200C Submission Paper for `benchtestr`

Igor Geyn and Shing Hon Lam

6/11/2021

Introduction

Benchmarking experimental findings with observational data – a practice famously encapsulated by Dehejia and Wahba's (1999) evaluation of Lalonde's (1986) National Supported Work (NSW) – is an important and informative exercise. More recent examples of this kind of analysis abound: see, for example, Su, et al (2012), Gaines and Kuklinski (2011). When executed properly, an observational benchmark of an experimental study can reveal key information about the generalizability of the experiment. Unfortunately, while high-quality benchmarking data exists and has been leveraged in a number of past benchmarking efforts as well as meta analyses of benchmarking, this data is not readily available to the researcher.

This package compiles data across substantive fields including political science, medicine, and education to allow for robust benchmarking of causal estimates. Specifically, `benchtestr` enables the user to leverage common estimation techniques – including a range of matching approaches – on the above-mentioned datasets. `benchtestr` presents a number of intuitive functions, helpful defaults, and diverse datasets that can be tested using existing classical estimators (e.g., naive difference-in-means) as well as those defined by the user.

Methodological Motivation

Experimentation, implemented practically by such forms as the randomized control trial (RCT), is widely regarded as the gold standard for causal inference. Through truly random assignment of treatment, the experimenter can have the most confidence in the assumptions that undergird valid causal interpretation: ignorability, monotonicity, and relevance. (To be sure, concerns about attrition and other threats to inference persist (Sovey and Green 2011).) In political science, as in other disciplines, experiments allow researchers to yield particularly valid estimates — for example, of the effect of get-out-the-vote (GOTV) efforts on voter turnout (Gerber and Green 2005), of the effects of negative campaign ads on presidential approval (Gaines and Kuklinski), and issue salience (Iyengar and Kinder 1987) — that are used as so-called benchmarks. In other words, they are understood to be less vulnerable to confounding/omitted variable bias, endogeneity, and other forms of bias relative to observational studies.

Still, experimentation is infeasible in many contexts. For example, studying the success of GOTV efforts is potentially vulnerable to attrition, two-directional noncompliance, and other implementation issues that could impact the validity of an estimated treatment effect. Kosuke Imai's (2005) challenge of findings in Gerber and Green (2000), in which he identifies "failure of randomization that would have been difficult to detect (and indeed were not detected) without my [the author's] methods" and ultimately shows the failure of the experiment to be implemented exactly as intended (and exactly as characterized by the authors of the study). In fact, Imai's use of observational methods in his challenge demonstrates benchmarking of experimental findings using observational methods rather than the other way around, which is the typical conception of benchmarking studies.

Indeed, practical obstacles to the ideal implementation of experiments abound throughout social science and even the biological and medical sciences. For example, patients or their clinicians may have such strong treatment preferences that they are recruited and self-select into their treatment group. They will also comply more than average if treated and comply less if not treated (Torgerson and Sibbald 1998). Thus, patient preference trials (PPTs) place patients into (1) willing to randomize plus preferring treatment (2) not willing

to randomize plus preferring treatment, and (3) willing to randomize plus no treatment preference. PPTs were thus conceived to allow randomization of preference and willingness (Torgerson and Sibbald 1998).

These and other practical concerns present obstacles to experimentation in many scenarios, but ethical considerations operate as well. Political scientists and policy evaluators study questions involving, in some cases, sensitive information: subjects' criminal justice records (Patrick and Marsh 2005), self-identified sexual orientation (Buser, et al 2018), and even health records (Gifford, et al 2002). In comparative studies often conducted in developing countries, scholars deploy experimental methods that study the effect of development aid on social cohesion (Fearon et al. 2009), support for combatants during wartime (Lyall, Blair and Imai 2013), and even health worker performance in Ebola outbreak (Christensen et al. 2020); all of which present opportunities for mismanaged or leaked subject-level data, threats to subjects' personal safety, and other concerns raised by ethicists (Haggerty 2004). Regardless of the mechanism, ethical concerns also play a role in motivating observational research in lieu of experimentation. So, we are left with both practical and ethical considerations.

Given the necessity of observational studies — let's say in political science — we are naturally interested in understanding the various threats to causal inference that exist in observational research.[1] One way of approaching this challenge is to compare, or benchmark, observational studies against those obtained via experimental methods. The intuition for this is straightforward: If an experimental finding represents the least biased estimate of some treatment effect, then differences between the experimental results and the estimate obtained through observational study should prompt further investigation. That is, the differences between observational and experimental findings should not be *prima facie* evidence of bias, but should be explainable through some principled argument. First, of course, is the requirement that the benchmarking take place at all.

One key constraint for conducting benchmarking is data availability. While LaLonde (1986) and Dehejia and Wahba (1999) present an exemplary case of observational data and experimental data both answering the same research question and estimating the same causal effect of interest, identifying other cases requires some effort. To begin, the observational data must allow the researcher to credibly argue that the observations are a representative sample of the population. In doing so, the covariates of the observations should be important enough and proportional to the covariates of the population. Additionally, the two sets of data must match, meaning that the coverage and measurements of their covariates should not be too distinct. A final practical consideration is the prevalence of protected health information (PHI) that is heavily regulated by the Health Insurance Portability and Accountability Act (HIPAA) of 1996, and which is rarely publicly available via replication materials.

Benchmarking Applications

Broadly speaking, the combination of data obtained via observational studies or other non-randomized evaluations with experimental data towards the goal of clarifying treatment effects can be understood as benchmarking. Still, there are multiple potential applications that fall under this broader umbrella. One well-known example is the Women's Health Initiative that studied "hormone replacement therapy" using randomized control trials (RCT) as well as observational approaches (Colnet, et al 2020). A discrepancy arose during benchmarking when the experimental data indicated findings that suggested rejection of the therapy as a treatment intervention while the observational data suggested much more positive results. While in this particular case the discrepancy was resolved after a close examination of the implementation protocols (Vandenbroucke 2009; Frieden 2017), we are interested in the general benchmarking process used to identify the discrepancy. In other words, in other applications — be they medical interventions or something situated more firmly in social science — there may not be a convergence between experimental findings and observational effects.

The implied alternative to the WHI scenario described above is, of course, an unresolved discrepancy in results. We suggest that part of a principled resolution to the choices produced by such an alternative (e.g., accepting the validity of the treatment effect, trading off between external validity and internal validity, and so on) is

[1] This is a vast and nuanced question, with applications in many diverse literatures, but see Angrist and Pischke 2010 for a useful summary.

a clear understanding of the difference in estimates between experimental and non-experimental methods. There exist several approaches to obtaining this information via benchmarking, which broadly fall into cases where the observational data has information on treatment assignment and those where treatment is unknown (Colnet, et al 2020). `benchtestr` endeavors to assist with the former scenario as using data without an indication of treatment status requires testing a number of rather advanced assumptions. Within the scenario in which treatment status is known in both the randomized treatment and observational cases, Colnet, et al (2020) outline a distinction between "nested" and "non-nested" designs (p.8) where nested designs deploy a single sampling from a population that is then bifurcated and non-nested designs sample separately for the experimental and observational approaches. `benchtestr` provides data and tools for non-nested analysis.

It is worth briefly mentioning an additional assumption in the `benchtestr` package. In most cases, users will find both observational and experimental data for the same study — for example, both are available for the National Supported Work (NSW) program and the Tennessee Student Teacher Achievement Ratio (STAR) ratio. The primary motivation for within-study comparisons of these datasets is their common origin: the observational and experimental datasets in each study were designed to be comparable. Still, formal methods exist for evaluating the comparability — or the degree of integrability (Yang, et al 2020; Kallus 2018; Wager and Athey 2018) — between observational and experimental datasets.[2] Users interested in more rigorous testing of integrability assumptions will have to conduct evaluations independently.

Data Collection Process

A primary goal of this project was to collect high-quality to be used for experimental benchmarking as described above. The following criteria were used:

Exhibit 1: Criteria used for data collection.

- The experimental design must be causally valid to be considered for use as a benchmark.
- Observational data must be available at the level of the individual subject as individual treatment assignment must be discernible.
- Observational and experimental data must be comparable within-study.

Following the above criteria, and given the authors' time and resource constraints, X data sources were identified and added to the `benchtestr` package (see Exhibit X for a detailed description of each). The data represent a diverse set of research areas including media and political behavior, workforce development program evaluation, and evaluation in education outcomes.

Exhibit 2: Data collected for this project. Below are dataset-level descriptions of the data included in `benchtestr`.

Tennessee Student Teacher Achievement Ratio (STAR) Evaluation

- `Comparison_Students`

A set of students observed during the same time as the longitudinal STAR study. These students are pupils at non-STAR schools (remember that study participation in the STAR study, though observations are collected at the student level) but are intended to be comparable along a number of covariates. A rich array of covariate data is included.

- `STAR_High_Schools`

Data on students at high schools that participated in the STAR study.

- `STAR_K-3_Schools`

Data on students in grades Kindergarten through Grade 3 at schools that participated in the STAR study.

- `STAR_Students`

[2]See Colnet, et al (2020) for a useful summary of these approaches.

Data on all students in school that participated in the STAR study.

- `star_tenn_comparison`

A dataset processed and formatted by the authors to serve as an observational benchmark against the findings in the STAR study. Includes a simplified set of covariates, outcomes of interest, and treatment status.

- `star_tenn_experiment`

A dataset processed and formatted by the authors to serve as the experimental findings in the STAR study. Includes a simplified set of covariates, outcomes of interest, and treatment status.

National Supported Work (NSW) Study

- `lalonde`

A dataset generated from the original Lalonde (1986) study evaluating the NSW program. More documentation available at the source: https://users.nber.org/~rdehejia/data/.nswdata2.html.

- `nsw_dehejia_wahba`

A dataset generated from the Dehejia and Wahba (1999) study evaluating the NSW program. Data have been simplified for use in benchmarking against Lalonde (1986) data. More documentation available at the source: https://users.nber.org/~rdehejia/data/.nswdata2.html.

- `psid_controls_dw`

A set of observations to serve as controls in NSW benchmarking. Generated from the Population Survey of Income Dynamics (PSID). More documentation available at the source: https://users.nber.org/~rdehejia/data/.nswdata2.html.

- `cps_controls_dw`

A set of observations to serve as controls in NSW benchmarking. Generated from the Current Population Survey (CPS). More documentation available at the source: https://users.nber.org/~rdehejia/data/.nswdata2.html.

Miscellaneous

- `ajps`

Data used to generate Table 2 of "Experimental Estimation of Heterogeneous Treatment Effects Related to Self-Selection" (Gaines and Kuklisnki 2011; see references below). A basic experimental dataset containing treatment status (receipt of a negative campaign advertisement) as well as two outcome variables of interest: a thermometer measures for Barack Obama and one for John McCain.

- `roaches_gelman`

A dataset of observations in a study on cockroach abatement first appearing in Gelman and Hill (2007).

Vignette / Examples

Having motivated the creation, utility, and intended uses of `benchtestr`, we would like to demonstrate its application. The vignette below describes three examples, and in so doing reveals a number of the useful functions contained in the package. We encourage users to explore further by using R's built-in help function to learn about the various functions and datasets (e.g., `?benchtestr::match_estimator`).

Example 1: The Kitchen Sink

Let's say you are working on an estimator that you think is effective for estimating the average treatment effect on the treated (i.e. ATT, ATET, etc.) You would like to see how this estimator performs in the experimental setting vs. in the observational setting. What would you do?

Follow a few easy steps:

1. **Examine the documentation on datasets.** Included with `benchtestr`, and made available on the package's Github repo, is documentation describing the data that's shipped with the package. This includes a summary of the accompanying paper's findings (where appropriate), a synopsis of the estimators and estimates in the original analysis, a brief description of the variables, and links to past benchmarking efforts (focused mostly on published academic work, for the time being).
2. **Make dataset selections.** Choose which datasets are most relevant and appropriate, or simply use the default option (all datasets) and examine the output.
3. **Make decision about matching.** Do you want to look at your balance on observed covariates across the selected datasets? Do you want to try different matching approaches? The default is to apply three classical matching techniques and output two summaries of balance – a balance table and a Love plot – for each of the three techniques.
4. **Run the corresponding `benchtestr` functions to get your results.** With your decisions in-hand, all you have to do is pass parameters to a series of `benchtestr` functions.

For example:

```
devtools::install_github('igorgeyn/benchtestr/benchtestr')
library(benchtestr)

### What's in here?
### A few datasets, all of which are documented.

# data('ajps')
data('Comparison_Students')
data('cps_controls_dw')
data('lalonde')
data('nsw_dehejia_wahba')
data('psid_controls_dw')
# data('roaches_gelman')
# data('STAR_High_Schools')
# data('STAR_K-3_Schools')
# data('STAR_Students')
data('star_tenn_comparison')
data('star_tenn_experiment')

# head(ajps)

### Let's look at the NSW study.

head(lalonde)
```

Code Chunk 1: Enumeration and exploration of available data.

```
##   nsw age educ black hisp married re74 re75 re78 u74 u75 u78
## 1   0  47   12     0    0       0    0    0    0   1   1   1
## 2   0  50   12     1    0       1    0    0    0   1   1   1
## 3   0  44   12     0    0       0    0    0    0   1   1   1
## 4   0  28   12     1    0       1    0    0    0   1   1   1
## 5   0  54   12     0    0       1    0    0    0   1   1   1
## 6   0  55   12     0    1       1    0    0    0   1   1   1
```

```
head(nsw_dehejia_wahba)
```

```
##                data_id treat age education black hispanic married nodegree re74
## 1 Dehejia-Wahba Sample     1  37        11     1        0       1        1    0
## 2 Dehejia-Wahba Sample     1  22         9     0        1       0        1    0
## 3 Dehejia-Wahba Sample     1  30        12     1        0       0        0    0
## 4 Dehejia-Wahba Sample     1  27        11     1        0       0        1    0
## 5 Dehejia-Wahba Sample     1  33         8     1        0       0        1    0
## 6 Dehejia-Wahba Sample     1  22         9     1        0       0        1    0
##   re75       re78
## 1    0  9930.0459
## 2    0  3595.8940
## 3    0 24909.4492
## 4    0  7506.1460
## 5    0   289.7899
## 6    0  4056.4939
```

```
head(cps_controls_dw)
```

```
##   data_id treat age education black hispanic married nodegree      re74
## 1    CPS1     0  45        11     0        0       1        1 21516.670
## 2    CPS1     0  21        14     0        0       0        0  3175.971
## 3    CPS1     0  38        12     0        0       1        0 23039.020
## 4    CPS1     0  48         6     0        0       1        1 24994.369
## 5    CPS1     0  18         8     0        0       1        1  1669.295
## 6    CPS1     0  22        11     0        0       1        1 16365.760
##        re75      re78 source
## 1 25243.551 25564.670   cps1
## 2  5852.565 13496.080   cps1
## 3 25130.760 25564.670   cps1
## 4 25243.551 25564.670   cps1
## 5 10727.610  9860.869   cps1
## 6 18449.270 25564.670   cps1
```

```
head(psid_controls_dw)
```

```
##   data_id treat age education black hispanic married nodegree re74 re75 re78
## 1    PSID     0  47        12     0        0       0        0    0    0    0
## 2    PSID     0  50        12     1        0       1        0    0    0    0
## 3    PSID     0  44        12     0        0       0        0    0    0    0
## 4    PSID     0  28        12     1        0       1        0    0    0    0
## 5    PSID     0  54        12     0        0       1        0    0    0    0
## 6    PSID     0  55        12     0        1       1        0    0    0    0
##   source
## 1  psid1
## 2  psid1
## 3  psid1
## 4  psid1
## 5  psid1
## 6  psid1
```

Having identified and examined — at least, at a quick glance — the data available in `benchtestr`, how do we move on to using it for actual benchmark evaluation? One place to start is with two classical and relatively straightforward estimator: the difference in means (DIM) estimator and the linear regression estimator, which we will also refer to as the lm estimator owing to its function call in the R language. Below, you will find code that implements and then compares both of these estimators.

```
## Generate estimator-based comparisons
## for a single study:

?benchtestr::dim_estimator
```

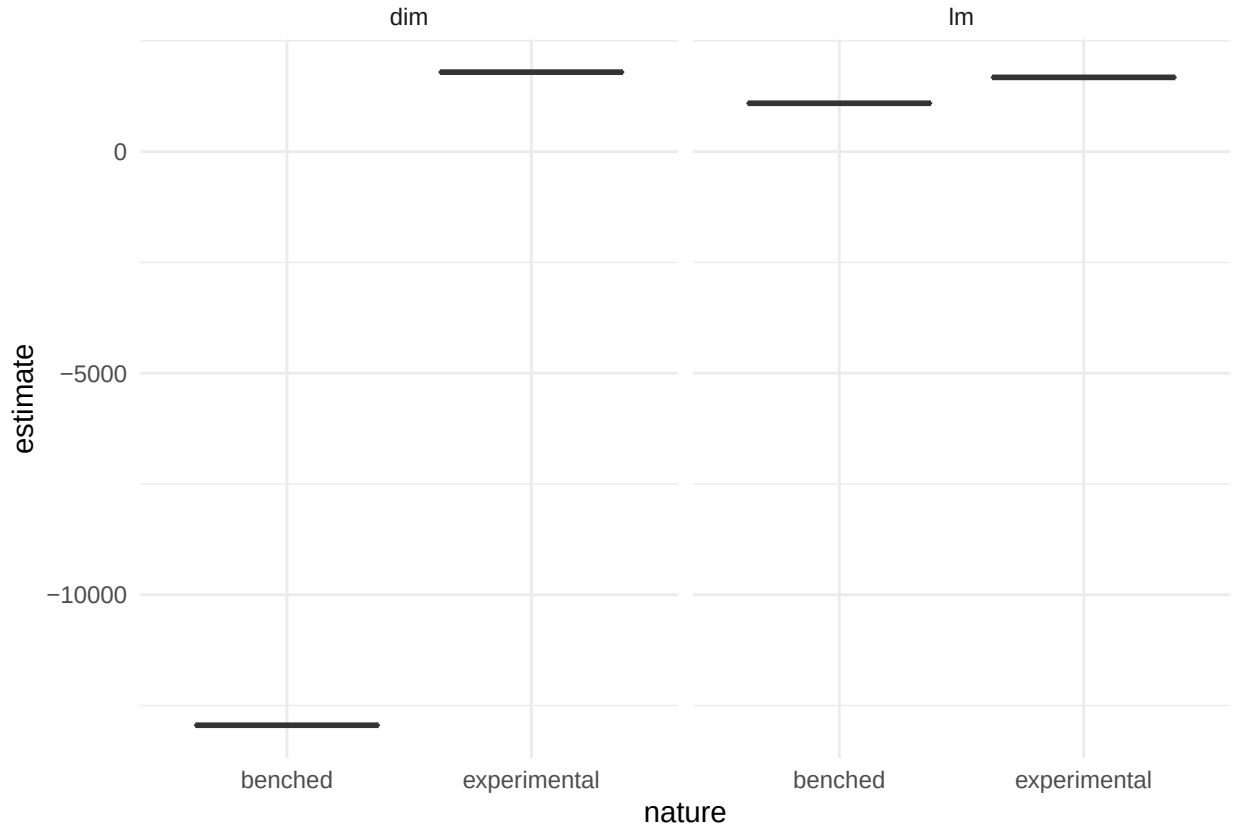
Code Chunk 2: Running and comparing linear regression and difference in means (DIM) estimators.

```
## starting httpd help server ... done
```

```
?benchtestr::lm_estimator
?benchtestr::match_estimator
?benchtestr::dim_estimator_multi
# Save space for matching
# ?benchtestr::iv_estimator ## Save for later

# Generate DIM estimate
dim_output <-
  dim_estimator(df_exp = nsw_dehejia_wahba, df_base = psid_controls_dw,
                treatment = 'treat', outcome = 're78')
```

```
## Loading required package: estimatr
```

```
# Generate regression estimate
lm_output <-
  lm_estimator(df_exp = nsw_dehejia_wahba, df_base = psid_controls_dw,
               treatment = 'treat', outcome = 're78') %>% filter(term == 'treat')

# Quickly compare estimates across estimators
comparison_df <- rbind(
  dim_output %>% select('estimator', 'nature', 'estimate', 'outcome'),
  lm_output %>% filter(term == 'treat') %>% select('estimator', 'nature', 'estimate', 'outcome')
  )

# Plot the different estimates
ggplot(data = comparison_df) +
  geom_boxplot(aes(x = nature, y = estimate)) +
  facet_wrap(.~estimator) +
  theme_minimal()
```



Example 2: The Balance Check (and Match-based Estimator)

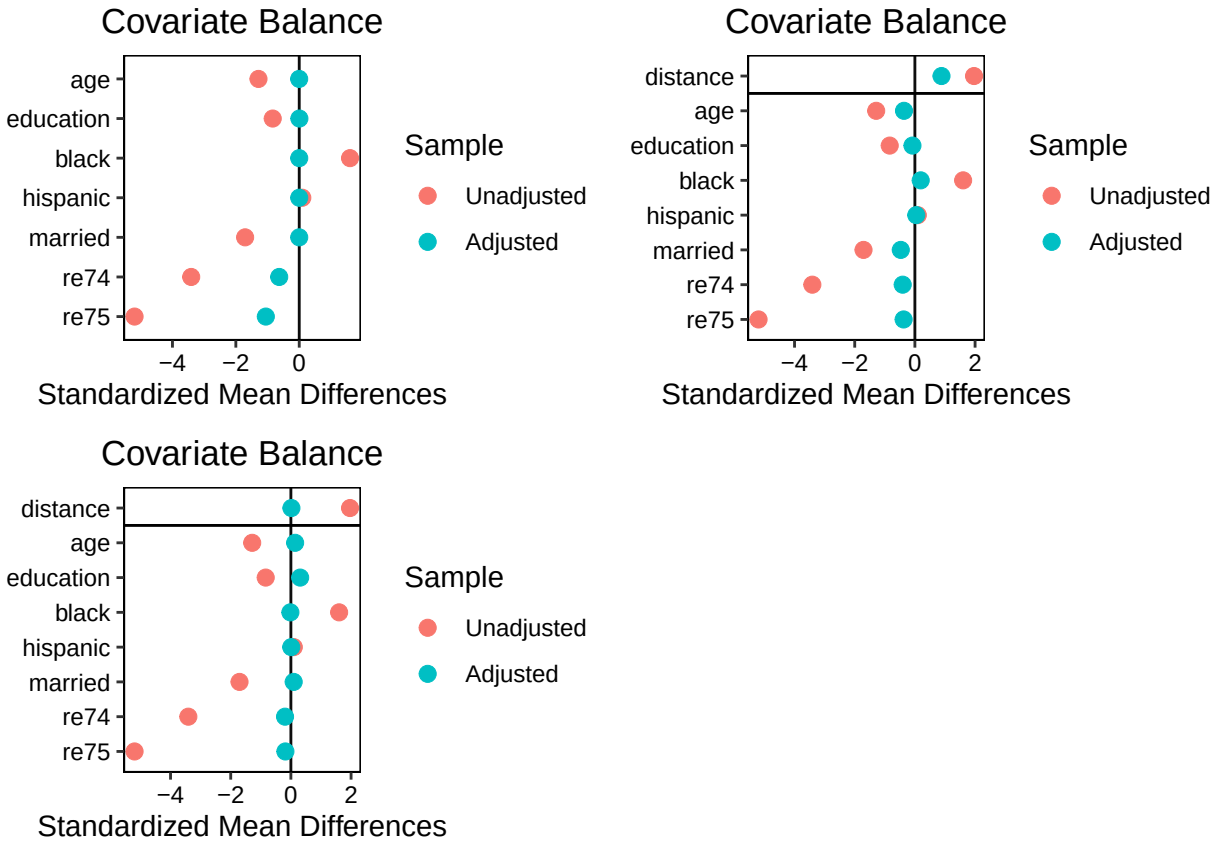
In the previous example, we compared a few different estimators' performance across several datasets. Already, we saw that two estimators that should yield identical, or at least similar, results (the linear regression and the DIM) in fact yielded different estimates. In other situations, these differences could be even more stark.

We can use `benchtestr` to explore the possibility of covariate imbalance, opportunities for matching-based estimation, and finally some comparison to non-matching estimators (e.g., DIM and linear regression).

```
### Of course, you are welcome to use MatchIt and other packages for
### manual matching and evaluation.
### But here is the `benchtestr` case:

## Let's say you've been running your estimator(s) on multiple datasets.
## You can test them all at the same time by passing them
## as arguments to a function.

# Set up your inputs
nsw_formula_reduced = treat ~ age + education + black + hispanic + married + re74 + re75
covars_list = c('age', 'education', 'black', 'hispanic', 'married', 're74', 're75')

# And then run the function with a few additional arguments
# Note that the function takes extra arguments, so you can do things like
# specify the number of subclasses (via subclass = X), and so on.
balance_assess(df_exp = nsw_dehejia_wahba, df_bench = psid_controls_dw,
```

```
              treat = 'treat', outcome = 're78', covars = covars_list,
              formula_arg = nsw_formula_reduced)
```

Code Chunk 3: Running a balance test and matching estimator ont he National Supported Work (NSW) study.




```
## Having observed imbalance in the covariates, you can do a number of things:

## 1) Settle on your preferred matching method and proceed with your analysis outside of `benchtestr`.

## 2) Extract the matched datasets and run classical estimators within `benchtestr` (i.e.
##    DIM, lm estimator, etc.).

## 3) Run the following function to get an estimate using `benchtestr` on each of the matched datasets.

match_estimator(df_exp = nsw_dehejia_wahba, df_bench = psid_controls_dw,
              treatment = 'treat', outcome = 're78', covars = covars_list,
              formula_arg = nsw_formula_reduced, reg_formula = re78 ~ treat)

m_out_df <- m_out_df %>% rename(match_method = V1, estimate = V2)
kableExtra::kable(x = m_out_df,
                  caption = 'Estimates by Match Method',
                  format = 'latex'
                  )
```

Table 1: Estimates by Match Method

	match_method	estimate
m_out_effect_est	cem	-2742.91
m_out_effect_est.1	nearest	100.7245
m_out_effect_est.2	full	-14409.07
m_out_effect_est.3	subclass	-14409.07

Example 3: Systematic Plotting

Finally, we can consider the case in which we're taking a granular look at a single dataset. We will take the NSW data as an example. From previous examples in the vignette, we know that we can simply apply individual estimators on a one-off basis. However, observing differences between pure experimental estimates and those benchmarked against observational data will yield the most comprehensive comparison.

Let us do this now.

Code Chunk 4a: Generating and comparing estimates using multiple estimators. The code below performs a difference-in-means (DIM) estimate, linear regression estimate, and then finally four match-based DIM estimates (using coarsened exact matching, optimal matching, nearest, and subclassification) on the NSW dataset using PSID controls for benchmarking. While users can specify all function arguments in the function call, we demonstrate how arguments can also be defined outside the call (e.g., by setting the list of covariates).

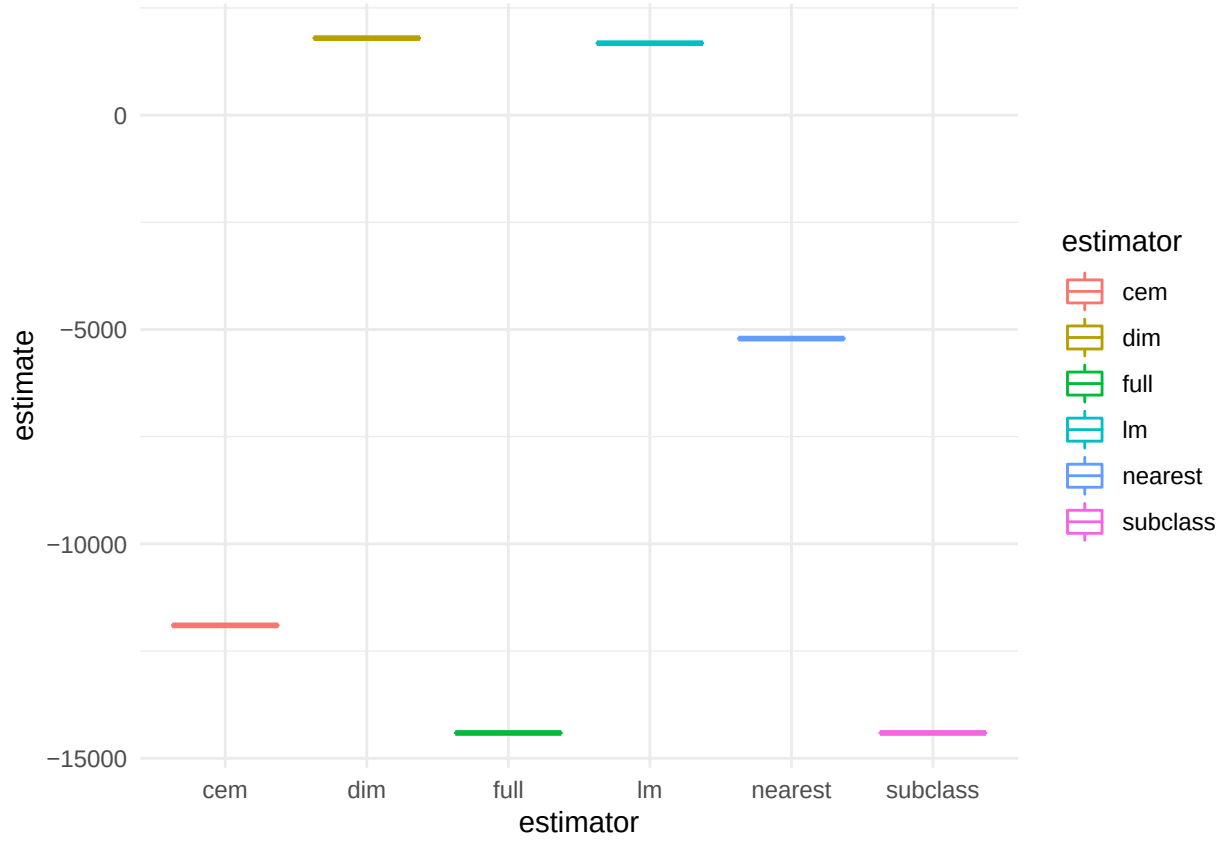
```
### Analyzing the case where we are only interested in matching on basic demographics

covars_list <- c('age', 'education', 'black', 'hispanic')

compar_ests(dim = 1, lm = 1, match = 1, df_exp = nsw_dehejia_wahba, df_base = psid_controls_dw,
            treatment = 'treat', outcome = 're78',
            covars = c('age', 'education', 'black', 'hispanic'),
            formula_arg = treat ~ age + education + black + hispanic,
            reg_formula = re78 ~ treat
            )
```

```
##     estimator       nature  term   estimate    conf.low  conf.high outcome
## 1:1       dim experimental treat   1794.342    474.0105   3114.674    re78
## 1:2       dim      benched treat -12945.249 -14221.6458 -11668.852    re78
## 2:1        lm experimental treat   1676.343    345.6483   3007.037    re78
## 2:2        lm      benched treat   1092.156   -222.2527   2406.566    re78
## 3:1       cem         <NA>  <NA> -11899.966          NA         NA    <NA>
## 3:2   nearest         <NA>  <NA>  -5211.862          NA         NA    <NA>
## 3:3      full         <NA>  <NA> -14409.074          NA         NA    <NA>
## 3:4  subclass         <NA>  <NA> -14409.074          NA         NA    <NA>
```

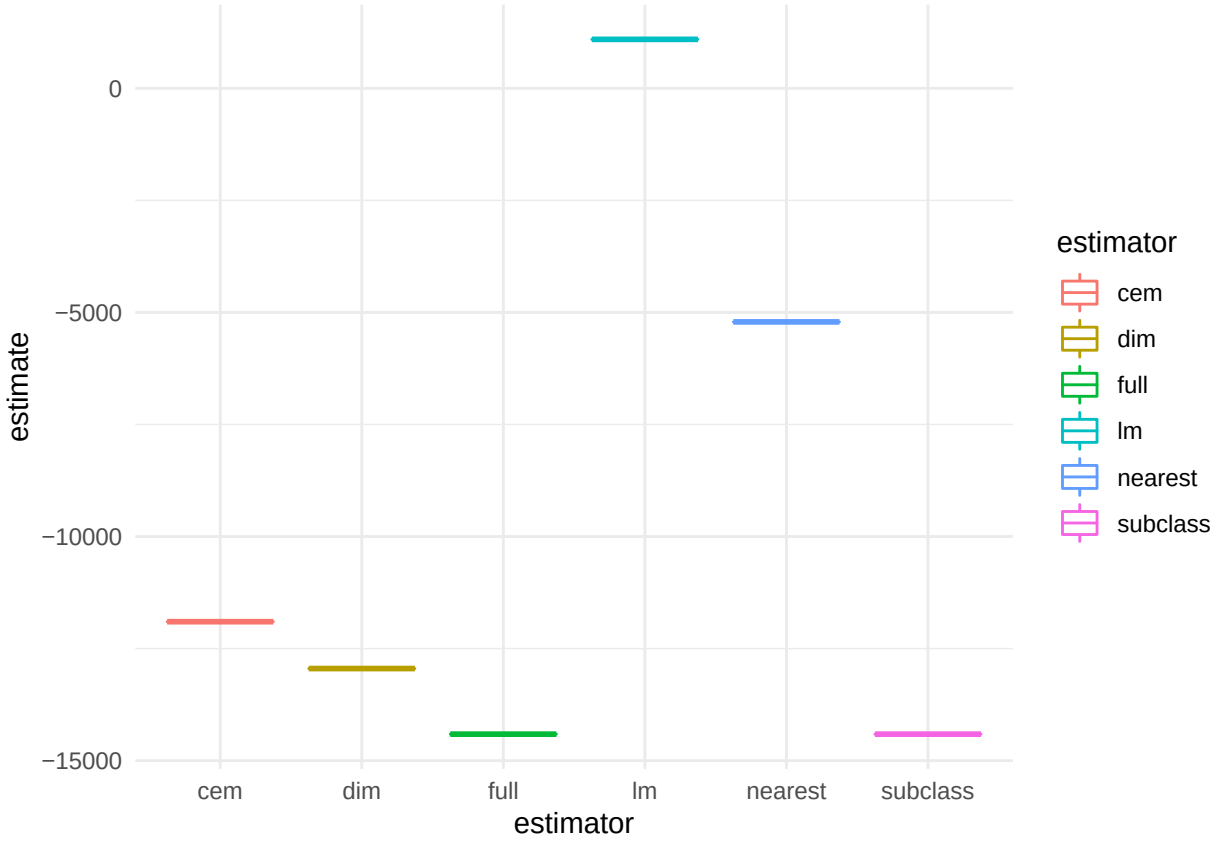
Code Chunk 4b: Manipulating data comparing multi-estimator estimates. We can now examine our comparison data, which is output from the `compare_ests` function as a datafame (df) object. An R user familiar with tidyverse plotting via `ggplot` will see how simply one can visualize within estimator categories (i.e. within experimental estimates or within benchmarked datasets) as well as across categories.

```
compare_out <-
  compar_ests(dim = 1, lm = 1, match = 1, df_exp = nsw_dehejia_wahba, df_base = psid_controls_dw,
              treatment = 'treat', outcome = 're78',
              covars = c('age', 'education', 'black', 'hispanic'),
              formula_arg = treat ~ age + education + black + hispanic,
```

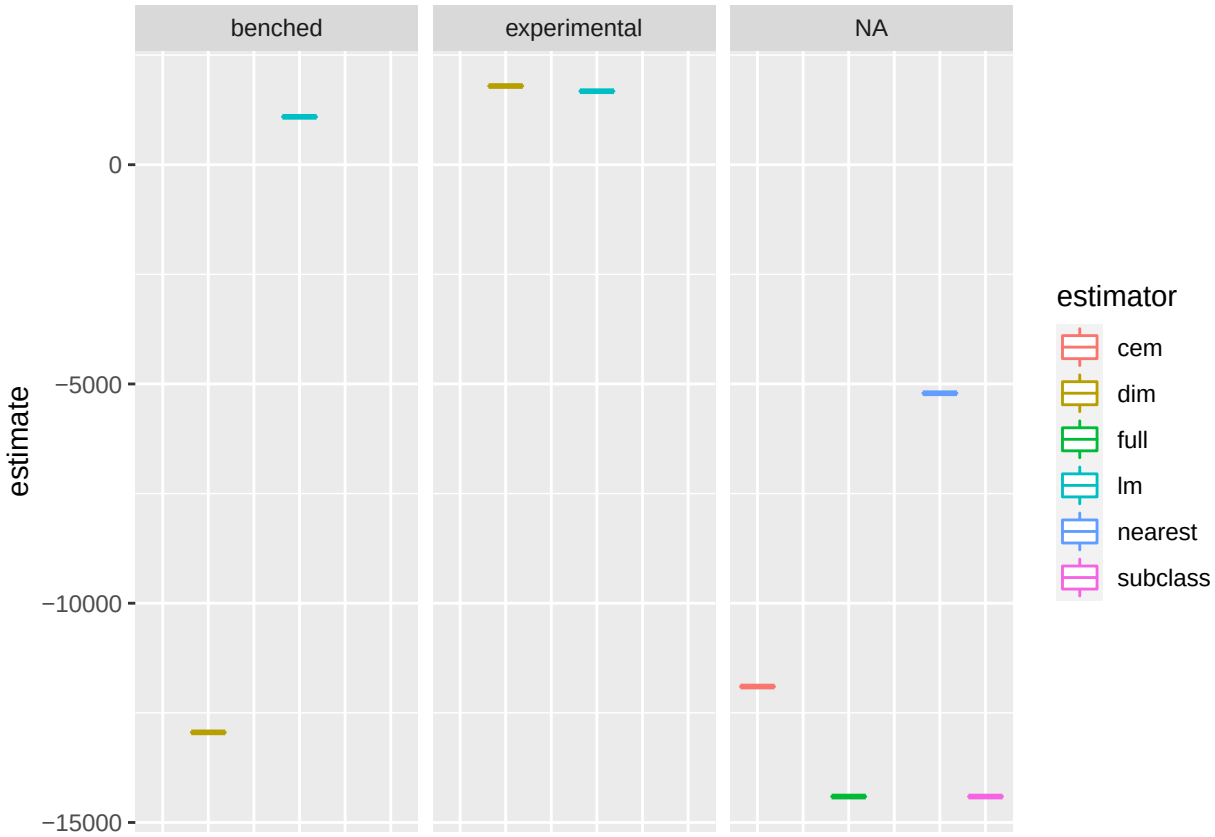
```
              reg_formula = re78 ~ treat)
# compare_out

## Plot just the experimental results with matched results

ggplot(compare_out %>% filter(nature == ('experimental') | is.na(nature))) +
  geom_boxplot(aes(x = estimator, y = estimate, color = estimator)) + theme_minimal()
```




```
## Plot just the obs. benchmark results with matched results

ggplot(compare_out %>% filter(nature == ('benched') | is.na(nature))) +
  geom_boxplot(aes(x = estimator, y = estimate, color = estimator)) + theme_minimal()
```




```
## plot all three

ggplot(compare_out) +
  geom_boxplot(aes(x = estimator, y = estimate, color = estimator)) +
  facet_wrap(.~nature) +
  theme(axis.title.x = element_blank(), axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



Discussion

`benchtestr` is a package designed to put information into the hands of researchers interested in causal inference. Past research (e.g., Colnet, et al 2020) shows the challenges of interpreting when observational and experimental data can be combined, the difficulty of assessing whether a nested or non-nested approach is appropriate, and the sensitivity of such benchmark analyses to matching techniques. Thus, neither this paper nor `benchtestr` purport to resolve the complicated process of drawing conclusions from benchmarking. Instead, we equip the user with data and estimators to answer the basic question of whether benchmarking *could* be a fruitful process. Using a variety of tools, including visualization and simple comparisons across the estimators, the researcher can then decide whether additional evauation, assumption testing, etc. is appropriate.

Acknowledgments

The authors would like to thank Professor Chad Hazlett for a helpful discussion that generated the idea for this package, and both Chad Hazlett and Ciara Sterbenz (PhD Student, UCLA) for instruction during the quarter that spawned `benchtestr`. We take full credit for poorly written R functions.

Appendices

Appendix A: Technical Documentation

`benchtestr` is available on Github via the following repository: LINK LINK LINK LINK. It is maintained under the MIT License.

Appendix B: Implementation Code

This appendix contains additional code that explores various features of `benchtestr`. It is also available as part of a standalone vignette on Github at: LINK LINK LINK LINK.

The chunk below uses data from a longitudinal study on classroom sizes and school outcomes conducted in Tennessee in the 1990s, known as the Student-Teacher Achievement Ratio (STAR) study. It applies a DIM estimator to numerous potential outcomes of interest, representing a case where the user might be interested in numerous outcomes.

Code Chunk B.1: Applying the DIM estimator to multiple outcomes of interest in the Tennessee Student-Teacher Achievement Ratio (STAR) data. Let's replicate the above analysis for a different study.

```
data("star_tenn_comparison")
data("star_tenn_experiment")

## Treatment Data

tenn_star_df <- tenn_star_df %>% mutate(treat = case_when(yearsstar == 4 ~ 1, yearsstar < 4 ~ 0))
tenn_compar_df <- tenn_star_df %>% mutate(treat = 0)

# kindergarten
tenn_star_df %>% group_by(treat) %>%
  summarise(avg_read_k = mean(gktreadss, na.rm = TRUE),
            avg_math_k = mean(gktmathss, na.rm = TRUE),
            avg_words_k = mean(gkwordskillss, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   treat avg_read_k avg_math_k avg_words_k
##   <dbl>      <dbl>      <dbl>       <dbl>
## 1     0       429.       473.        426.
## 2     1       444.       498.        442.
```

```
# 1st g
tenn_star_df %>% group_by(treat) %>%
  summarise(avg_read_1g = mean(g1treadss, na.rm = TRUE),
            avg_math_1g = mean(g1tmathss, na.rm = TRUE),
            avg_words_1g = mean(g1wordskillss, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   treat avg_read_1g avg_math_1g avg_words_1g
##   <dbl>       <dbl>       <dbl>        <dbl>
## 1     0        506.        519.         500.
## 2     1        538.        543.         530.
```

```
# 2nd g
tenn_star_df %>% group_by(treat) %>%
  summarise(avg_read_2d = mean(g2treadss, na.rm = TRUE),
            avg_math_2d = mean(g2tmathss, na.rm = TRUE),
            avg_words_2g = mean(g2wordskillss, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   treat avg_read_2d avg_math_2d avg_words_2g
##   <dbl>       <dbl>       <dbl>        <dbl>
## 1     0        572.        570.         571.
## 2     1        596.        591.         596.
```

```
## Comparison (Control) Data

# kindergarten
# comparison data begins in 1st grade

# 1st g
tenn_compar_df %>%
  # group_by(treat) %>%
  summarise(avg_read_1g = mean(g1treadss, na.rm = TRUE),
            avg_math_1g = mean(g1tmathss, na.rm = TRUE),
            avg_words_1g = mean(g1wordskillss, na.rm = TRUE))
```

```
##   avg_read_1g avg_math_1g avg_words_1g
## 1    520.7873    530.5279     513.4357
```

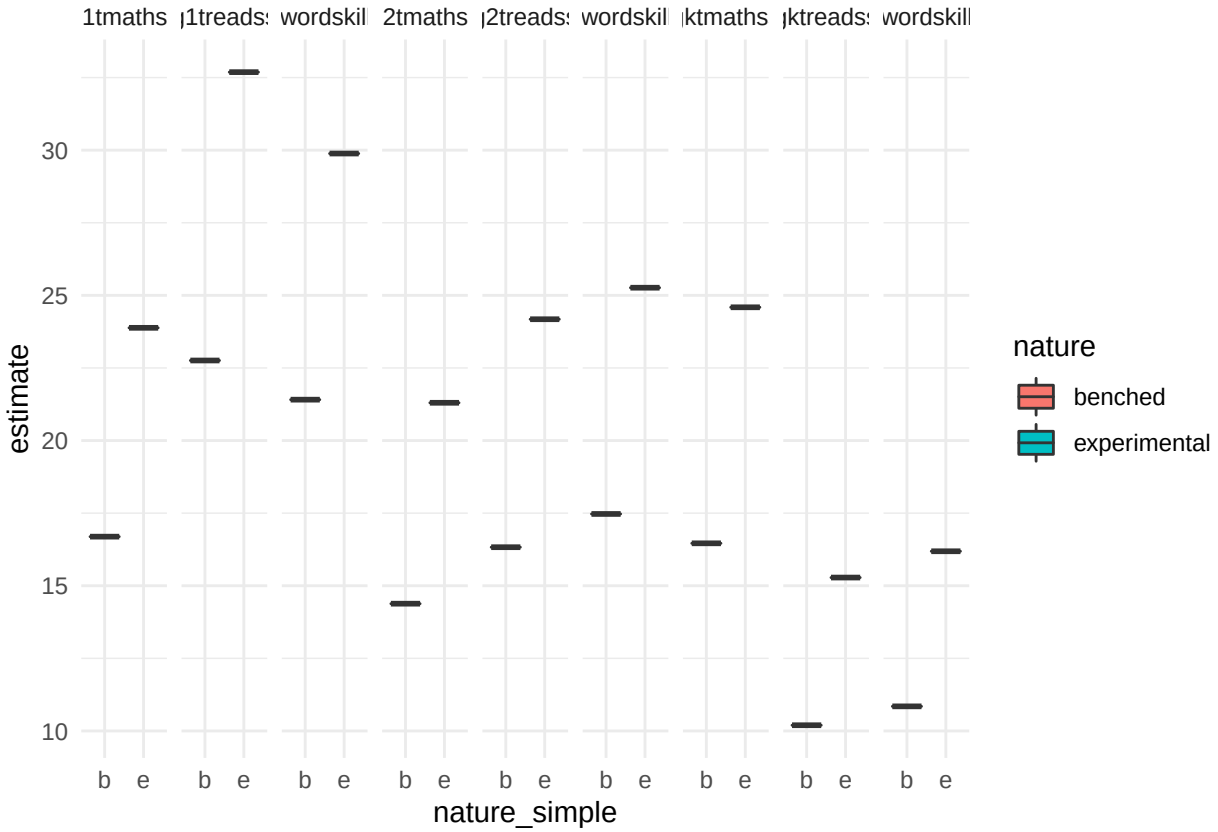
```
# 2nd g
tenn_compar_df %>%
  # group_by(treat) %>%
  summarise(avg_read_2d = mean(g2treadss, na.rm = TRUE),
            avg_math_2d = mean(g2tmathss, na.rm = TRUE),
            avg_words_2g = mean(g2wordskillss, na.rm = TRUE))
```

```
##   avg_read_2d avg_math_2d avg_words_2g
## 1    583.9348    580.6125     582.9855
```

```
### Start with DIM and regression estimates.

dim_output_star <- dim_estimator(df_exp = tenn_star_df, df_base = tenn_compar_df,
                                 treatment = 'treat', outcome = 'g2treadss')
dim_output_star
```

```
##   estimator       nature  term estimate std.error statistic      p.value
## 1       dim experimental treat 24.17786 1.1401602  21.20567 2.473852e-96
## 3       dim      benched treat 16.32763 0.9445075  17.28692 3.824807e-65
##   conf.low conf.high       df   outcome
## 1 21.94274  26.41298 6065.450 g2treadss
## 3 14.47600  18.17926 5189.959 g2treadss
```

```
## Plotting multi

dim_estimator_multi(df_exp = tenn_star_df, df_base = tenn_compar_df,
                    treatment = 'treat',
                    # outcomes_list = c('g2treadss', 'g2tmathss', 'g2wordskillss'),
                    outcomes_list = c('g1treadss', 'g1tmathss' ,
                                      'g1wordskillss', 'g2treadss','g2tmathss', 'g2wordskillss',
                                      'gkwordskillss', 'gktreadss', 'gktmathss'
                                      ),
                    plot = 1) ## toggles on and off (turn off for a df)
```



Assuming you decided to use `benchtestr` to the fullest extent (i.e. you ran Step 3 above), you should now have a table of matching-based estimates of the observational treatment effect using your observational benchmarking data. This can be compared against your experiment benchmark – that is, against the estimate of the treatment effect using whatever estimator you brought to bear on the experimental data.

Depending on your results, you can either be satisfied that your observational approach yields a relatively similar results when compared to experimental findings, or you might engage in additional analysis. Currently, `benchtestr` is not equipped for additional analysis, so these next steps would need to take place in another environment.

References

Publications

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Journal of economic perspectives, 24(2), 3-30.

Brian J. Gaines and James H. Kuklinski, (2011), "Experimental Estimation of Heterogeneous Treatment Effects Related to Self-Selection," American Journal of Political Science 55(3): 724-736, doi:10.1111/j.1540-5907.2011.00518.x.

Anastasopoulos, L. Jason. 2020. "Principled Estimation of Regression Discontinuity Designs." arXiv:1910.06381 [econ, stat]. http://arxiv.org/abs/1910.06381 (April 13, 2021).

"Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis | NBER." https://www.nber.org/papers/w22791 (April 22, 2021).

Benson, Kjell, and Arthur J. Hartz. 2000. "A Comparison of Observational Studies and Randomized, Controlled Trials." New England Journal of Medicine 342(25): 1878–86.

Brad. 2012. "Available Public Use Files." MDRC. https://www.mdrc.org/available-public-use-files (May 27, 2021).

Buser, T., Geijtenbeek, L., & Plug, E. (2018). Sexual orientation, competitiveness and income. Journal of Economic Behavior & Organization, 151, 191-198. "CausalImpact." https://google.github.io/CausalImpact/CausalImpact.html (April 30, 2021).

Christensen, Darin, et al. "Building Resilient Health Systems: Experimental Evidence from Sierra Leone and the 2014 Ebola Outbreak." University of Chicago, Becker Friedman Institute for Economics Working Paper 2020-28 (2020).

Colnet, Bénédicte et al. 2020. "Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review." https://hal.archives-ouvertes.fr/hal-03008276 (May 3, 2021).

Cranmer, Skyler J. "Introduction to the Virtual Issue: Machine Learning in Political Science." : 9.

D'Amour, Alexander et al. 2021. "Overlap in Observational Studies with High-Dimensional Covariates." Journal of Econometrics 221(2): 644–54.

Fearon, James D., Macartan Humphreys, and Jeremy M. Weinstein. "Can development aid contribute to social cohesion after civil war? Evidence from a field experiment in post-conflict Liberia." American Economic Review 99.2 (2009): 287-91.

Frieden, Thomas R. "Evidence for health decision making—beyond randomized, controlled trials." New England Journal of Medicine 377.5 (2017): 465-475.

Gerber, Alan S., & Green, Donald P. (2005). Correction to Gerber and Green (2000), Replication of Disputed Findings, and Reply to Imai (2005). American Political Science Review, 99(2): 301–313.

Gerber, Alan S., and Donald P. Green. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." The American Political Science Review 94.3 (2000): 653-663.

Gifford, A. L., Cunningham, W. E., Heslin, K. C., Andersen, R. M., Nakazono, T., Lieu, D. K., ... & Bozzette, S. A. (2002). Participation in research and access to experimental treatments by HIV-infected patients. New England Journal of Medicine, 346(18), 1373-1382.

Haggerty, K. D. (2004). Ethics creep: Governing social science research in the name of ethics. Qualitative sociology, 27(4), 391-414.

Imai, Kosuke. "Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." American Political Science Review 99.2 (2005): 283-300.

Iyengar, S., & Kinder, D. R. (1987). American politics and political economy. News that matters: Television and American opinion. University of Chicago Press.

Kallus, N., X. Mao, and M. Udell (2018). Causal inference with noisy and missing covariates via matrix factorization. In Advances in neural information processing systems, pp. 6921–6932.

"Lalonde: The Lalonde Data Set in Sbw: Stable Balancing Weights for Causal Inference and Estimation with Incomplete Outcome Data." https://rdrr.io/cran/sbw/man/lalonde.html (April 30, 2021).

Lyall, Jason, Graeme Blair, and Kosuke Imai. "Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan." American Political Science Review 107.4 (2013): 679-705.

Patrick, S., & Marsh, R. (2005). Juvenile diversion: Results of a 3-year experimental study. Criminal Justice Policy Review, 16(1), 59-73.

Reis, Sally M. et al. 2008. "Using Enrichment Reading Practices to Increase Reading Fluency, Comprehension, and Attitudes." The Journal of Educational Research 101(5): 299–315.

Schatz, Edward, and Irwin J. Schatz. 2003. "Medicine and Political Science: Parallel Lessons in Methodological Excess." PS: Political Science and Politics 36(3): 417–22.

Soriano, Dan et al. 2021. "Interpretable Sensitivity Analysis for Balancing Weights." arXiv:2102.13218 [stat]. http://arxiv.org/abs/2102.13218 (April 30, 2021).

Torgerson, David J., and Bonnie Sibbald. "Understanding controlled trials. What is a patient preference trial?." BMJ: British Medical Journal 316.7128 (1998): 360.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113 (523), 1228–1242.

Vandenbroucke, Danny, et al. "A Network Perspective on Spatial Data Infrastructures: Application to the Sub-national SDI of Flanders (Belgium)." Transactions in GIS 13 (2009): 105-122.

Yang, Shu, Donglin Zeng, and Xiaofei Wang. "Elastic Integrative Analysis of Randomized Trial and Real-World Data for Treatment Heterogeneity Estimation." arXiv preprint arXiv:2005.10579 (2020).

software

`GK2011`, Gaines and Kuklinski (2011) Estimators for Hybrid Experiments, *Github user*: `leeper`, *Name on Github*: Thomas J. Leeper. https://github.com/leeper/GK2011.