-2600,2900)	(-77,88)	0.52
Sim. ZIF Outcome	(-570,3500)	(-17,100)	0.57

Table 1: 95% Confidence intervals for the causal effect of bombings on the number of positive answers among the those interviewed after the bombing. Intervals reflect the best covariance adjustment model found during the tuning parameter search. Numbers shown to two significant digits.

The rightmost column compares the widths of the confidence intervals calculated with and without adjustment. In four of the five cases, we see an improvement of between 40 and 50% (ratios from about .6 to about .5). We did not see this improvement for community efficacy, although we speculate that 1.1 signals merely lack of improvement and simulation error and not decrement.

Using this approach we see that, of the roughly 3300 people interviewed after the bombing, we see that of the 3300 people interviewed after the bombing, we saw 6100 responses in favor of neighborhood cohesiveness. If these people had been interviewed before the bombing, we would have expected as many as 220 to 44 **more** responses in favor of neighborhood cohesiveness. That is, for every 100 people interviewed after the bombing, we should not be surprised to see between 7 and 1 more cohesive responses after the bombing. The effect of the bombing on community efficacy was weak: for every 100 people interviewed after the bombing, we would easily have seen about -12 fewer trusting responses to about 13 more trusting responses.

4.3.2 Summary

The overall lesson of this section is that one may use prognostic covariates to increase the precision of statistical inferences about causal inferences without repeatedly reexamining the causal effect itself. Our strategy confines attention to the control group and makes use of machine learning approaches to search for the best covariance adjustment specification (where ”best” means ”most powerful while maintaining nominal Type I error”).

5 Discussion and Conclusion

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. However, common practice in analyzing experiments risks seemingly significant results appearing by chance because of the process of hunting for a statistically powerful covariance adjustment specification while testing hypotheses about the treatment effect itself. The newer practice of pre-registering analyses will help maintain the error rates of the statistical tests used in experiments — by declaring in advance that one will use a set of variables in a certain way, one is able to execute one test with a known false rejection rate. However, the newer practice will probably leave power on the table by ignoring the relationships occurring within any given dataset. The noise in an outcome arises, in part, from idiosyncratic processes within a given moment of data collection (what is often called the "natural variation" in the outcome): even if a covariate tends to be strongly related to an outcome in general, it may or may not be strongly prognostic for that outcome in a given dataset — and it is this strength of prediction which determines it's utility in increasing the precision of our confidence intervals.

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*b*) 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.

In this paper we have demonstrated a way to use only data from the control group to train a machine learning algorithm. We used the elastic net algorithm in this paper, but any other such model can be used.[14]

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. We discovered that the common methods of tuning parameter selection in machine learning (mean squared error targeted k-fold cross-validation) tended to overfit the control group and thus produce invalid statistical tests. In this paper we advocate power analysis verified by false rejection rate error evaluation for tuning parameter selection in covariance adjustment specification searches for experiments.

Our application to the UK 2005 Home Office Survey demonstrated statistical inference for a causal quantity, the attributable effect, that is well suited to binary and count data. We produced confidence intervals for the attributable effect without requiring any model of the probability generating process of the observed outcomes. Rather than use a negative binomial model for one outcome and a Poisson model for

[14] One could even learn from the Super Learner ideas of Van der Laan and Rose (2011), in which one may fit many different models to produce an ensemble prediction (see Grimmer on ensemble prediction in machine learning).

another, our approach allows us to treat all count and binary outcomes in a unified framework that derives from a simple and flexible individual specific causal model.

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. First, without looking at outcomes, we bolstered our argument in that the London Bombings of July 2005 could be seen as a part of a natural experiment within the UK 2005 Home Office Survey by creating a stratified research design and by providing some evidence to show that our design cannot be distinguished from an equivalent randomized experiment. Second, we did look at outcomes, but only in the control group, and we changed common practice in machine learning by choosing tuning parameters based on power and size of tests. Third, our statistical inference for causal effects occurred only after the first two modules had been completed. Thus, our method side-steps worries about multiple testing while also enhancing precision. Because we verified that our confidence intervals have correct coverage as a part of the process, we could provide some evidence that this step in the process was valid as well as powerful.

Appendix A The 2005 UK Home Office Survey as a natural experiment

The survey was designed to be a nationally representative probability sample. The order of interviews within the survey, however, was not random. Figure 5 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 administered 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.

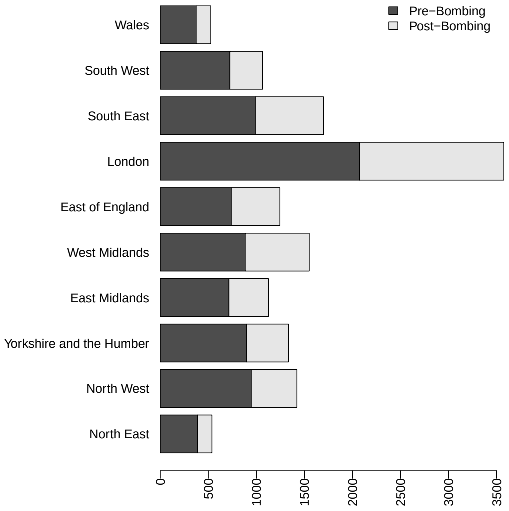

Figure 5: The counts of respondents interviewed before and after the London bombings, grouped by the governmental regions used for sampling and organization of the fieldwork for the 2005 Home Office survey across the UK. While the overall survey was designed to be a representative sample of the population, 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 6 plots the standardized mean differences across many covariates for pre and post bombing respondents. With many differences in excess of 0.05 pooled standard deviations (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 .15 standard deviations. 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 2 standard deviations.

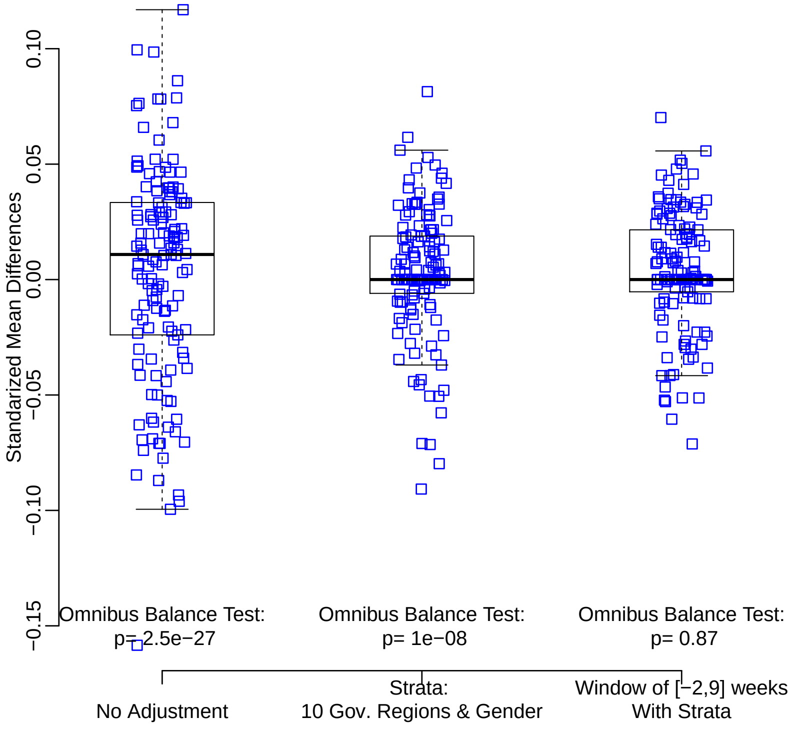

Figure 6: Covariate similarity assessments for 140 covariate terms before versus after the London 2005 bombings. Left panel shows difference in proportions or means for binary and continuous covariates respectively. The p-value for the Hansen and Bowers (2008) d^2 omnibus balance test is reported in 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.

Is a difference of .15 standard deviations large? In frequentist statistics, 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 .15 standard deviations?" If it would be very strange to see a difference of .15 standard deviations from the perspective of a randomized experiment, then we suspect that comparisons pre-vs-post bombing will not provide the clear interpretable comparisons offered by a randomized experiment.[15] If differences of .15 standard deviations would be typical of a randomized experiment then, 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).

Now, although we could answer this question about one covariate here we have chosen to inspect balance on 140 covariates. In practice, of course, even a randomized experiment will 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 6 also shows the p-value for a χ^2 based on the d^2 statistic.

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. 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 6 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

[15] Kinder and Palfrey (1993) cite Campbell and Stanley (1966) to provide us with the idea that the point of an experiment is to produce clearly interpretable comparisons.

14078 respondents, after shrinking the window 6451 subjects remain.

Appendix B Outcome Questions

Figure 7 shows the distribution of the number of positive responses to the outcome questions compared before and after the bombing of July 7, 2005.

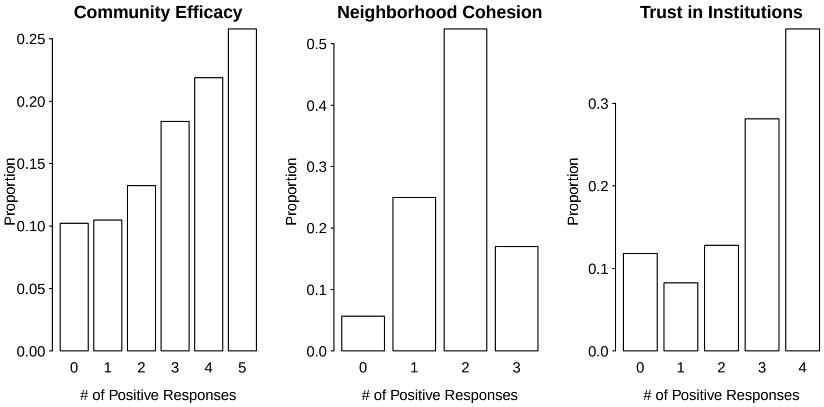

Figure 7: Perceptions of social cohesion and trust in institutions for respondents interviewed in the 2 weeks before the London Bombing of July 7, 2005 ($n_{\text{pre}} = 1195$) in the UK Home Office Survey 2005.

Appendix B.1 Trust in Institutions

- How much do you trust the police? [PTPolc] (1) A lot, (2) A fair amount, (3) Not very much, (4) Not at all
- How much do you trust the Courts (Magistrates & Crown Courts) [PTCrt]
- How much do you trust Parliament [PTParl]
- How much do you trust your local council? [PTCncl]

The trust in institutions variable that we use is the number of ”a lot” answers.

Appendix B.2 Community Efficacy

- If a group of local children were playing truant from school and hanging around on a street corner, how likely is it that people in your neighbourhood would do something about it? [STruant]

  (1) Very likely, (2) Likely, (3) Unlikely, (4) Very unlikely, (5) Don’t know

- If some children were spray-painting graffiti on a local building, how likely is it that people in your neighbourhood would do something about it? [SGraff]

- If there was a fight near your home and someone was being beaten up or threatened, how likely is it that people in your neighbourhood would do something about it? [SFight]
- If a child was being rude to an adult, how likely is it that people in your neighbourhood would tell that child off? [SRude]
- How likely is it that people in your neighbourhood would participate if they were asked by a local organisation to help solve a community problem? [SProb]

The community efficacy variables is the number of ”Very likely” or ”Likely” responses

Appendix B.3 Neighborhood Cohesion

- People in this neighbourhood are willing to help their neighbours? [SHelp] (1) Strongly agree, (2) Agree (3) Disagree (4) Strongly disagree (5) Don’t know
- This is a close-knit neighbourhood? [SClose]
- People in this neighbourhood do not share the same values? [SValue]

The neighborhood cohesion variable is the number of ”Strongly agree” or ”Agree” responses.

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/*

Belson, W.A. 1956. “A technique for studying the effects of a television broadcast.” *Applied Statistics* pp. 195–202.

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet Sekhon and Bin Yu. 2015. “Lasso adjustments of treatment effect estimates in randomized experiments.” *arXiv preprint arXiv:1507.03652* .

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.

Campbell, D.T. and J.C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research.* Houghton Mifflin.

Cox, David R. 1958. *The Planning of Experiments.* John Wiley.

Efron, Bradley. 2010. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction.* Vol. 1 Cambridge University Press.

Faraway, Julian J. 1992. “On the cost of data analysis.” *Journal of Computational and Graphical Statistics* 1(3):213–229.

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.

Gelman, Andrew and Eric Loken. 2013. “The garden of forking paths: Why multiple comparisons can be a problem, even when there is no Śfishing expeditionŠor Śp-hackingŠand the research hypothesis was posited ahead of time.” *Downloaded January* 30:2014.

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 Adam Sales. 2015. “Comment on Cochrans “Observational Studies”.” *Observational Studies* 1(1):184–193.

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.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman and James Franklin. 2005. “The elements of statistical learning: data mining, inference and prediction.” *The Mathematical Intelligencer* 27(2):83–85.
URL: *http://www-stat.stanford.edu/ tibs/ElemStatLearn/*

Humphreys, Macartan, Raul Sanchez de la Sierra and Peter Van der Windt. 2013. “Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration.” *Political Analysis* 21(1):1–20.

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

Lin, Winston et al. 2013. “Agnostic notes on regression adjustments to experimental data: Reexamining FreedmanŠs critique.” *The Annals of Applied Statistics* 7(1):295–318.

Lohr, S. 2001. *Sampling: Design and Analysis*. 2nd ed ed. Brooks/Cole.

Neyman, J. 1923 [1990]. “On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (1923).” *Statistical Science* 5:463–480. reprint. Transl. by Dabrowska and Speed.

Peters, C.C. 1941. “A method of matching groups for experiment with no loss of population.” *The Journal of Educational Research* pp. 606–612.

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*. “Attributing effects to treatment in matched observational studies.” *Journal of the American Statistical Association* 97(457):183–192.

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

Rosenbaum, Paul R. 2008. “Testing hypotheses in order.” *Biometrika* 95(1):248–252.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. Springer.
URL: *http://www.springer.com/statistics/statistical+theory+and+methods/book/978-1-4419-1212-1*

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

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

Schochet, Peter Z. 2010. "Is regression adjustment supported by the Neyman model for causal inference?" *Journal of Statistical Planning and Inference* 140(1):246–259.

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

Van der Laan, Mark J and Sherri Rose. 2011. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

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