Simulating data for mediation

Alessio Crippa

10/16/2017

Aim

The aim of this R Markdown document is to describe how to the data for the workshop on *mediation analysis* has been generated and to actual simulate them.

The simulation design is based on the Stata code written by Andrea Bellavia and available here.

Simulation design

The following code generates a sample population of 10.000 individuals with information on race/ethnicity (binary), fast-food consumption (binary), exposure to a certain chemicals (e.g. DiNP, metabolite of diisononyl phthalate, continuous), and diabetes (binary).

Despite some of the associations are chosen based on real data, the sample does not represent a real population but only a simplified situation. The purpose of the data is to illustrate the estimation and interpretation of results from mediation analysis.

Notes:

- Generate race/ethnicity (x) as a binary covariate with 19% of black-American.
- Generate the binary mediator (yes/no) of fast-food consumption. Use results from Zota et al, 2016, showing a proportion of 44% fast-food consumers among black-American, and 33% among other groups.
- A second continuous mediator represents the urinary concentration of a specific chemical. We assume that this covariate is associated with both race/ethnicity (main effect: $\beta_1 = 1.5$) and fast-food consumption (main effect: $\beta_2 = 0.9$). We also assume that an interaction between race/ethnicity and fast-food consumption is present ($\beta_3 = 0.5$).
  The chosen coefficients will provide a covariate an average DiNP concentration of 11 ug/l in the entire population, 10.5 ug/l among non black-American, and 12.6 ug/l among black-American.
  OBS: (Please note that in real situations environmental chemicals are seldom normally distributed).
- Generates diabetes (yes/no) as a function of race/ethnicity ($\exp(\gamma) = \text{OR} = 1.1$), fast-food consumption ($\exp(\gamma) = \text{OR} = 1.2$), and DiNP urinary concentration ($\exp(\gamma) = \text{OR} = 1.3$ for each unit increase of DiNP).

Code

```
library(tidyverse)
# seed for reproducibility
set.seed(1234)
n <- 10000
# x: race/ethnicity
x <- rbinom(n = n, size = 1, prob = .19)
# m1: fast-food consumption (first mediator)
m1 <- rbinom(n = n, size = 1, prob = ifelse(x == 1, .44, .33))
```

```r
# m2: DiNP concentration (second mediator)
beta <- c(10.5, 1.5, 0.9, 0.5)
m2 <- rnorm(n, cbind(1, x, m1, x*m1) %*% beta)
# y: diabetes
gamma <- c(-4.3, log(1.1), log(1.2), log(1.2))
invlogit <- function(x) exp(x)/(1 + exp(x))
y <- rbinom(n = n, size = 1, invlogit(cbind(1, x, m1, m2) %*% gamma))
dat <- data_frame(
  race = factor(x, labels = c( "Other", "Black-American")),
  fastfood = factor(m1, labels = c("no", "yes")),
  dinp = m2,
  diabetes = factor(y, labels = c("no", "yes"))
)
dat1 <- data.frame(x, m1, m2, y)
save(dat, file = "data/dat.Rda")
save(dat1, file = "data/dat1.Rda")
```

Descriptive of the simulated data

```r
# print data
dat
```

```
# A tibble: 10,000 x 4
             race fastfood      dinp diabetes
           <fctr>   <fctr>     <dbl>   <fctr>
 1          Other       no  8.683102       no
 2          Other       no 11.127167       no
 3          Other       no 11.018092       no
 4          Other      yes 11.540922       no
 5 Black-American       no 13.457272       no
 6          Other       no 10.006403       no
 7          Other       no  8.377756       no
 8          Other       no 10.366433       no
 9          Other       no 10.072400       no
10          Other       no 10.587795       no
# ... with 9,990 more rows
```

```r
# table by diabetes and race
dat %>%
  group_by(diabetes, race) %>%
  tally %>%
  group_by(diabetes) %>%
  mutate(pct = (100*n)/sum(n))
```

```
# A tibble: 4 x 4
# Groups:   diabetes [2]
  diabetes           race     n      pct
    <fctr>         <fctr> <int>    <dbl>
1       no          Other  7350 82.17800
2       no Black-American  1594 17.82200
3      yes          Other   780 73.86364
4      yes Black-American   276 26.13636
```

```r
# table by diabetes and fastfood
dat %>%
  group_by(diabetes, fastfood) %>%
  tally %>%
  group_by(diabetes) %>%
  mutate(pct = (100*n)/sum(n))
```

```
# A tibble: 4 x 4
# Groups:   diabetes [2]
  diabetes fastfood     n      pct
    <fctr>   <fctr> <int>    <dbl>
1       no       no  5837 65.26163
2       no      yes  3107 34.73837
3      yes       no   595 56.34470
4      yes      yes   461 43.65530
```

```r
# mean and std tables by diabetes and race
dat %>%
  group_by(diabetes, race) %>%
  summarise(mean = mean(dinp), std = sd(dinp))
```

```
# A tibble: 4 x 4
# Groups:   diabetes [?]
  diabetes           race     mean      std
    <fctr>         <fctr>    <dbl>    <dbl>
1       no          Other 10.77756 1.088614
2       no Black-American 12.54066 1.219341
3      yes          Other 11.01710 1.112565
4      yes Black-American 12.94945 1.327768
```