Rough Draft of the `benchtestr` vignette for 200C - Spring 2021

Igor Geyn

5/26/2021

Introduction

This document serves to introduce the R package `benchtestr`. It is co-authored and co-mainted by Igor Geyn (PhD Student, Department of Political Science UCLA) and Shing Hon Lam (PhD Student, Department of Political Science UCLA). Original inspiration for the project comes from Professor Chad Hazlett (UCLA Political Science and Statistics) who taught 200C with Ciara Sterbenz in Spring 2021.

Purpose

As has been made famous by Dehejia and Wahba (1999), [citation], and [citation], benchmarking experimental findings with observational data is an important and informative exercise. 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.

Vignette / Examples

The best way to learn `benchtestr` is by doing. Let's take a look at a couple of scenarios.

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 anlaysis, 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 technques.

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')
```

```
## Skipping install of 'benchtestr' from a github remote, the SHA1 (bfb8ed81) has not changed since las
##   Use `force = TRUE` to force installation
```

```
library(benchtestr)
# library(rlang)

### 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)
```

```
##   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
```

```
## What can we do?
```

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

?benchtestr::dim_estimator
```

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

```
?benchtestr::lm_estimator
# 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
```

```
dim_output
```

```
##   estimator       nature  term   estimate std.error  statistic      p.value
## 1       dim experimental treat   1794.342  670.9965   2.674145 7.892978e-03
## 3       dim      benched treat -12945.249  648.5126 -19.961446 2.362821e-56
##      conf.low  conf.high       df outcome
## 1    474.0105   3114.674 307.1325    re78
## 3 -14221.6458 -11668.852 289.5199    re78
```

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

```
##           term      estimate    std.error   statistic      p.value
## 1  (Intercept)  7.850614e+02 3.306397e+03  0.23743712 8.124295e-01
## 2        treat  1.676343e+03 6.770491e+02  2.47595413 1.366768e-02
## 3          age  5.531668e+01 4.140877e+01  1.33586856 1.822907e-01
## 4    education  3.957343e+02 1.982766e+02  1.99587002 4.657111e-02
## 5        black -2.159522e+03 1.020849e+03 -2.11541758 3.496220e-02
## 6     hispanic  1.640327e+02 1.381731e+03  0.11871536 9.055557e-01
## 7      married -1.387253e+02 8.787522e+02 -0.15786622 8.746355e-01
## 8     nodegree -7.068064e+01 1.025594e+03 -0.06891676 9.450875e-01
## 9         re74  8.214121e-02 1.107324e-01  0.74179919 4.586093e-01
## 10        re75  5.276410e-02 1.277466e-01  0.41303739 6.797827e-01
## 21 (Intercept)  3.714487e+02 1.739569e+03  0.21352915 8.309285e-01
## 22       treat  1.092156e+03 6.703634e+02  1.62920072 1.033737e-01
## 23         age -8.301101e+01 1.858407e+01 -4.46678210 8.228725e-06
## 24   education  5.493741e+02 1.111098e+02  4.94442574 8.049382e-07
## 25       black -8.132900e+02 4.009504e+02 -2.02840534 4.260554e-02
## 26    hispanic  1.555913e+03 9.154442e+02  1.69962611 8.930314e-02
## 27     married  1.194791e+03 4.358437e+02  2.74132891 6.154770e-03
## 28    nodegree  3.586720e+02 5.473725e+02  0.65526126 5.123489e-01
## 29        re74  2.780944e-01 5.821839e-02  4.77674488 1.865578e-06
## 30        re75  5.713067e-01 6.252102e-02  9.13783281 1.131360e-19
##         conf.low    conf.high   df outcome       nature
## 1  -5713.4388019 7283.5616487  435    re78 experimental
## 2    345.6482818 3007.0369690  435    re78 experimental
## 3    -26.0694668  136.7028197  435    re78 experimental
## 4      6.0350622  785.4335480  435    re78 experimental
## 5  -4165.9321043 -153.1122108  435    re78 experimental
## 6  -2551.6663733 2879.7317684  435    re78 experimental
## 7  -1865.8533443 1588.4027714  435    re78 experimental
## 8  -2086.4171844 1945.0558986  435    re78 experimental
## 9     -0.1354958    0.2997783  435    re78 experimental
## 10    -0.1983131    0.3038413  435    re78 experimental
## 21 -3039.3962312 3782.2936365 3053    re78      benched
## 22  -222.2526697 2406.5656188 3053    re78      benched
## 23  -119.4495701  -46.5724471 3053    re78      benched
## 24   331.5165747  767.2317198 3053    re78      benched
## 25 -1599.4501073  -27.1299208 3053    re78      benched
## 26  -239.0364135 3350.8622617 3053    re78      benched
## 27   340.2142030 2049.3678057 3053    re78      benched
```

```
## 28  -714.5838855 1431.9278701 3053    re78      benched
## 29     0.1639432    0.3922456 3053    re78      benched
## 30     0.4487191    0.6938942 3053    re78      benched
##                                                                  control
## 1  age + education + black + hispanic + married + nodegree + re74 + re75
## 2  age + education + black + hispanic + married + nodegree + re74 + re75
## 3  age + education + black + hispanic + married + nodegree + re74 + re75
## 4  age + education + black + hispanic + married + nodegree + re74 + re75
## 5  age + education + black + hispanic + married + nodegree + re74 + re75
## 6  age + education + black + hispanic + married + nodegree + re74 + re75
## 7  age + education + black + hispanic + married + nodegree + re74 + re75
## 8  age + education + black + hispanic + married + nodegree + re74 + re75
## 9  age + education + black + hispanic + married + nodegree + re74 + re75
## 10 age + education + black + hispanic + married + nodegree + re74 + re75
## 21 age + education + black + hispanic + married + nodegree + re74 + re75
## 22 age + education + black + hispanic + married + nodegree + re74 + re75
## 23 age + education + black + hispanic + married + nodegree + re74 + re75
## 24 age + education + black + hispanic + married + nodegree + re74 + re75
## 25 age + education + black + hispanic + married + nodegree + re74 + re75
## 26 age + education + black + hispanic + married + nodegree + re74 + re75
## 27 age + education + black + hispanic + married + nodegree + re74 + re75
## 28 age + education + black + hispanic + married + nodegree + re74 + re75
## 29 age + education + black + hispanic + married + nodegree + re74 + re75
## 30 age + education + black + hispanic + married + nodegree + re74 + re75
##    estimator
## 1         lm
## 2         lm
## 3         lm
## 4         lm
## 5         lm
## 6         lm
## 7         lm
## 8         lm
## 9         lm
## 10        lm
## 21        lm
## 22        lm
## 23        lm
## 24        lm
## 25        lm
## 26        lm
## 27        lm
## 28        lm
## 29        lm
## 30        lm
```

```
# 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')
  )
comparison_df
```

```
##    estimator       nature   estimate outcome
## 1        dim experimental   1794.342    re78
```

```
## 3        dim      benched -12945.249    re78
## 11        lm experimental   1676.343    re78
## 2         lm      benched   1092.156    re78
```

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

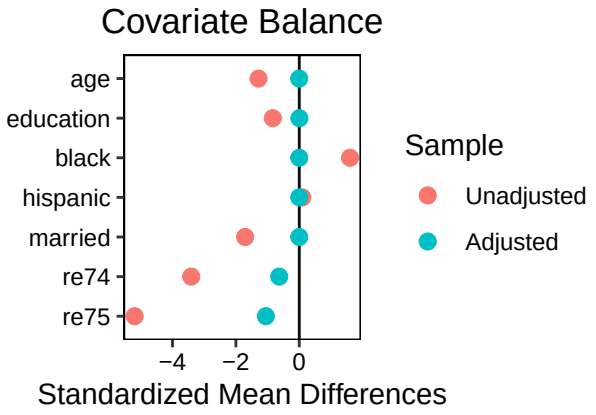
In the previous example, we compared a few different estimators' performance across several datasets. Already, we saw that two estimators that should yield identical, or at least similar, results (the linear regression and the DIM) in fact yielded different estimates. In other situations, these differences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

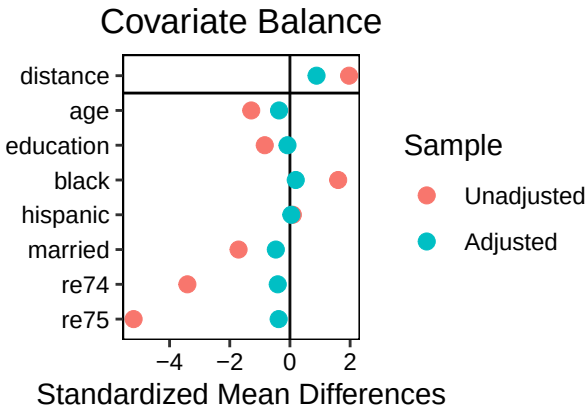
We can use `benchtestr` to explore the possibility of covariate imbalance, opportunities for matching-based estimation, and finally some comparison to non-matching estimators (e.g., DIM and linear regression).

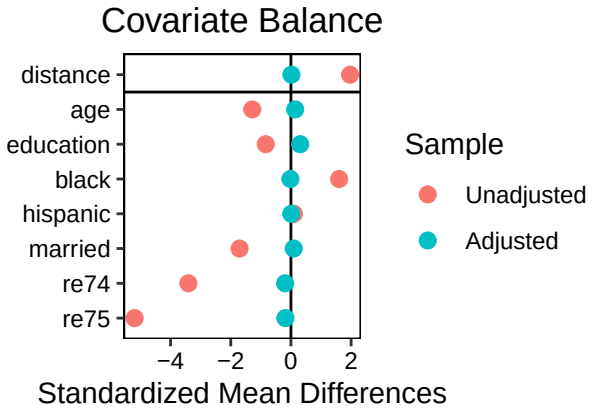
```
### Of course, you are welcome to use MatchIt and other packages for
### manual matching and evaluation.
### But here is the `benchtestr` case:

## 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)
```

```
## Loading required package: cobalt
##  cobalt (Version 4.3.1, Build Date: 2021-03-30 09:50:18 UTC)
##
## Attaching package: 'cobalt'
## The following object is masked _by_ '.GlobalEnv':
##
##     lalonde
## The following object is masked from 'package:benchtestr':
##
##     lalonde
## Loading required package: gridExtra
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
## Loading required package: MatchIt
##
## Attaching package: 'MatchIt'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     lalonde

## The following object is masked from 'package:cobalt':
##
##     lalonde

## The following object is masked from 'package:benchtestr':
##
##     lalonde

## Warning: Large mean differences detected; you may not be using standardized mean
## differences for continuous variables.

## Warning: Large mean differences detected; you may not be using standardized mean
## differences for continuous variables.

## Warning: Large mean differences detected; you may not be using standardized mean
## differences for continuous variables.
```








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

## 1) Settle on your preferred matching method and proceed with your analyis 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 datases.

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

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 additoinal analysis, so these next steps would need to take place in another environment.

References

Publications

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.

software

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