Correlated Counfounder and Propensity Score Matching

Miao Cai

2017-11-27

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
```

```
set.seed(666)

# 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
##REVISED: rowSums(x)-3.8 to reduce proportion treated
pt = GMCM:::inv.logit(rowSums(x)-3.8)
## REVISED: to confirm that mean(pt) is near 0.2
mean(pt)
```

```
## [1] 0.1916124
```

```
# mean(pt) is around 0.2 to make sure there are sufficient
# number of comparison groups to choose from.

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

# y: outcome - POPULATION PARAMETER for treatment is 3
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 the 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.*

Section 1.1 Nine covariates without matching

1.1.1 `y ~ tr` on unmatched data

```
library(effsize)
```

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

```
lm1.1.1 <- lm(y ~ tr, data = dat)

# summary the output
knitr::kable(
  summary(lm1.1.1)$coefficients,
  caption = 'Linear regression between y and treatment on unmatched data',
  digits = 3
)
```

Table 1: Linear regression between y and treatment on unmatched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.644	0.237	-11.161	0
tr	16.195	0.542	29.882	0

```
# get the Cohen's d for this model
cohen.d(dat$y,as.factor(dat$tr))$estimate
```

```
##         0
## -2.403914
```

1.1.2 `y ~ tr + 9` covariates on unmatched data

```
lm1.1.2 <- lm(y ~ tr + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
              data = dat)

# summary the output
knitr::kable(
  summary(lm1.1.2)$coefficients,
```

```
  caption = 'Linear regression between y and treatment, 9 covariates on unmatched data',
  digits = 3
)
```

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.027	0.037	0.732	0.464
tr	3.047	0.101	30.134	0.000
X1	2.963	0.034	87.260	0.000
X2	1.989	0.033	59.382	0.000
X3	0.986	0.033	29.627	0.000
X4	2.995	0.039	76.521	0.000
X5	1.953	0.039	50.346	0.000
X6	1.038	0.039	26.948	0.000
X7	3.053	0.059	52.107	0.000
X8	2.154	0.057	37.819	0.000
X9	0.816	0.060	13.571	0.000

```
# get the Cohen's d for this model
cohen.d(dat$y,as.factor(dat$tr))$estimate
```

```
##         0
## -2.403914
```

1.1.3 Cohen's d for each covariate by tr

```
cohen.d(dat$X1,as.factor(dat$tr))$estimate
```

```
##         0
## -0.589994
```

```
cohen.d(dat$X2,as.factor(dat$tr))$estimate
```

```
##          0
## -0.6271197
```

```
cohen.d(dat$X3,as.factor(dat$tr))$estimate
```

```
##          0
## -0.4717218
```

```
cohen.d(dat$X4,as.factor(dat$tr))$estimate
```

```
##          0
## -0.7108146
```

```
cohen.d(dat$X5,as.factor(dat$tr))$estimate
```

```
##          0
## -0.7753352
```

```
cohen.d(dat$X6,as.factor(dat$tr))$estimate
```

```
##          0
## -0.7488767
```

```
cohen.d(dat$X7,as.factor(dat$tr))$estimate
```

```
##         0
## -1.010066
```

```
cohen.d(dat$X8,as.factor(dat$tr))$estimate
```

```
##          0
## -0.9972595
```

```
cohen.d(dat$X9,as.factor(dat$tr))$estimate
```

```
##         0
## -1.009796
```

Section 1.2 Nine covariates with matching

1.2.1 `y ~ tr` on matched data

```
#1 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 = dat,
          method = "nearest",
          ratio = 1))

#2 linear regression - y and treatment on matched data
lm1.2.1 <- lm(y ~ tr, data = matcheddata1)
knitr::kable(
  summary(lm1.2.1)$coefficients,
  caption = 'Linear regression between y and treatment on matched data',
  digits = 2
)
```

Table 3: Linear regression between y and treatment on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.29	0.32	16.41	0
tr	8.26	0.46	18.14	0

```
cohen.d(matcheddata1$y, as.factor(matcheddata1$tr))$estimate
```

```
##        0
## 1.856162
```

1.2.2 y ~ tr + 9 covariates on matched data

```
#3.2 linear regression - y, treatment and covariates
lm1.2.2 <- lm(y ~ tr + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
            data = matcheddata1)
knitr::kable(
  summary(lm1.2.2)$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)	0.09	0.11	0.87	0.38
tr	3.03	0.12	24.26	0.00
X1	3.02	0.06	50.42	0.00
X2	1.86	0.06	29.77	0.00
X3	1.02	0.06	17.25	0.00
X4	3.00	0.07	42.87	0.00
X5	1.95	0.07	27.87	0.00
X6	0.96	0.07	13.38	0.00
X7	3.01	0.10	28.97	0.00
X8	2.20	0.09	24.22	0.00
X9	0.85	0.10	8.19	0.00

```
#4 effect size - Cohen's d
cohen.d(matcheddata1$y, as.factor(matcheddata1$tr))$estimate
```

```
##        0
## 1.856162
```

1.2.3 Cohen's d for each covariate by tr

```
cohen.d(matcheddata1$X1,as.factor(matcheddata1$tr))$estimate
```

```
##        0
## 0.255222
```

```
cohen.d(matcheddata1$X2,as.factor(matcheddata1$tr))$estimate
```

```
##         0
## 0.3259172
```

```
cohen.d(matcheddata1$X3,as.factor(matcheddata1$tr))$estimate
```

```
##         0
## 0.2019597
```

```
cohen.d(matcheddata1$X4,as.factor(matcheddata1$tr))$estimate
```

```
##         0
## 0.2646747
```

```
cohen.d(matcheddata1$X5,as.factor(matcheddata1$tr))$estimate
```

```
##         0
## 0.2625669
```

```
cohen.d(matcheddata1$X6,as.factor(matcheddata1$tr))$estimate
```

```
##         0
## 0.1502806
```

```
cohen.d(matcheddata1$X7,as.factor(matcheddata1$tr))$estimate
```

```
##         0
## 0.4512691
```

```
cohen.d(matcheddata1$X8,as.factor(matcheddata1$tr))$estimate
```

```
##         0
## 0.4743845
```

```
cohen.d(matcheddata1$X9,as.factor(matcheddata1$tr))$estimate
```

```
##         0
## 0.4069146
```

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.

Section 2.1 Three uncorrelated covariates on unmatched data

2.1.1 `y ~ tr` on unmatched data

```
lm2.1.1 <- lm(y ~ tr, data = dat)

# summary the output
knitr::kable(
  summary(lm2.1.1)$coefficients,
  caption = 'Linear regression between y and treatment on unmatched data',
  digits = 3
)
```

Table 5: Linear regression between y and treatment on unmatched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.644	0.237	-11.161	0
tr	16.195	0.542	29.882	0

```
# get the Cohen's d for this model
cohen.d(dat$y,as.factor(dat$tr))$estimate

##         0
## -2.403914
```

2.1.2 `y ~ tr + 3` uncorrelated covariates on unmatched data

```
lm2.1.2 <- lm(y ~ tr + X1 + X4 + X7, data = dat)

# summary the output
knitr::kable(
  summary(lm2.1.2)$coefficients,
  caption = 'Linear regression between y and treatment, 3 uncorrelated covariates on unmatched data',
  digits = 3
)
```

Table 6: Linear regression between y and treatment, 3 uncorrelated covariates on unmatched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.971	0.124	-7.826	0
tr	7.144	0.322	22.173	0
X1	3.330	0.113	29.396	0
X4	4.049	0.115	35.064	0
X7	4.748	0.120	39.592	0

```
# get the Cohen's d for this model
cohen.d(dat$y,as.factor(dat$tr))$estimate
```

```
##         0
## -2.403914
```

2.1.3 Cohen's d for each covariate by tr

```
cohen.d(dat$X1,as.factor(dat$tr))$estimate
```

```
##         0
## -0.589994
```

```
cohen.d(dat$X4,as.factor(dat$tr))$estimate
```

```
##          0
## -0.7108146
```

```
cohen.d(dat$X7,as.factor(dat$tr))$estimate
```

```
##         0
## -1.010066
```

Section 2.2 Three uncorrelated covariates on matched data

2.2.1 `y ~ tr` on matched data

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

#2 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 on matched data',
  digits = 2
)
```

Table 7: Linear regression between y and treatment on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.72	0.36	10.21	0
tr	9.83	0.51	19.11	0

```
#3 Cohen's d
cohen.d(matcheddata2$y, as.factor(matcheddata2$tr))$estimate
```

```
##        0
## 1.955425
```

2.2.2 y ~ tr + X1 + X4 + X7 on matched data

```
#1 linear regression - y  treatment and covariates
lm2.2.2 <- lm(y ~ tr + X1 + X4 + X7,
            data = matcheddata2)
knitr::kable(
  summary(lm2.2.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 3 covariates on matched data',
  digits = 2
)
```

Table 8: Linear regression between y ,treatment and 3 covariates on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.82	0.29	-2.79	0.01
tr	7.86	0.33	24.06	0.00
X1	3.08	0.17	17.74	0.00
X4	3.49	0.20	17.57	0.00
X7	4.17	0.21	19.71	0.00

```
#2 Cohen's d
cohen.d(matcheddata2$y, as.factor(matcheddata2$tr))$estimate
```

```
##        0
```

```
## 1.955425
```

2.2.3 Cohen's d for each covariate by tr

```
cohen.d(matcheddata2$X1,as.factor(matcheddata2$tr))$estimate
```

```
##         0
## 0.1959693
```

```
cohen.d(matcheddata2$X4,as.factor(matcheddata2$tr))$estimate
```

```
##        0
## 0.134114
```

```
cohen.d(matcheddata2$X7,as.factor(matcheddata2$tr))$estimate
```

```
##         0
## 0.2675217
```

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] Pripal component analysis is conducted using the base R function *prcomp()*

Section 3.1 Regression on unmatched data

3.1.1 `y ~ tr` on unmatched data

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

#2 Linear regression on unmatched data
lm3.1.1 <- lm(y ~ tr, data = dat)

# summary the output
knitr::kable(
  summary(lm3.1.1)$coefficients,
  caption = 'Linear regression between y and treatment on unmatched data',
  digits = 3
)
```

Table 9: Linear regression between y and treatment on unmatched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.644	0.237	-11.161	0
tr	16.195	0.542	29.882	0

```
# get the Cohen's d for this model
cohen.d(pca3data$y, as.factor(pca3data$tr))$estimate
```

```
##         0
## -2.403914
```

3.1.2 y ~ tr + 3PCs on unmatched data

```
#2 Linear regression on unmatched data
lm3.1.2 <- lm(y ~ tr + PC1 + PC2 + PC3, data = pca3data)

# summary the output
knitr::kable(
  summary(lm3.1.2)$coefficients,
  caption = 'Linear regression between y and treatment and 3 PCs on unmatched data',
  digits = 3
)
```

Table 10: Linear regression between y and treatment and 3 PCs on unmatched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.161	0.070	-2.306	0.021
tr	3.197	0.193	16.604	0.000
PC1	3.147	0.041	76.026	0.000
PC2	3.291	0.046	71.748	0.000
PC3	-3.855	0.054	-70.753	0.000

```
# get the Cohen's d for this model
cohen.d(pca3data$y, as.factor(pca3data$tr))$estimate

##         0
## -2.403914
```

3.1.3 Cohen's d for each covariate by tr

```
cohen.d(pca3data$PC1,as.factor(pca3data$tr))$estimate

##         0
## -1.021873

cohen.d(pca3data$PC2,as.factor(pca3data$tr))$estimate

##         0
## -0.919134

cohen.d(pca3data$PC3,as.factor(pca3data$tr))$estimate

##         0
## 0.9820388
```

Section 3.2 Regression on matched data

3.2.1 `y ~ tr` on matched data

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

#2 linear regression - y and treatment
lm3.2.1 <- lm(y ~ tr, data = matcheddata3)
knitr::kable(
  summary(lm3.2.1)$coefficients,
  caption = 'Linear regression between y and treatment on matched data',
  digits = 2
)
```

Table 11: Linear regression between y and treatment on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.40	0.32	17.06	0
tr	8.15	0.45	18.20	0

```
#2 Cohen's d
cohen.d(matcheddata3$y, as.factor(matcheddata3$tr))$estimate
```

```
##        0
## 1.862391
```

3.2.2 `y ~ tr + 3PC` on matched data

```
#1 linear regression - y  treatment and covariates
lm3.2.2 <- lm(y ~ tr + PC1 + PC2 + PC3,
            data = matcheddata3)
knitr::kable(
  summary(lm3.2.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 3 PCs on matched data',
  digits = 2
)
```

Table 12: Linear regression between y ,treatment and 3 PCs on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.24	0.19	-1.25	0.21
tr	3.16	0.22	14.09	0.00
PC1	3.19	0.09	36.25	0.00
PC2	3.27	0.10	32.60	0.00
PC3	-3.96	0.11	-35.43	0.00

```
#2 Cohen's d
cohen.d(matcheddata3$y, matcheddata3$tr)$estimate
```

```
## Treatment
##  2.115921
```

3.2.3 Cohen's d for each covariate by tr

```
cohen.d(matcheddata3$PC1,as.factor(matcheddata3$tr))$estimate
```

```
##         0
## 0.4299604
```

```
cohen.d(matcheddata3$PC2,as.factor(matcheddata3$tr))$estimate
```

```
##         0
## 0.3144533
```

```
cohen.d(matcheddata3$PC3,as.factor(matcheddata3$tr))$estimate
```

```
##          0
## -0.4313365
```

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.

Section 4.1 Regression on unmatched data

4.1.1 `y ~ tr` on unmatched data

```
#1 principal component analysis - 3 sets
pca4.1 <- prcomp(dat[,paste("X", 1:3, sep = "")], scale = FALSE)
pca4.2 <- prcomp(dat[,paste("X", 4:6, sep = "")], scale = FALSE)
pca4.3 <- prcomp(dat[,paste("X", 7:9, sep = "")], scale = FALSE)
pca4data <- data.frame(
  dat$y,
  dat$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")

#2 Linear regression on unmatched data
lm4.1.1 <- lm(y ~ tr, data = pca4data)

# summary the output
knitr::kable(
  summary(lm4.1.1)$coefficients,
  caption = 'Linear regression between y and treatment on unmatched data',
  digits = 3
)
```

Table 13: Linear regression between y and treatment on unmatched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.644	0.237	-11.161	0
tr	16.195	0.542	29.882	0

```
#3 get the Cohen's d for this model
```

```
cohen.d(pca4data$y, as.factor(pca4data$tr))$estimate
```

```
##         0
## -2.403914
```

4.1.2 y ~ tr + 3PCs on unmatched data

```
#2 Linear regression on unmatched data
lm4.1.2 <- lm(y ~ tr + PC1 + PC2 + PC3, data = pca4data)

# summary the output
knitr::kable(
  summary(lm4.1.2)$coefficients,
  caption = 'Linear regression between y and treatment and 3 PCs on unmatched data',
  digits = 3
)
```

Table 14: Linear regression between y and treatment and 3 PCs on unmatched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.152	0.069	-2.209	0.027
tr	3.147	0.189	16.629	0.000
PC1	-3.496	0.053	-66.370	0.000
PC2	3.408	0.045	75.068	0.000
PC3	3.456	0.041	83.323	0.000

```
# get the Cohen's d for this model
cohen.d(pca4data$y, as.factor(pca4data$tr))$estimate
```

```
##         0
## -2.403914
```

4.1.3 Cohen's d for each covariate by tr

```
cohen.d(pca4data$PC1,as.factor(pca4data$tr))$estimate
```

```
##         0
## 0.8299555
```

```
cohen.d(pca4data$PC2,as.factor(pca4data$tr))$estimate
```

```
##          0
## -0.9327787
```

```
cohen.d(pca4data$PC3,as.factor(pca4data$tr))$estimate
```

```
##         0
## -1.096185
```

Section 4.2 Regression on matched data

4.2.1 `y ~ tr` on matched data

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

#2 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 on matched data',
  digits = 2
)
```

Table 15: Linear regression between y and treatment on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.35	0.32	16.77	0
tr	8.20	0.45	18.18	0

```
#3 Cohen's d
cohen.d(matcheddata4$y, as.factor(matcheddata4$tr))$estimate
```

```
##       0
## 1.86005
```

4.2.2 `y ~ tr + 3PC` on matched data

```
#1 linear regression - y  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 on matched data',
  digits = 2
)
```

Table 16: Linear regression between y ,treatment and 3 PCs on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.36	0.19	-1.94	0.05
tr	3.10	0.22	14.10	0.00
PC1	-3.62	0.10	-34.71	0.00
PC2	3.44	0.10	34.42	0.00
PC3	3.56	0.09	39.43	0.00

```
#2 Cohen's d
cohen.d(matcheddata4$y, as.factor(matcheddata4$tr))$estimate
```

```
##       0
## 1.86005
```

4.2.3 Cohen's d for each covariate by tr

```
cohen.d(matcheddata4$PC1,as.factor(matcheddata4$tr))$estimate
```

```
##          0
## -0.3351799
```

```
cohen.d(matcheddata4$PC2,as.factor(matcheddata4$tr))$estimate
```

```
##         0
## 0.3166284
```

```
cohen.d(matcheddata4$PC3,as.factor(matcheddata4$tr))$estimate
```

```
##         0
## 0.4917816
```