Correlated Counfounder and Propensity Score Matching

Miao Cai

2017-10-13

Creating simulation data

Random variables X_1 - X_3 have the correlation coefficient of 0.3; random variables X_4 - X_6 have the correlation coefficient of 0.5; random variables X_7 - X_9 have the correlation coefficient of 0.8. The true population parameters for X_1, X_4, X_7 is 2, parameters for X_2, X_5, X_8 is 3, parameters for X_3, X_6, X_9 is 1.

```
library(MASS)
library(Matrix)
library(GMCM)
```

```
## Warning: package 'GMCM' was built under R
## version 3.4.2
```

```
library(MatchIt)
```

```
## Warning: package 'MatchIt' was built under R
## version 3.4.2
```

```
# correlations
r1 = 0.3
r2 = 0.5
r3 = 0.8

# block diagnoal correlation matrix
m1 = matrix(r1, nrow=3, ncol=3)
diag(m1) = 1
m2 = matrix(r2, nrow=3, ncol=3)
diag(m2) = 1
m3 = matrix(r3, nrow=3, ncol=3)
diag(m3) = 1

cmat = bdiag(m1, m2, m3)

# covariates
x = data.frame(mvrnorm(n=1000, mu=rep(0,9), Sigma=cmat))

# pt: the probability to draw the binary treatment
pt = GMCM:::inv.logit(rowSums(x))
```

```r
# tr: treatment
tr = rbinom(n = 1000, size = 1, prob = pt)

# y
y = rnorm(n = 1000,
          mean = tr * 3 + 3*x$X1 + 2*x$X2 + x$X3 + 3*x$X4 + 2*x$X5 + x$X6 + 3*x$X7 + 2*x$X8 + x$X9,
          sd = 1)

# constructing the data.frame
dat <- data.frame(x, tr, y)
```

Part 1 Nine Covariates

This part firstly uses all 9 correlated covariates to match the treatment and comparison group[1]. Then I use linear regression to estimate the coefficients of $X_1 \sim X_9$, and Cohen's d is used to test effect size.[2]

[1] Propensity score method is used to match the treatment group and the comparison group. I use the **MatchIt** package to do propensity score matching

[2] Cohen's d is calculated using the following formula:

$$Cohen's\ d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}$$

,with the *cohen.d()* function in the **effsize** package. When paired is set, the effect size is computed using the approach suggested in (Gibbons et al. 1993) *Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. Journal of Educational Statistics, 18, 271-279.*

```
library(effsize)
```

```
## Warning: package 'effsize' was built under R
## version 3.4.2
```

```
set.seed(666)
```

```
#1 select all comparison cases
#  and randomly select 100 treatment cases
#  and all the comparison group cases
data1 <- rbind(dat[dat$tr == 0,],
               dat[sample(which(dat$tr == 1), 100),])
```

```
#2 match the treatment and comparison groups - 1 to 1 match
matcheddata1 <- match.data(
  matchit(tr ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
          data = data1,
          method = "nearest",
          ratio = 1))
```

```
#3.1 linear regression - y and treatment
lm1.1 <- lm(y ~ tr, data = matcheddata1)
knitr::kable(
  summary(lm1.1)$coefficients,
  caption = 'Linear regression between y and treatment',
  digits = 2
)
```

Table 1: Linear regression between y and treatment

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.52	0.54	0.96	0.34
tr	8.93	0.77	11.61	0.00

```
#3.2 linear regression - y, treatment and covariates
lm1.2 <- lm(y ~ tr + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
            data = matcheddata1)
knitr::kable(
```

```
  summary(lm1.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 9 covariates',
  digits = 2
)
```

Table 2: Linear regression between y ,treatment and 9 covariates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.02	0.10	0.20	0.84
tr	2.85	0.16	17.52	0.00
X1	3.10	0.07	43.77	0.00
X2	2.04	0.07	27.53	0.00
X3	0.99	0.08	12.55	0.00
X4	3.08	0.08	36.33	0.00
X5	1.96	0.08	23.10	0.00
X6	1.05	0.09	11.21	0.00
X7	2.98	0.13	23.09	0.00
X8	2.18	0.13	16.39	0.00
X9	0.89	0.14	6.54	0.00

```
#4 effect size - Cohen's d
cohen.d(matcheddata1$y[matcheddata1$tr == 1],
             matcheddata1$y[matcheddata1$tr == 0],
             paired = TRUE)$estimate
```

```
## [1] 1.22433
```

Part 2 Three Covariates

This part firstly uses 3 correlated covariates to match the treatment and comparison group. Then propensity scores are used to match the treatment groups and comparison groups. Linear regression and Cohen's d are conducted after propensity score matching.

```
#1 select all comparison cases
#  and randomly select 100 treatment cases
#  and all the comparison group cases
data2 <- rbind(dat[dat$tr == 0,],
               dat[sample(which(dat$tr == 1), 100),])

#2 match the treatment and comparison groups - 1 to 1 match
matcheddata2 <- match.data(
  matchit(tr ~ X1 + X4 + X7,
          data = data2,
          method = "nearest",
          ratio = 1))

#3.1 linear regression - y and treatment
lm2.1 <- lm(y ~ tr, data = matcheddata2)
knitr::kable(
  summary(lm2.1)$coefficients,
  caption = 'Linear regression between y and treatment',
  digits = 2
)
```

Table 3: Linear regression between y and treatment

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.42	0.55	-0.76	0.45
tr	10.44	0.78	13.40	0.00

```
#3.2 linear regression - yï¼ treatment and covariates
lm2.2 <- lm(y ~ tr + X1 + X4 + X7,
            data = matcheddata2)
knitr::kable(
  summary(lm2.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 9 covariates',
  digits = 2
)
```

Table 4: Linear regression between y ,treatment and 9 covariates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.62	0.34	-7.77	0
tr	7.95	0.46	17.25	0
X1	3.30	0.24	13.56	0
X4	3.79	0.24	15.95	0
X7	4.36	0.30	14.40	0

```
#4 Cohen's d
cohen.d(matcheddata2$y[matcheddata2$tr == 1],
             matcheddata2$y[matcheddata2$tr == 0],
             paired = TRUE)$estimate
```

```
## [1] 1.262105
```

Part 3 Integrating 9 Covariates into 3 Principal Components

This part integrates the 9 covariates into 3 principal components using one principal component analysis.[3] Then propensity scores are used to match the treatment groups and comparison groups using the 3 principal components. Linear regression and Cohen's d are conducted after propensity score matching.

[3] Prinpal component analysis is conducted using the base R function *prcomp()*

```
#1 select all comparison cases
#  and randomly select 100 treatment cases
#  and all the comparison group cases
data3 <- rbind(dat[dat$tr == 0,],
               dat[sample(which(dat$tr == 1), 100),])

#2 principal component analysis
pca3 <- prcomp(data3[,paste("X", 1:9, sep = "")], scale = FALSE)
pca3data <- data.frame(
  data3$y,
  data3$tr,
  pca3$x[,1:3]
  )#extract the  three PCs, y and tr
names(pca3data) <- c("y", "tr", "PC1", "PC2", "PC3")

#3 propensity score matching - one to one match
matcheddata3 <- match.data(
  matchit(tr ~ PC1 + PC2 + PC3,
          data = pca3data,
          method = "nearest",
          ratio = 1))

#4.1 linear regression - y and treatment
lm3.1 <- lm(y ~ tr, data = matcheddata3)
knitr::kable(
  summary(lm3.1)$coefficients,
  caption = 'Linear regression between y and treatment',
  digits = 2
)
```

Table 5: Linear regression between y and treatment

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.96	0.52	1.85	0.07
tr	8.46	0.73	11.57	0.00

```
#4.2 linear regression - yï¼ treatment and covariates
lm3.2 <- lm(y ~ tr + PC1 + PC2 + PC3,
            data = matcheddata3)
knitr::kable(
  summary(lm3.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 3 PCs',
  digits = 2
)
```

Table 6: Linear regression between y ,treatment and 3 PCs

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.01	0.26	-15.57	0
tr	2.91	0.34	8.59	0
PC1	-1.03	0.09	-10.90	0
PC2	2.33	0.11	20.51	0
PC3	-5.36	0.18	-30.00	0

```
#5 Cohen's d
cohen.d(matcheddata3$y[matcheddata3$tr == 1],
            matcheddata3$y[matcheddata3$tr == 0],
            paired = TRUE)$estimate
```

```
## [1] 1.132767
```

Part 4 Separately Integrating 9 Covariates into 3 sets of Principal Components

This part separately integrates the 9 covariates into 3 sets principal components.[4] Then propensity scores are used to match the treatment groups and comparison groups using the 3 sets of principal components. Linear regression and Cohen's d are conducted after propensity score matching.

[4] Different from part 3, this part uses 3 principal component analyses and integrates $X_1 - X_3$ into PC_1, integrates $X_4 - X_6$ into PC_2, and integrates $X_7 - X_9$ into PC_3.

```
#1 select all comparison cases
#  and randomly select 100 treatment cases
#  and all the comparison group cases
data4 <- rbind(dat[dat$tr == 0,],
               dat[sample(which(dat$tr == 1), 100),])

#2 principal component analysis - set 1
pca4.1 <- prcomp(data4[,paste("X", 1:3, sep = "")], scale = FALSE)
pca4.2 <- prcomp(data4[,paste("X", 4:6, sep = "")], scale = FALSE)
pca4.3 <- prcomp(data4[,paste("X", 7:9, sep = "")], scale = FALSE)
pca4data <- data.frame(
  data4$y,
  data4$tr,
  pca4.1$x[,1],
  pca4.2$x[,1],
  pca4.3$x[,1]
  )#extract the  three PCs, y and tr
names(pca4data) <- c("y", "tr", "PC1", "PC2", "PC3")

#3 propensity score matching - one to one match
matcheddata4 <- match.data(
  matchit(tr ~ PC1 + PC2 + PC3,
          data = pca4data,
          method = "nearest",
          ratio = 1))

#4.1 linear regression - y and treatment
lm4.1 <- lm(y ~ tr, data = matcheddata4)
knitr::kable(
  summary(lm4.1)$coefficients,
  caption = 'Linear regression between y and treatment',
  digits = 2
)
```

Table 7: Linear regression between y and treatment

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.98	0.53	1.87	0.06
tr	8.78	0.74	11.81	0.00

```
#4.2 linear regression - yï¼ treatment and covariates
lm4.2 <- lm(y ~ tr + PC1 + PC2 + PC3,
            data = matcheddata4)
knitr::kable(
  summary(lm4.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 3 PCs',
  digits = 2
)
```

Table 8: Linear regression between y ,treatment and 3 PCs

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.06	0.24	-17.06	0
tr	2.73	0.32	8.57	0
PC1	3.62	0.12	29.00	0
PC2	-3.47	0.13	-27.44	0
PC3	-3.49	0.11	-30.77	0

```
#5 Cohen's d
cohen.d(matcheddata4$y[matcheddata4$tr == 1],
            matcheddata4$y[matcheddata4$tr == 0],
            paired = TRUE)$estimate
```

```
## [1] 1.328392
```