How to increase the precision of causal inferences in experiments using machine learning (but without data snooping).

Jake Bowers[1] Mark Fredrickson[1] Ben B. Hansen[2]
Escuela de Invierno de Métodos
Departamento de Ciencias Sociales y Políticas
Universidad Católica de Uruguay

[1] Political Science & Statistics & NCSA @ University of Illinois
jwbowers@illinois.edu — http://jakebowers.org

[2] Statistics, University of Michigan

The Challenge: Funky Outcomes and Many Covariates

The General Motivation: The civic effects of terrorism.
The Specific Study: The London Bombings of 2005 and voluntary activities (from the 2005 UK Home Office Survey).

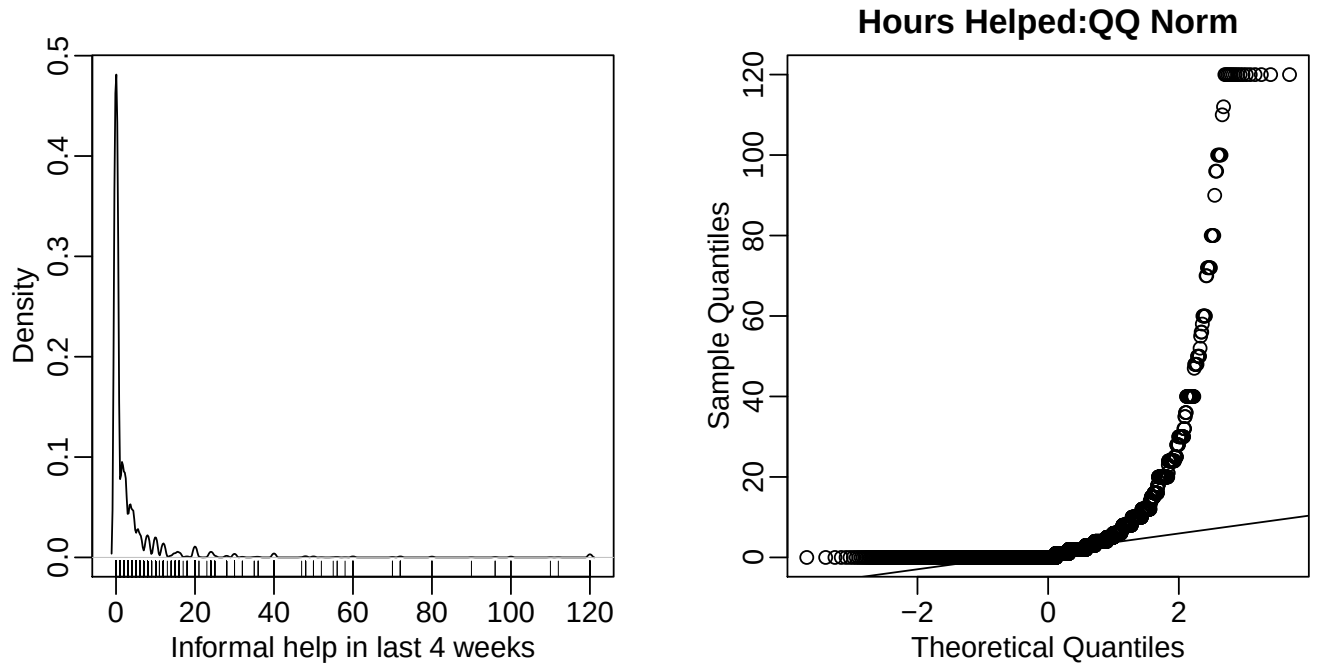


Figure 1: Hours of informal help for non-relatives in the last four weeks (UK Home Office Survey 2005). These observations from a window around the London Bombing of 2005 (-2 weeks $n_{\text{pre}} = 1195$, +8 weeks $n_{\text{post}} = 3341$).

The Challenge: Funky Outcomes and Many Covariates

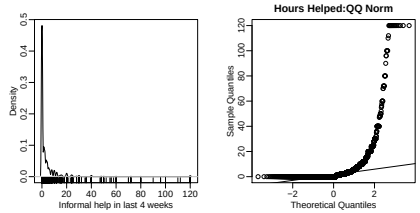


Figure 1: Hours of informal help for non-relatives in the last four weeks (UK Home Office Survey 2005). These observations from a window around the London Bombing of 2005 (-2 weeks $n_{\text{pre}} = 1195$, +8 weeks $n_{\text{post}} = 3341$).

Opportunity: A natural experiment; many covariates with possibility to increase precision of statistical inferences about causal effects (1000+)
The Methodological Challenge: We want a procedure for using many covariates plus substantive knowledge to increase precision. Specifically we want a work-flow that is:

Powerful ...so that we can shrink confidence intervals as much as possible to detect small effects.(avoid conservative tests)
Valid ...so that we can use covariates to increase precision without compromising test validity.(avoid bias)
Transparent ...so that others can replicate and we can register analyses *before* data collection. (avoid snooping)
Design-based ...so that we focus on causal comparisons rather than on probability models of outcomes.(avoid debates about stochastic processes and functional forms)

A Proposal

Power from machine learning Use machine learning to select a most powerful covariance adjustment strategy (but constrained to maintain unbiased tests).

Separate adjustment from effect assessment Restrict specification search to the control group to avoid bias and enable transparency (Peters, 1941; Belson, 1956).

Design-based inference Use randomization inference to produce confidence intervals about the causal effect of the bombing (Hansen and Bowers, 2009).

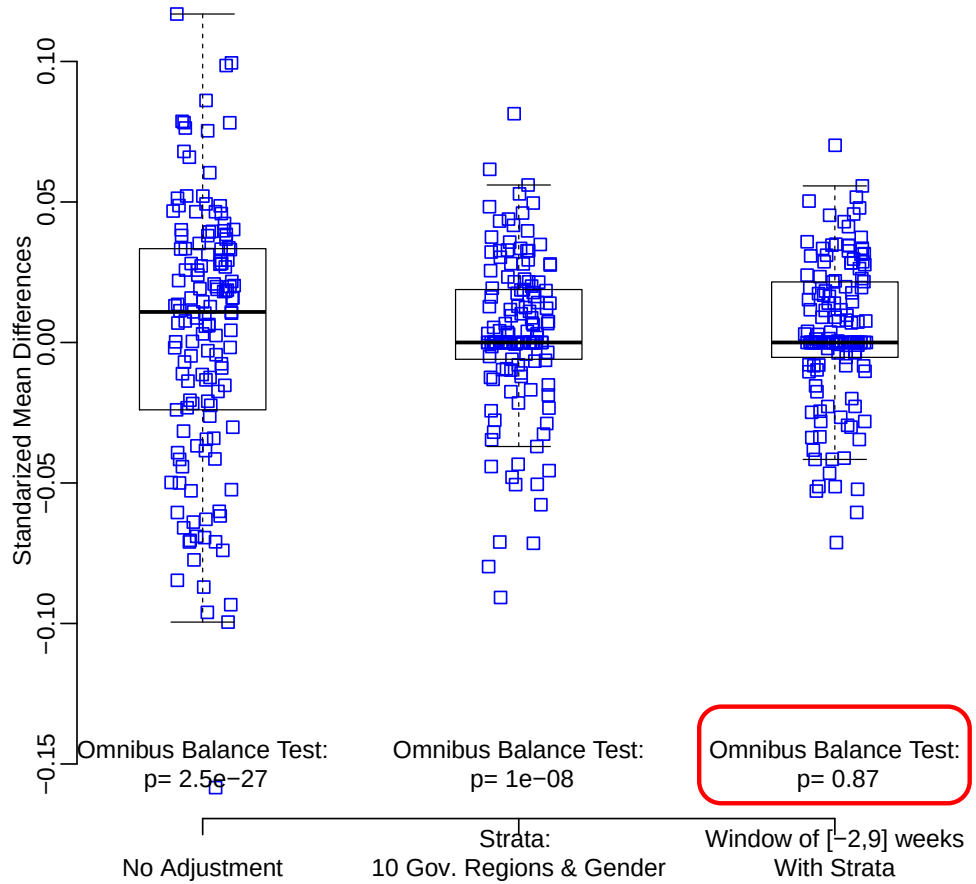
Add Sensitivity Analysis To restrict range of alternative explanations. A natural experiment is not a randomized experiment.

A (natural) experiment?

The randomized experiment is a minimal standard for clear comparisons.

Task: Create a design observationally indistinguishable from a randomized experiment.

Details: We measure closeness to randomized experiment with the d^2 omnibus balance test on 140 background covariate terms (Hansen and Bowers, 2008). Final design $N = 4536$ with 1195 interviews pre-bombing and 3341 post-bombing.



What do we want to know?

- Y_i is the observed outcome, $Z_i \in \{0 = \text{control}, 1 = \text{treatment}\}$.
- $y_{i,Z_i=1} \equiv y_{i,1}$ is partially observed potential response (+SUTVA).
- Generalized Attributable Effect is $A = \sum_i Z_i \tau_i$, where $\tau_i = y_{i,1} - y_{i,0}$ and $y \geq 0$. (Original attributable effects for $y \in \{0, 1\}$ (Rosenbaum, 2002).)
- U is experimental pool, $C \subseteq U$, so $\sum_{i \in C} Y_i - y_{i,0} = 0$

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

$$= \sum_{i \notin C} Y_i - \sum_{i \notin C} y_{i,0} = \sum_{i=1}^{N} Y_i - \sum_{i=1}^{N} y_{i,0} = t_U - t_C \tag{2}$$

$$= \underbrace{\text{observed total overall}}_{\text{fixed and observed}} - \underbrace{\text{total outcome under control}}_{\text{unobserved, to estimate}} \tag{3}$$

A is the number of hours volunteered after the bombing that we would not have seen in the absence of the bombing.

How can we use what we observe to learn about A?

Recall:

$$A = \sum_{i=1}^{N} Z_i \tau_i = t_U - t_C \tag{4}$$

$$= \underbrace{\text{observed total overall}}_{\text{fixed and observed}} - \underbrace{\text{total outcome under control}}_{\text{unobserved, to estimate}} \tag{5}$$

An approximate confidence interval for $\hat{A}$

Extend Hansen and Bowers (2009) from binary to count outcomes:

1. Observed total outcomes, t_U, is fixed across randomizations.
2. Survey regression estimator $\hat{t}_C = \sum_{i \in U} \hat{Y}_i + \sum_{i \in C} (Y_i - \hat{Y}_i)$ with $\hat{Y}_i = f(\mathbf{X}_i, \beta)$
3. As $N \to \infty$, $\text{CI}(\hat{t}_C) \approx \hat{t}_C \pm z_{\alpha/2} \text{SE}(\hat{t}_c)$.
4. $\widehat{\text{SE}}(\hat{t}_c)$ from standard sampling theory+adjustment for model fitting (Hastie et al., 2005, Chap 7) .
5. **So, $\widehat{\text{CI}}(\hat{A}) \approx t_U - \widehat{\text{CI}}(\hat{t}_C)$.**

Note 1: No parameterized stochastic model of outcomes (i.e. no presumption of zero-inflated poisson, etc . . .).

Note 2: Penalty for incorrect model of $Y_{i \in C}$ is lack of power not bias (Hansen and Bowers (2009), Särndal and Swensson (2003), Lohr (1999))

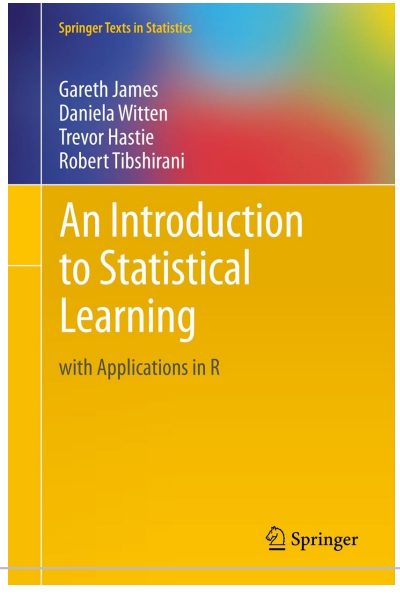
How to choose β and $\mathbf{X}$?

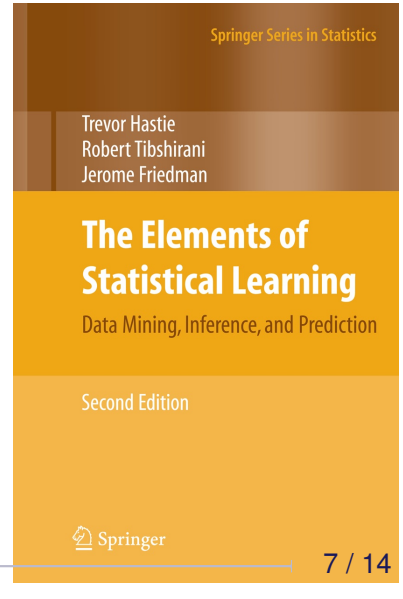
Example Options

1. Chose $\mathbf{X}$ with $n >> k$ and choose β to maximize fit to data: $\hat{\beta}(\text{OLS}) = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - \mathbf{X}_i\beta)^2$.
2. Chose a large $\mathbf{X}$ (perhaps $n < k$). Penalize overly large β_k and make overly small $\beta_k = 0$. (Lasso and variants). For example, the elastic net (Zou and Hastie, 2004):

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

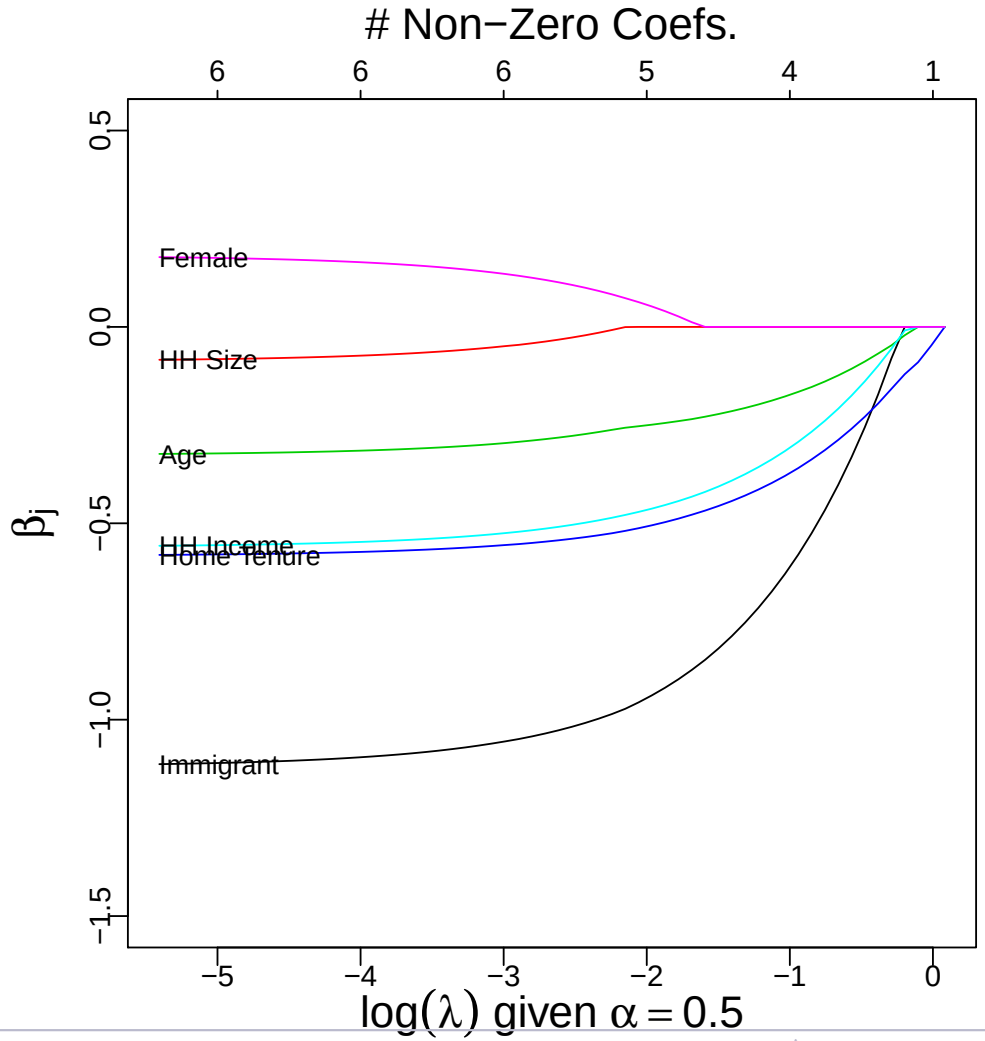
3. Many others! (Adaptive elastic net, random forests)





How to choose β and $\mathbf{X}$?

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



Procedure to choose tuning parameters

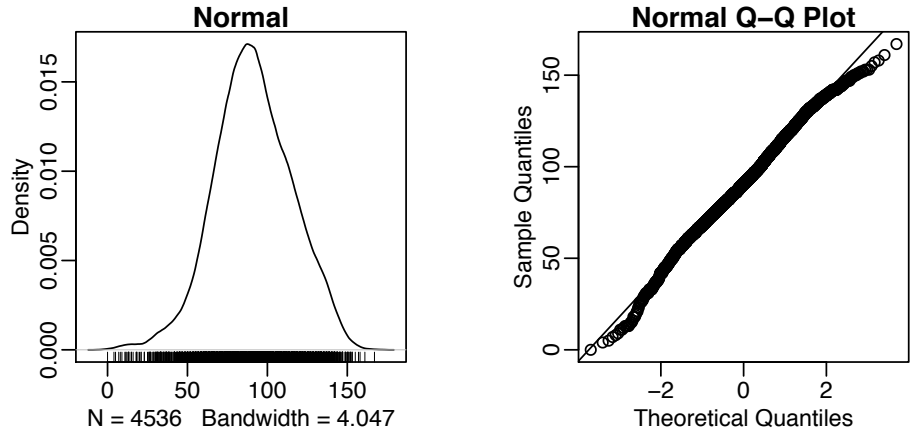
Repeat B times:

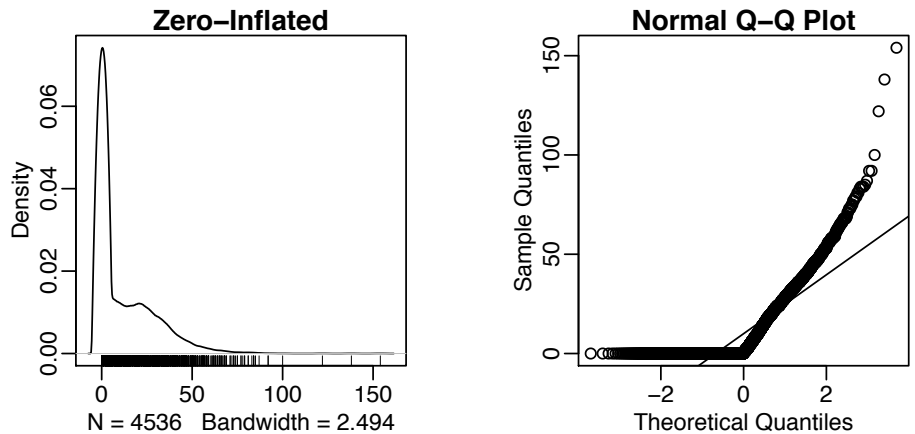
1. Bootstrap the control group to make synthetic data $N \times p$.
2. Assign $\mathbf{Z}$ in the synthetic data following the design.
3. Choose tuning parameters and fit a penalized model to the pseudo-control group ($\mathbf{Z} = 0$).
4. Produce $\hat{t}_C = \sum_{i \in U} \hat{Y}_i + \sum_{i \in C} (Y_i - \hat{Y}_i)/\pi_i$ where π_i is probability of assignment to C (varies by stratum) and $\widehat{\text{CI}}(\hat{A}) = t_U - \widehat{\text{CI}}(\hat{t}_C)$.
5. Record size of the test — the proportion of B simulations that the truth is outside the CI. (Should be rare.)
6. Record power of the test — the proportion of the B simulations that an alternative is outside the CI. (Should be often.)

Choose tuning parameters with maximum power for size $\leq \alpha$.

Can we do as well as a known covariate model?

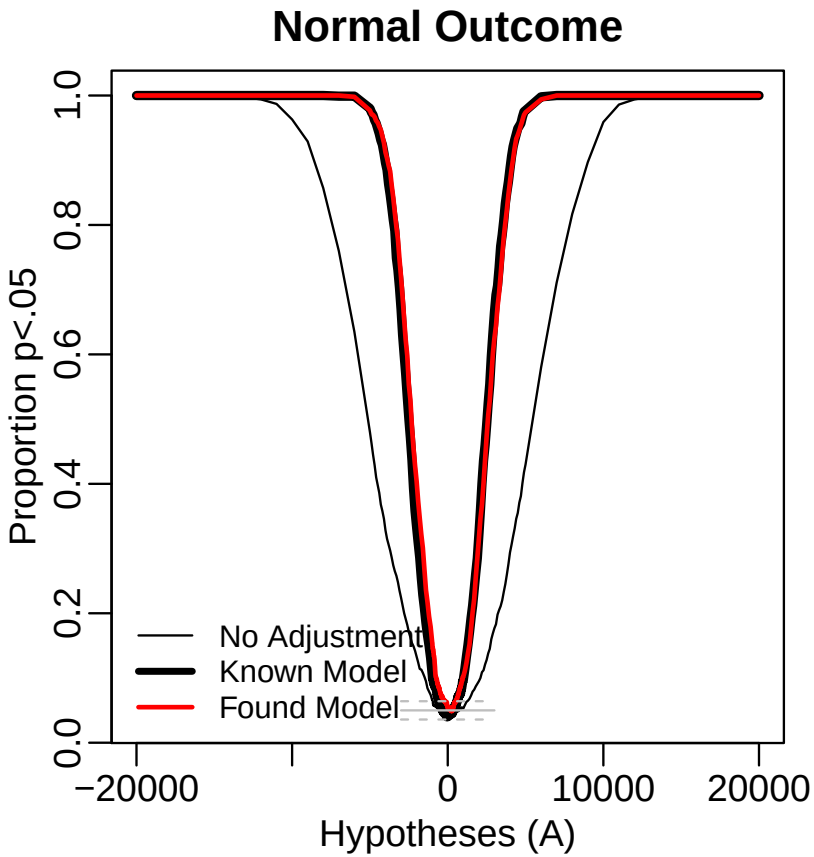
Known Outcomes: Simulated outcomes with known covariance relationships using data from the UK 2005 Home Office Survey: for each person i and Gender$\times$Region strata j set $Y_i = \beta_{0j} + 16 \times \text{Age}_i + 12 \times \text{Income}_i + e_i$ with $e \sim N(0, \sigma^2)$ or $e_i \sim \pi_i \text{Geom}(.7) + (1 - \pi_i)\text{Geom}(.07)$ where $\pi_i \sim \text{Bernoulli}(.5)$.

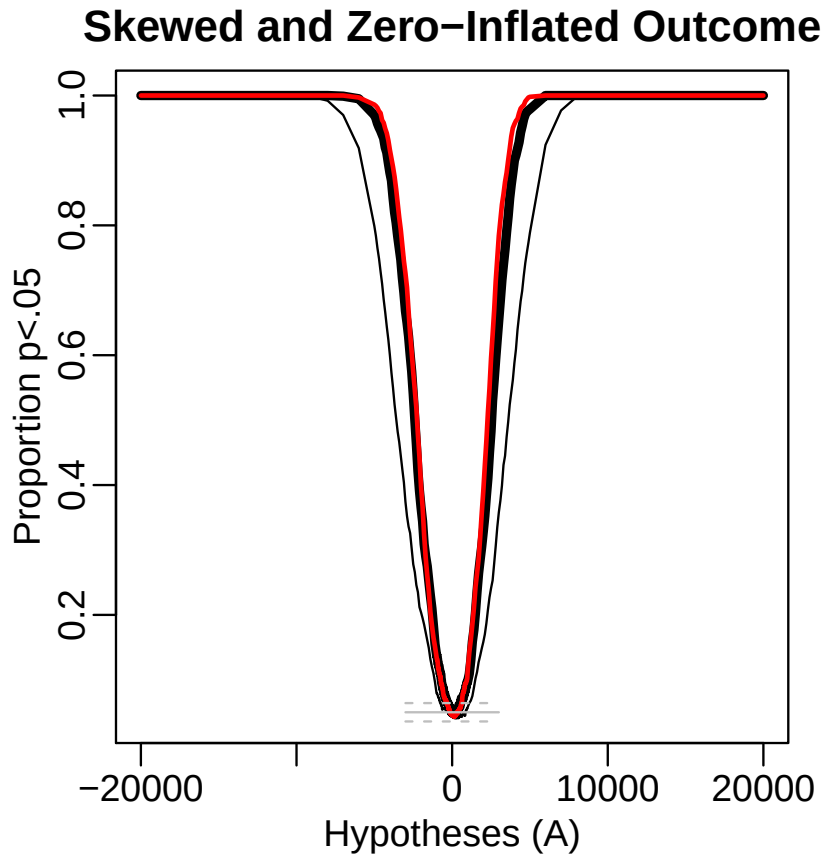




We can do as well as a known model!

Details: Power curves for $\alpha = .05$ for CI($\hat{A}$). Elastic net model search inputs: 143 covariate terms (Age and Income as 4 piece natural cubic spline bases) with $B = 1000$ simulations per tuning parameter proposal (γ, λ). R^2 for the best models were .71 and .58 for the Normal and Zero-inflated outcome respectively.





95% Confidence intervals for $\hat{A}$

	Unadjusted	Adjusted	Difference in CI Widths
Normal Outcome	(-4900,5800)	(-1400,3600)	5700
ZIF Outcome	(-5800,1400)	(-3400,1800)	2000

Table 1: Confidence intervals (95%) for simulated outcomes. "Unadjusted" intervals do not involve covariates. "Adjusted" intervals reflect the best covariance adjustment model found during the tuning parameter search. Numbers shown to two significant digits. Adjusted intervals are 47% or 72% the size of the unadjusted intervals but equally valid (i.e. have correct coverage).

How to interpret $\hat{A}$?

1. Implausible if more than 1800 hours of civic work were donated after the bombing that would not have been donated before the bombing among the 4500 respondents in this survey.
2. Divide CI by n_t to get Effect of Treatment on the Treated (ETT) per treated unit: (-1 hour $= -3400/3341, 0.54$ hour $= 1800/3341$).

Summary

We shrank valid, design-based, confidence intervals for funky outcomes by about 30% without looking at our treatment effect.

We brought insights from machine learning (1996-now) together with insights from survey sampling (1950s-1960s) and the analysis of randomized experiments (1920s-1930s, 1990s-now) to enable applied researchers to use prior knowledge and data.

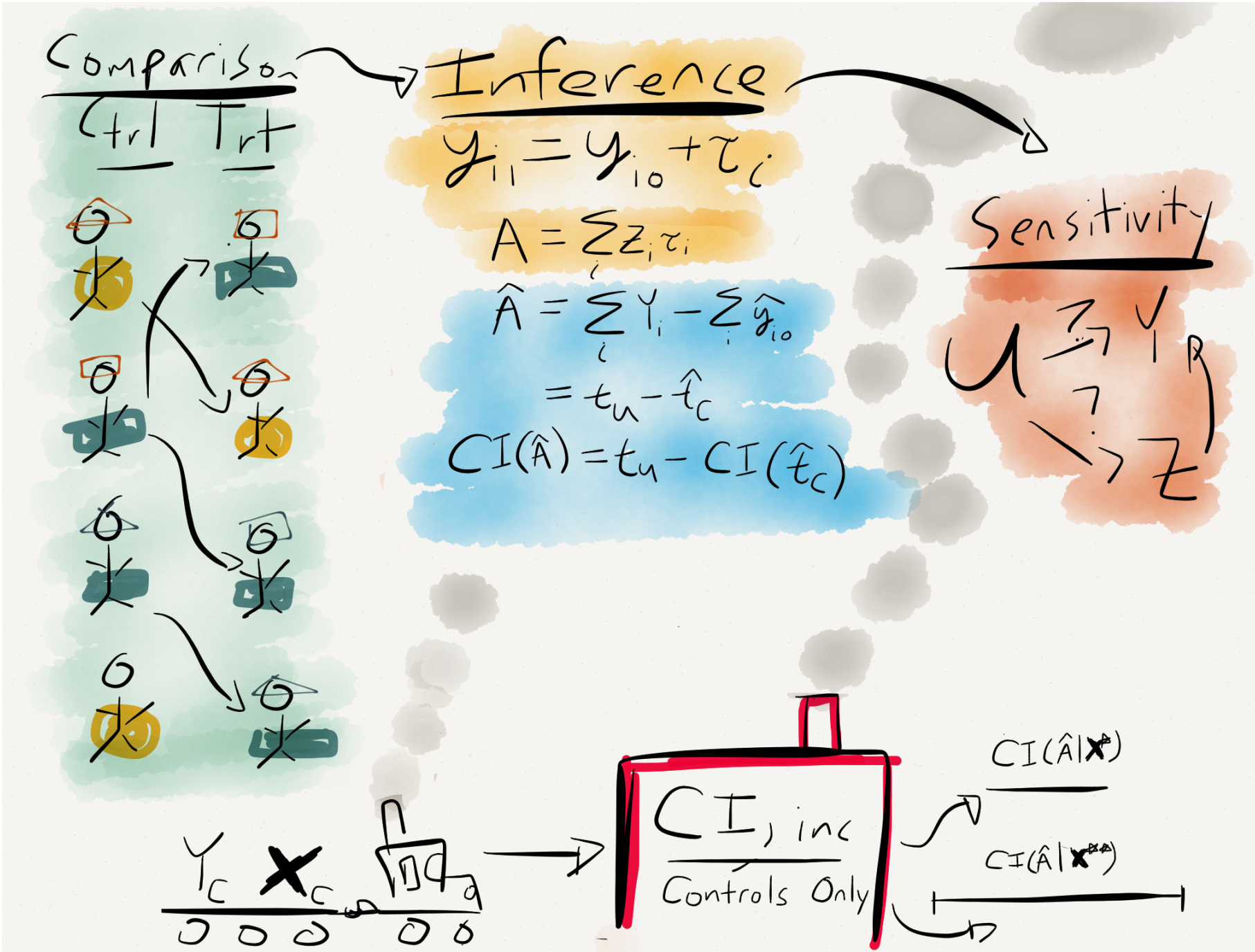
So What? What else?

A modular approach

- allows the application of different standards to different analytic tasks.
- allows the use of linear models while preventing data snooping and multiple testing.
- allows precision to be increased while maintaining the promises of design-based frequentist testing procedures.

Other Contributions

- We extended the definition of $A = \sum_i (y_{i,1} - y_{i,0})$ to cover count variables and showed how to produce confidence intervals for this new, generalized, attributable effect. We thus contribute to the analysis of zero-inflated count outcomes.
- A novel tuning parameter selection procedure using operating characteristics of assessments of treatment effects to choose models.
- We did not specify a parameter stochastic model for Y_i; our statistical inference is design-based.


Comparison
Ctrl Trt
Inference
$y_{i1} = y_{i0} + \tau_i$
$A = \sum_i z_i \tau_i$
$\hat{A} = \sum_i Y_i - \sum_i \hat{y}_{i0}$
$= t_u - \hat{t}_c$
$CI(\hat{A}) = t_u - CI(\hat{t}_c)$
Sensitivity
U
Y
R
Z
Y_c X_c
CI, inc
Controls Only
$CI(\hat{A}|X)$
$CI(\hat{A}|X^{**})$

Belson, W. (1956), "A technique for studying the effects of a television broadcast," *Applied Statistics*, 195–202.

Hansen, B. B. and Bowers, J. (2009), "Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign." *Journal of the American Statistical Association*, 104, 873—885.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005), "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, 27, 83–85.

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

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

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

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

Zou, H. and Hastie, T. (2004), "Regression shrinkage and selection via the elastic net, with applications to microarrays," *JR Statist. Soc. B*.

The London Bombings: July 7, 2005

ON THIS DAY 1950 2005

7 July

BBC NEWS

Search ON THIS DAY by date 7 July GO

Front Page | Years | Themes | Witness

About This Site | Text Only

2005: Bomb attacks on London

A series of bomb attacks on London's transport network has killed more than 30 people and injured about 700 others.

Three explosions on the Underground left 35 dead and two died in a blast on a double decker bus.

The first three bombs went off at 0850 on underground trains just outside Liverpool Street and Edgware Road stations, and on another travelling between King's Cross and Russell Square.

The final explosion was around an hour later on a number 30 double-decker bus in Tavistock Square, not far from King's Cross

Foreign Secretary Jack Straw said the bombings had "the hallmarks of an al-Qaeda-related attack".

Prime Minister Tony Blair promised the "most intense police and security service action to make sure we bring those responsible to justice".

Mr Blair, who flew back to London from the G8 summit in Gleneagles, condemned the terrorists and paid tribute to the stoicism and

Watch/Listen

A medic helps an injured woman after the attack on Edgware Road tube station

PLAY VIDEO

Scenes across London after the terror attacks

In Context

In total 52 people lost their lives in the London bombings, 700 people were injured.

The attacks were carried out by four suicide bombers.

The presumed ringleader, Mohammed Siddique Khan, had recently visited Pakistan and was later found to have made a video message in which he claimed British foreign policy was oppressing Muslims.

Al-Qaeda issued a videotaped statement in September claiming it was behind the London

Stories From 7 Jul

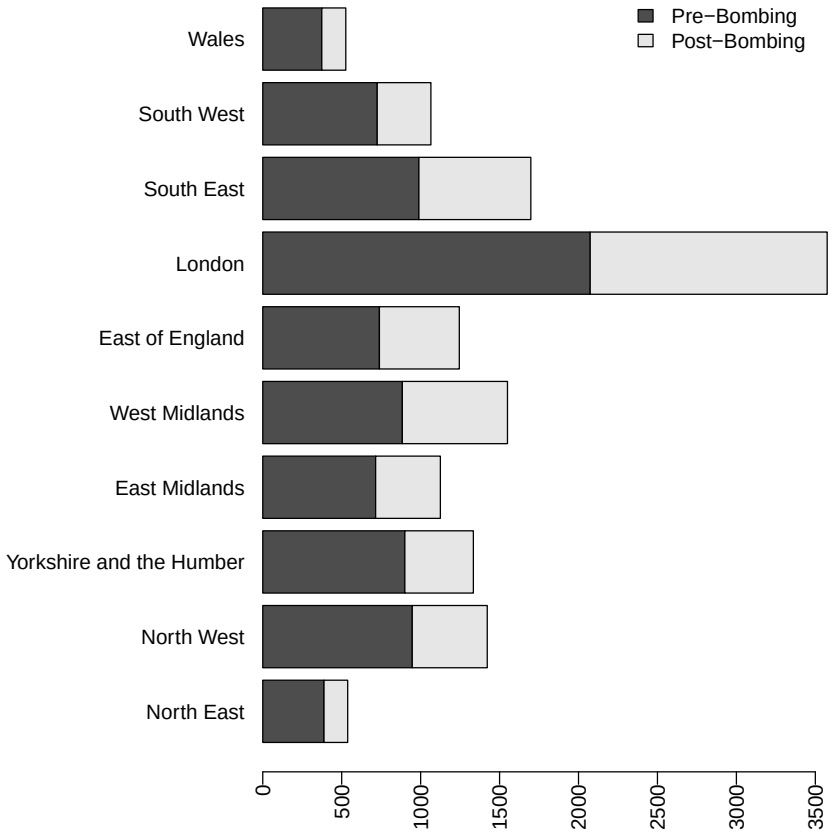
- 2005: Bomb attacks on London
- 2001: Two stabbed in Bradford race riots
- 1976: British grandmother missing in Uganda
- 1985: Boris Becker wins Wimbledon at 17
- 1969: Brian Jones died of 'drink and drugs'
- 1998: Chief's death sparks turmoil in Nigeria

BBC News >>

- **In depth**
  London terror attacks: 7 July and 21 July

The 2005 Home Office Citizenship Survey

In-person interviews, March 8 – September 30, 2005, $N = 14,000$, Stratified sample by governmental region.

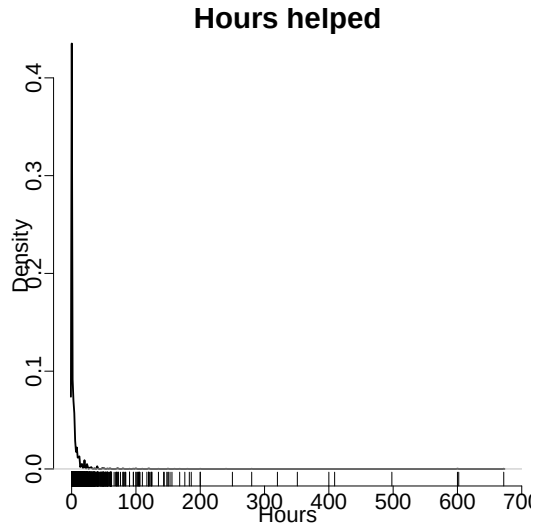


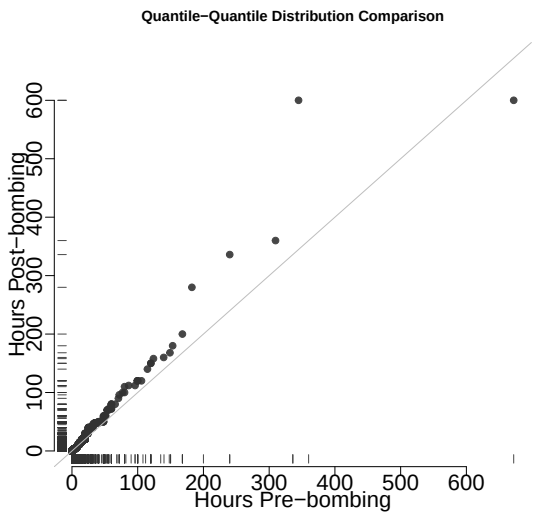
Outcome: Social capital

In the last 12 months have you, as an individual, done any unpaid help . . . any unpaid help you may have given to other people, that is apart from any help given through a group, club or organisation. This could be help for a friend, neighbour or someone else but not a relative.

Now just thinking about the past 4 weeks. Approximately how many hours have you spent doing this kind of thing/these kind of things in the past 4 weeks?

The 2005 Home Office Citizenship Survey



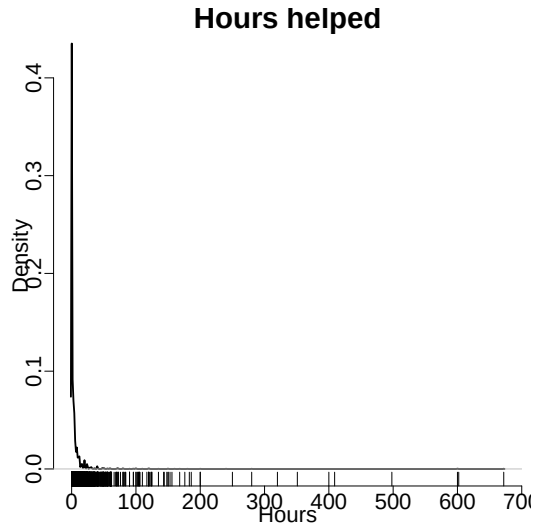


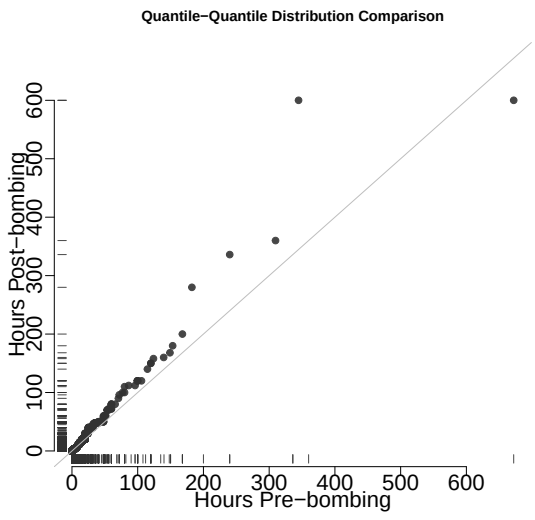
Outcome: Social capital

In the last 12 months have you, as an individual, done any unpaid help . . . any unpaid help you may have given to other people, that is apart from any help given through a group, club or organisation. This could be help for a friend, neighbour or someone else but not a relative.

Now just thinking about the past 4 weeks. Approximately how many hours have you spent doing this kind of thing/these kind of things in the past 4 weeks?

The 2005 Home Office Citizenship Survey





Outcome: Social capital

In the last 12 months have you, as an individual, done any unpaid help . . . any unpaid help you may have given to other people, that is apart from any help given through a group, club or organisation. This could be help for a friend, neighbour or someone else but not a relative.

Now just thinking about the past 4 weeks. Approximately how many hours have you spent doing this kind of thing/these kind of things in the past 4 weeks?

Analytic Challenges

Clarify comparison How can we make pre-vs-post bombing comparisons focus on differences caused by the bombing — not differences in types of people?

Define effects What counterfactual quantity is our target of causal and statistical inference? An average? Something else?

Use what we know to enhance precision How can we use the fact that we know that education ought to predict the outcome to shrink our confidence intervals without worrying about (1) multiple comparisons/testing and/or (2) data snooping?

Assess effects How can we make confidence intervals which exclude the true null rarely and in the controlled fashion (and which have an easy to describe target of statistical inference)?

Engage assumptions Since the bombings were not randomized, how can we learn about how our results might change if we were to discover a new predictor of interview week that also predicts the outcome?

Analytic Challenges

Clarify comparison How can we make pre-vs-post bombing comparisons focus on differences caused by the bombing — not differences in types of people?

Define effects What counterfactual quantity is our target of causal and statistical inference? An average? Something else?

Use what we know to enhance precision How can we use the fact that we know that education ought to predict the outcome to shrink our confidence intervals without worrying about (1) multiple comparisons/testing and/or (2) data snooping?

Assess effects How can we make confidence intervals which exclude the true null rarely and in the controlled fashion (and which have an easy to describe target of statistical inference)?

Engage assumptions Since the bombings were not randomized, how can we learn about how our results might change if we were to discover a new predictor of interview week that also predicts the outcome?

Analytic Challenges

Clarify comparison How can we make pre-vs-post bombing comparisons focus on differences caused by the bombing — not differences in types of people?

Define effects What counterfactual quantity is our target of causal and statistical inference? An average? Something else?

Use what we know to enhance precision How can we use the fact that we know that education ought to predict the outcome to shrink our confidence intervals without worrying about (1) multiple comparisons/testing and/or (2) data snooping?

Assess effects How can we make confidence intervals which exclude the true null rarely and in the controlled fashion (and which have an easy to describe target of statistical inference)?

Engage assumptions Since the bombings were not randomized, how can we learn about how our results might change if we were to discover a new predictor of interview week that also predicts the outcome?

Analytic Challenges

Clarify comparison How can we make pre-vs-post bombing comparisons focus on differences caused by the bombing — not differences in types of people?

Define effects What counterfactual quantity is our target of causal and statistical inference? An average? Something else?

Use what we know to enhance precision How can we use the fact that we know that education ought to predict the outcome to shrink our confidence intervals without worrying about (1) multiple comparisons/testing and/or (2) data snooping?

Assess effects How can we make confidence intervals which exclude the true null rarely and in the controlled fashion (and which have an easy to describe target of statistical inference)?

Engage assumptions Since the bombings were not randomized, how can we learn about how our results might change if we were to discover a new predictor of interview week that also predicts the outcome?

Analytic Challenges

Clarify comparison How can we make pre-vs-post bombing comparisons focus on differences caused by the bombing — not differences in types of people?

Define effects What counterfactual quantity is our target of causal and statistical inference? An average? Something else?

Use what we know to enhance precision How can we use the fact that we know that education ought to predict the outcome to shrink our confidence intervals without worrying about (1) multiple comparisons/testing and/or (2) data snooping?

Assess effects How can we make confidence intervals which exclude the true null rarely and in the controlled fashion (and which have an easy to describe target of statistical inference)?

Engage assumptions Since the bombings were not randomized, how can we learn about how our results might change if we were to discover a new predictor of interview week that also predicts the outcome?

Entertain "what if" questions about the design.

As-if-randomized story: For strata s, prob$(Z_{is} = 1) = m_S/n_s$ for all i in s.

Not-really-randomized question: What if prob$(Z_{is} = 1) = \pi(m_S/n_s)$, where $\pi > 1$? How large of a π can we entertain without a change in our confidence intervals? (Rosenbaum)

Omitted variable question: What if we discovered a new covariate, U_i, which predicted both timing of interview *and* the outcome. How much would our confidence intervals change? (Hosman, Hansen, Holland; Clarke; Imbens)

A sensitive design shows substantively large shifts in statistical inference for small changes in π or for U_i with mild relationships to treatment and outcomes.

Is there a standard for minimal insensitivity?