Setup Instructions for "Introduction to Double Robust Estimation for Causal Inference"

Laura B. Balzer, PhD

ISES-ISEE 2018 Pre-Conference Course:
Big data, machine learning techniques to investigate health effects
in environmental health studies
August 26, 2018

1 Introduction

We are looking forward to sharing the pre-conference course "Big data, machine learning techniques to investigate health effects in environmental health studies". These setup instructions focus on the portion of the course dedicated to "Introduction to Double Robust Estimation for Causal Inference". To get the most of the course, we recommend that you bring your laptop with the necessary software installed and the data set downloaded.

We will be using `R` in the workshop [R Core Team, 2018]. `R` is a language and environment for statistical computing and graphics. It provides a wide variety of statistical techniques (e.g. linear and nonlinear modeling, statistical tests, plotting, classification, clustering). `R` can also easily be extended via packages, covering a wide variety of modern statistics. There are about eight packages supplied with the `R` distribution, and many more are freely available online.

The following instructions describe how to install `R` and the required packages, as well as how to read in and explore the data set for the workshop. The whole process should take no more than 15 minutes.

2 Download and install `R`

`R` can be freely downloaded from the Comprehensive `R` Archive Network (CRAN) at `http://cran.r-project.org`. Contributed packages can be found at this website as well. Click the appropriate link for your operating system and follow the installation instructions. Running the most recent version of `R` is recommended.

3 Start an `R` session

Start `R` by double clicking the `R` icon on the desktop, looking for `R` in the Start menu, or finding it in your Applications menu. When `R` starts, the only window opened by default is the console, in

which you can type code to execute specific tasks. For instance, you can use R as a simple calculator by typing:

```
> 4+6/2+2*3

[1] 13

> # Hint 1. Do not type or copy the >. This represents the prompt in an R console.
> # Hint 2. Press "Enter" to run the code.
> # Hint 3. Hashtags are used in R for making comments. #YayTMLE!
```

As a second example, we can assign our previous calculations to object `x`:

```
> x<- 4+6/2+2*3
> x

[1] 13
```

If you are unfamiliar with R, we encourage you to check out some of the online tutorials (Section 6).

4 Install packages

To install additional R packages from CRAN, use the `install.packages()` function. For this workshop, we will need the Super Learner package, which allows us to apply and combine a variety of algorithms to our data [Polley et al., 2018]. Install `SuperLearner`, along with the packages on which it depends, by typing:

```
> install.packages("SuperLearner", dependencies = TRUE)
```

We will also want to install the package to implement targeted maximum likelihood estimation (TMLE) [Lendle et al., 2017]:

```
> install.packages("ltmle", dependencies = TRUE)
```

For a given version of R, you only need to install a package once. However, anytime you open a new R session, you need to remind R to use that package:

```
> library(SuperLearner)
> library(ltmle)
```

5 Read in and explore the data

Create a folder on your computer for the files that you will use and create during the workshop. Set this folder as your working directory in R with the `setwd()` function. For example, if I created a folder "BigData_2018" on my Dropbox, I would type

```
> setwd("~/Dropbox/BigData_2018")
```

R will use this folder for reading and writing files. You can check that your working directory with

```
> getwd()
```

Download and save the data set "CausalWorkshop.csv" to your workshop folder. The file is in comma separated values (csv) format. After setting your working directory, read the data set into R using

```
> data <- read.csv("CausalWorkshop.csv")
```

This reads in the data and stores it in an object named `data`. To make sure the data has been read in properly, you can view the first six lines of the data using

```
> head(data)
```

```
  W1          W2         W3         W4 A           Y
1  1 0.680496872 -0.4430059 -0.3693323 1 0.076577008
2  1 0.006966161 -2.0723602 -2.2561853 1 0.000035900
3  1 0.269141939 -1.3921258 -1.4211069 1 0.005613424
4  0 0.824778590  1.7246493  1.2792155 1 0.000059700
5  1 0.095466105 -2.0389724 -2.3132384 1 0.000025200
6  0 0.772451564 -2.2415619 -1.2843908 1 0.000072100
```

the last six lines using

```
> tail(data)
```

```
    W1         W2          W3         W4 A           Y
95   0 0.43866653  0.62099901  0.6736281 0 0.041108280
96   0 0.47763995  0.40945733  0.8789507 0 0.087362023
97   1 0.38601726  0.10916835 -0.9633303 0 0.045684975
98   1 0.51518289  1.02451436  1.6171894 1 0.000042700
99   0 0.05015548 -1.62360003 -1.1573370 0 0.005801751
100  1 0.88454275  0.05359586 -0.6562194 0 0.103371039
```

The `dim` function provides us with the dimensions of the data set, and the `summary` function provides a summary of the data set:

```
> dim(data)
```

```
[1] 100   6
```

```
> summary(data)
```

```
       W1              W2                  W3                  W4          
 Min.   :0.00   Min.   :0.006966   Min.   :-2.24156   Min.   :-2.31324  
 1st Qu.:0.00   1st Qu.:0.320116   1st Qu.:-0.66465   1st Qu.:-0.63671  
 Median :1.00   Median :0.536026   Median :-0.17106   Median :-0.04623  
 Mean   :0.54   Mean   :0.542049   Mean   :-0.09806   Mean   :-0.01003  
```

```
 3rd Qu.:1.00   3rd Qu.:0.804983   3rd Qu.: 0.59019   3rd Qu.: 0.62323
 Max.   :1.00   Max.   :0.975733   Max.   : 3.17559   Max.   : 2.69499
       A              Y
 Min.   :0.00   Min.   :0.00000
 1st Qu.:0.00   1st Qu.:0.01686
 Median :0.00   Median :0.06175
 Mean   :0.32   Mean   :0.06467
 3rd Qu.:1.00   3rd Qu.:0.09694
 Max.   :1.00   Max.   :0.22128
```

There are $n = 100$ observations in the data set. Now we are ready to start the pre-conference course!

6 Additional R resources

For additional help learning R, try the following resources:

- UCLA Institute for Digital Research and Education: https://stats.idre.ucla.edu/r/
- DataCamp R Programming: https://www.datacamp.com/
- Dalgaard, Peter. Introductory Statistics with R. New York, NY: Springer Science+Business Media, 2008.
- Teetor, Paul. R Cookbook. Sabastopol, CA: O'Reilly Media, 2011.

References

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL http://www.R-project.org.

E. Polley, E. LeDell, C. Kennedy, and M. van der Laan. *SuperLearner: Super Learner Prediction*, 2018. URL http://CRAN.R-project.org/package=SuperLearner. R package version 2.0-24.

S.D. Lendle, J. Schwab, M.L. Petersen, and M.J. van der Laan. ltmle: An R package implementing targeted minimum loss-based estimation for longitudinal data. *Journal of Statistical Software*, 81 (1):1–21, 2017.