Correlated Counfounder and Propensity Score Matching

Miao Cai

2017-12-03

Corrections in this 20171202 version

1 Keep the directions of Cohen's d consistent across all comparisons[1]

2 Cohen's d for all 9 covariates, for 3 covariates and 3 PCs

3 Adding the 7 covariates and 5 covariates part

3 Display the loading factors

4 Adding the Sufficient Reduction Method

[1] This is correctified by explicitly assigning the levels of the factors `levels = c(0, 1)` in `cohen.d(dat$y,factor(dat$tr, levels = c(0, 1)))$estimate`.

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[2]. 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.[3]

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$y,factor(dat$tr, levels = c(0, 1)))$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,
```

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

[3] 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.*

```
  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

1.1.3 Cohen's d for each covariate by tr

```
cohen.d(dat$X1,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(dat$X2,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(dat$X3,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(dat$X4,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(dat$X5,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(dat$X6,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(dat$X7,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(dat$X8,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(dat$X9,factor(dat$tr, levels = c(0, 1)))$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 = 3
)
```

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.288	0.322	16.414	0
tr	8.264	0.456	18.139	0

```
cohen.d(matcheddata1$y, factor(matcheddata1$tr, levels = c(0, 1)))$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 = 3
)
```

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.092	0.106	0.873	0.383
tr	3.026	0.125	24.256	0.000
X1	3.019	0.060	50.421	0.000
X2	1.863	0.063	29.766	0.000
X3	1.023	0.059	17.253	0.000
X4	2.998	0.070	42.866	0.000
X5	1.954	0.070	27.872	0.000
X6	0.956	0.071	13.384	0.000
X7	3.009	0.104	28.973	0.000
X8	2.196	0.091	24.224	0.000
X9	0.847	0.103	8.191	0.000

```
#4 effect size - Cohen's d
cohen.d(matcheddata1$y, factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

1.2.3 Cohen's d for each covariate by tr

```
cohen.d(matcheddata1$X1,factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata1$X2,factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata1$X3,factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata1$X4,factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata1$X5,factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata1$X6,factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata1$X7,factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata1$X8,factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata1$X9,factor(matcheddata1$tr, levels = c(0, 1)))$estimate
```

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

Part 2 Seven Covariates (X1 ~ X7)

This part compares the results regressing X_1 - X_7 on y and the regression after propensity score matching.

Section 2.1 Seven covariates without matching

2.1.1 `y ~ tr + X1~X7` on unmatched data

```
library(effsize)
lm2.1.1 <- lm(y ~ tr + X1 + X2 + X3 + X4 + X5 + X6 + X7,
              data = dat)

# summary the output
knitr::kable(
  summary(lm2.1.1)$coefficients,
  caption = 'Linear regression between y and treatment, 7 covariates on unmatched data',
  digits = 3
)
```

Table 5: Linear regression between y and treatment, 7 covariates on unmatched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.184	0.069	-2.658	0.008
tr	3.811	0.188	20.231	0.000
X1	2.866	0.064	44.838	0.000
X2	1.940	0.063	30.736	0.000
X3	0.973	0.063	15.509	0.000
X4	2.924	0.073	39.782	0.000
X5	2.033	0.073	27.839	0.000
X6	1.011	0.073	13.924	0.000
X7	5.274	0.066	79.762	0.000

```
# get the Cohen's d for this model
cohen.d(dat$y,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

Section 2.2 Seven covariates with matching

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 + X2 + X3 + X4 + X5 + X6 + X7,
          data = dat,
          method = "nearest",
          ratio = 1))

#2 linear regression - y and treatment on matched data
lm2.2.1 <- lm(y ~ tr, data = matcheddata2)
knitr::kable(
  summary(lm2.2.1)$coefficients,
  caption = 'Linear regression between y and treatment on matched data',   digits = 3
)
```

Table 6: Linear regression between y and treatment on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.950	0.327	15.150	0
tr	8.602	0.462	18.616	0

```
cohen.d(matcheddata2$y, factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##        0
## 1.904958
```

2.2.2 y ~ tr + X1~X7 on matched data

```
#2.2 linear regression - y, treatment and 7 covariates
lm2.2.2 <- lm(y ~ tr + X1 + X2 + X3 + X4 + X5 + X6 + X7,
             data = matcheddata2)
knitr::kable(
  summary(lm2.2.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 7 covariates',
  digits = 3
)
```

Table 7: Linear regression between y ,treatment and 7 covariates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.259	0.193	-1.340	0.181
tr	4.085	0.220	18.556	0.000
X1	2.835	0.107	26.535	0.000
X2	1.786	0.113	15.780	0.000
X3	0.889	0.106	8.357	0.000

	Estimate	Std. Error	t value	Pr(>|t|)
X4	2.906	0.125	23.199	0.000
X5	2.004	0.125	15.977	0.000
X6	1.026	0.132	7.760	0.000
X7	5.184	0.146	35.492	0.000

```
#4 effect size - Cohen's d
cohen.d(matcheddata2$y, factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##        0
## 1.904958
```

2.2.3 Cohen's d for each covariate by tr

```
cohen.d(matcheddata2$X1,factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##        0
## 0.282992
```

```
cohen.d(matcheddata2$X2,factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2704297
```

```
cohen.d(matcheddata2$X3,factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2364499
```

```
cohen.d(matcheddata2$X4,factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2009865
```

```
cohen.d(matcheddata2$X5,factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2764484
```

```
cohen.d(matcheddata2$X6,factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.1153608
```

```
cohen.d(matcheddata2$X7,factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.4303553
```

```
cohen.d(matcheddata2$X8,factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.6766256
```

```
cohen.d(matcheddata2$X9,factor(matcheddata2$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.6441325
```

Part 3 Five Covariates (X1 ~ X4, X7)

This part compares the results regressing X_1 - X_4 and X_7 on y and the regression after propensity score matching.

Section 3.1 Five covariates without matching

3.1.1 `y ~ tr + X1~X4 + X7` on unmatched data

```
library(effsize)
lm3.1.1 <- lm(y ~ tr + X1 + X2 + X3 + X4 + X7,
              data = dat)

# summary the output
knitr::kable(
  summary(lm3.1.1)$coefficients,
  caption = 'Linear regression between y and treatment, 5 covariates on unmatched data',
  digits = 3
)
```

Table 8: Linear regression between y and treatment, 5 covariates on unmatched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.613	0.104	-5.882	0
tr	5.581	0.277	20.125	0
X1	2.788	0.098	28.569	0
X2	1.747	0.096	18.207	0
X3	0.870	0.096	9.089	0
X4	4.274	0.096	44.339	0
X7	5.022	0.100	50.060	0

```
# get the Cohen's d for this model
cohen.d(dat$y,factor(dat$tr, levels = c(0, 1)))$estimate
```

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

Section 3.2 Five covariates with matching

3.2.1 `y ~ tr` on matched data

```
#1 match the treatment and comparison groups - 1 to 1 match
matcheddata3 <- match.data(
  matchit(tr ~ X1 + X2 + X3 + X4 + X7,
          data = dat,
          method = "nearest",
          ratio = 1))

#2 linear regression - y and treatment on matched data
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 = 3
)
```

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.312	0.344	12.527	0
tr	9.240	0.487	18.982	0

```
cohen.d(matcheddata3$y, factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##        0
## 1.942417
```

3.2.2 y ~ tr + X1~X7 on matched data

```
# linear regression - y, treatment and 5 covariates
lm3.2.2 <- lm(y ~ tr + X1 + X2 + X3 + X4 + X7,
           data = matcheddata3)
knitr::kable(
  summary(lm3.2.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 9 covariates',
  digits = 3
)
```

Table 10: Linear regression between y ,treatment and 9 covariates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.291	0.257	-1.135	0.257
tr	6.158	0.293	20.996	0.000
X1	2.537	0.149	16.974	0.000
X2	1.677	0.158	10.583	0.000
X3	0.715	0.155	4.621	0.000
X4	3.747	0.174	21.566	0.000
X7	4.497	0.192	23.383	0.000

```
# effect size - Cohen's d
cohen.d(matcheddata3$y, factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##        0
## 1.942417
```

3.2.3 Cohen's d for each covariate by tr

```
cohen.d(matcheddata3$X1,factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##        0
## 0.221695
```

```
cohen.d(matcheddata3$X2,factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2253956
```

```
cohen.d(matcheddata3$X3,factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.1260765
```

```
cohen.d(matcheddata3$X4,factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2304387
```

```
cohen.d(matcheddata3$X5,factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.7392719
```

```
cohen.d(matcheddata3$X6,factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.6766296
```

```
cohen.d(matcheddata3$X7,factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.3356225
```

```
cohen.d(matcheddata3$X8,factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.6067564
```

```
cohen.d(matcheddata3$X9,factor(matcheddata3$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.5228617
```

Part 4 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 4.1 Three uncorrelated covariates on unmatched data

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

```
lm4.1.1 <- lm(y ~ tr + X1 + X4 + X7, data = dat)

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

Table 11: 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

Section 4.2 Three uncorrelated covariates on matched data

4.2.1 `y ~ tr` on matched data

```
#1 match the treatment and comparison groups - 1 to 1 match
matcheddata4 <- match.data(
  matchit(tr ~ X1 + X4 + X7,
          data = dat,
          method = "nearest",
          ratio = 1))

#2 linear regression - y and treatment
lm4.2.1 <- lm(y ~ tr, data = matcheddata4)
knitr::kable(
  summary(lm4.2.1)$coefficients,
  caption = 'Linear regression between y and treatment on matched data',
  digits = 3
)
```

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

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.717	0.364	10.213	0
tr	9.835	0.515	19.109	0

```
#3 Cohen's d
cohen.d(matcheddata4$y, factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

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

4.2.2 y ~ tr + X1 + X4 + X7 *on matched data*

```
#1 linear regression - y  treatment and covariates
lm4.2.2 <- lm(y ~ tr + X1 + X4 + X7,
           data = matcheddata4)
knitr::kable(
  summary(lm4.2.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 3 covariates on matched data',
  digits = 3
)
```

Table 13: Linear regression between y ,treatment and 3 covariates on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.819	0.294	-2.786	0.006
tr	7.861	0.327	24.057	0.000
X1	3.080	0.174	17.744	0.000
X4	3.492	0.199	17.574	0.000
X7	4.174	0.212	19.711	0.000

```
#2 Cohen's d
cohen.d(matcheddata4$y, factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

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

4.2.3 Cohen's d for each covariate by tr

```
cohen.d(matcheddata4$X1,factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata4$X2,factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.7579189
```

```
cohen.d(matcheddata4$X3,factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.5814118
```

```
cohen.d(matcheddata4$X4,factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata4$X5,factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.6270239
```

```
cohen.d(matcheddata4$X6,factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.6438917
```

```
cohen.d(matcheddata4$X7,factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata4$X8,factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.5792386
```

```
cohen.d(matcheddata4$X9,factor(matcheddata4$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.4788618
```

Part 5 Integrating 9 Covariates into 3 Principal Components

This part integrates the 9 covariates into 3 principal components using one principal component analysis.[4] 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.

Section 5.1 Regression on unmatched data

5.1.1 Principle component analysis

```
#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 The standard deviation of the principle components
knitr::kable(
  pca3$sdev,
  caption = 'The standard deviation of the principle components',
  digits = 3
)
```

```
## Warning in kable_markdown(x =
## structure(c("1.589", "1.412", "1.200",
## "0.892", : The table should have a header
## (column names)
```

1.589
1.412
1.200
0.892
0.840
0.712
0.695
0.458
0.430

```
#3 The matrix of variable loadings (columns are eigenvectors)
knitr::kable(
```

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

```
  pca3$rotation,
  caption = 'The matrix of variable loadings (columns are eigenvectors)',
  digits = 3
)
```

Table 15: The matrix of variable loadings (columns are eigenvectors)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
X1	-0.040	-0.014	-0.605	0.039	-0.793	0.044	-0.004	0.019	-0.003
X2	-0.029	-0.065	-0.593	0.637	0.482	-0.068	0.030	-0.009	0.001
X3	-0.021	-0.017	-0.526	-0.767	0.366	0.013	-0.033	-0.011	-0.001
X4	-0.034	0.566	-0.035	0.018	0.048	0.546	0.609	0.044	-0.065
X5	0.002	0.589	-0.033	0.064	0.034	0.242	-0.765	-0.057	0.015
X6	-0.022	0.572	-0.026	-0.034	-0.036	-0.797	0.185	-0.012	0.004
X7	0.580	-0.007	-0.030	0.003	-0.019	-0.006	0.029	-0.600	-0.549
X8	0.578	0.021	-0.031	0.002	0.012	-0.025	-0.050	0.777	-0.240
X9	0.570	0.034	-0.034	0.002	-0.009	0.033	0.068	-0.175	0.798

```
#4 The variable means
knitr::kable(
  pca3$center,
  caption = 'The variable means',
  digits = 3
)
```

Table 16: The variable means

X1	-0.057
X2	-0.047
X3	-0.024
X4	0.020
X5	-0.015
X6	-0.026
X7	0.034
X8	0.000
X9	0.020

5.1.2 y ~ tr + 3PCs on unmatched data

```
#2 Linear regression on unmatched data
lm5.1.2 <- lm(y ~ tr + PC1 + PC2 + PC3, data = pca3data)
```

```
# summary the output
knitr::kable(
  summary(lm5.1.2)$coefficients,
  caption = 'Linear regression between y and treatment and 3 PCs on unmatched data',
  digits = 3
)
```

Table 17: 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$y, factor(pca3data$tr, levels = c(0, 1)))$estimate
```

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

5.1.3 Cohen's d for each covariate by tr

```
cohen.d(pca3data$PC1,factor(pca3data$tr, levels = c(0, 1)))$estimate
```

```
##         0
## -1.021873
```

```
cohen.d(pca3data$PC2,factor(pca3data$tr, levels = c(0, 1)))$estimate
```

```
##         0
## -0.919134
```

```
cohen.d(pca3data$PC3,factor(pca3data$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.9820388
```

Section 5.2 Regression on matched data

5.2.1 `y ~ tr` on matched data

```
#1 propensity score matching - one to one match
matcheddata5 <- match.data(
  matchit(tr ~ PC1 + PC2 + PC3,
          data = pca3data,
          method = "nearest",
          ratio = 1))

#2 combine the PCs data (including PC1 - PC3) and original data (including X1 - X9)
matcheddata5 <- cbind(matcheddata5, dat[rownames(matcheddata5),-c(which(colnames(dat) == "y"), which(col

#3 linear regression - y and treatment
lm5.2.1 <- lm(y ~ tr, data = matcheddata5)
knitr::kable(
  summary(lm5.2.1)$coefficients,
  caption = 'Linear regression between y and treatment on matched data',
  digits = 3
)
```

Table 18: Linear regression between y and treatment on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.402	0.317	17.061	0
tr	8.150	0.448	18.200	0

```
#4 Cohen's d
cohen.d(matcheddata5$y, factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

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

5.2.2 y ~ tr + 3PC on matched data

```
#1 linear regression - y  treatment and covariates
lm5.2.2 <- lm(y ~ tr + PC1 + PC2 + PC3,
           data = matcheddata5)
knitr::kable(
  summary(lm5.2.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 3 PCs on matched data',
  digits = 3
)
```

Table 19: Linear regression between y ,treatment and 3 PCs on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.238	0.191	-1.246	0.213
tr	3.156	0.224	14.093	0.000
PC1	3.188	0.088	36.246	0.000
PC2	3.268	0.100	32.597	0.000
PC3	-3.955	0.112	-35.433	0.000

```
#2 Cohen's d
cohen.d(matcheddata5$y, matcheddata5$tr)$estimate
```

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

5.2.3 Cohen's d for each covariate by tr

```
#1 Cohen's d for each PCs by tr
cohen.d(matcheddata5$PC1,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata5$PC2,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata5$PC3,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

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

```
#2 Cohen's d for each covariate by tr
cohen.d(matcheddata5$X1,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.1628375
```

```
cohen.d(matcheddata5$X2,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.3913767
```

```
cohen.d(matcheddata5$X3,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.1850683
```

```
cohen.d(matcheddata5$X4,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2790684
```

```
cohen.d(matcheddata5$X5,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2301129
```

```
cohen.d(matcheddata5$X6,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2492141
```

```
cohen.d(matcheddata5$X7,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.4864719
```

```
cohen.d(matcheddata5$X8,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.4135808
```

```
cohen.d(matcheddata5$X9,factor(matcheddata5$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.3681363
```

Part 6 Separately Integrating 9 Covariates into 3 sets of Principal Components

This part separately integrates the 9 covariates into 3 sets principal components.[5] 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.

Section 6.1 Regression on unmatched data

6.1.1 principal component analysis - 3 sets

```
# principal component analysis - 3 sets
pca6.1 <- prcomp(dat[,paste("X", 1:3, sep = "")], scale = FALSE)
pca6.2 <- prcomp(dat[,paste("X", 4:6, sep = "")], scale = FALSE)
pca6.3 <- prcomp(dat[,paste("X", 7:9, sep = "")], scale = FALSE)
pca6data <- data.frame(
  dat$y,
  dat$tr,
  pca6.1$x[,1],
  pca6.2$x[,1],
  pca6.3$x[,1]
  )#extract the  three PCs, y and tr
names(pca6data) <- c("y", "tr", "PC1", "PC2", "PC3")

# Set 1: X1 ~ X3
## 1.1 The standard deviation of the principle components
knitr::kable(
  pca6.1$sdev,
  caption = 'The standard deviation of the principle components set 1',
  digits = 3
)
```

```
## Warning in kable_markdown(x =
## structure(c("1.202", "0.892", "0.840"), .Dim
## = c(3L, : The table should have a header
## (column names)
```

1.202
0.892
0.840

```
## 1.2 The matrix of variable loadings (columns are eigenvectors)
```

[5] 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.

```r
knitr::kable(
  pca6.1$rotation,
  caption = 'The matrix of variable loadings set1',
  digits = 3
)
```

Table 21: The matrix of variable loadings set1

	PC1	PC2	PC3
X1	-0.606	0.036	0.795
X2	-0.598	0.639	-0.485
X3	-0.525	-0.769	-0.366

```r
## 1.3 The variable means
knitr::kable(
  pca6.1$center,
  caption = 'The variable means set 1',
  digits = 3
)
```

Table 22: The variable means set 1

X1	-0.057
X2	-0.047
X3	-0.024

```r
# Set 2: X4 ~ X6
## 2.1 The standard deviation of the principle components
knitr::kable(
  pca6.2$sdev,
  caption = 'The standard deviation of the principle components set 2',
  digits = 3
)
```

```
## Warning in kable_markdown(x =
## structure(c("1.410", "0.713", "0.695"), .Dim
## = c(3L, : The table should have a header
## (column names)
```

1.410
0.713
0.695

```
## 2.2 The matrix of variable loadings (columns are eigenvectors)
knitr::kable(
  pca6.2$rotation,
  caption = 'The matrix of variable loadings set2',
  digits = 3
)
```

Table 24: The matrix of variable loadings set2

	PC1	PC2	PC3
X4	0.568	-0.509	-0.646
X5	0.590	-0.294	0.752
X6	0.573	0.809	-0.133

```
## 2.3 The variable means
knitr::kable(
  pca6.2$center,
  caption = 'The variable means set 2',
  digits = 3
)
```

Table 25: The variable means set 2

X4	0.020
X5	-0.015
X6	-0.026

```
# Set 3: X7 ~ X9
## 3.1 The standard deviation of the principle components
knitr::kable(
  pca6.3$sdev,
  caption = 'The standard deviation of the principle components set 3',
  digits = 3
)
```

```
## Warning in kable_markdown(x =
## structure(c("1.587", "0.460", "0.433"), .Dim
## = c(3L, : The table should have a header
## (column names)
```

1.587
0.460
0.433

```
## 3.2 The matrix of variable loadings (columns are eigenvectors)
knitr::kable(
  pca6.3$rotation,
  caption = 'The matrix of variable loadings set 3',
  digits = 3
)
```

Table 27: The matrix of variable loadings set 3

	PC1	PC2	PC3
X7	0.581	-0.589	0.562
X8	0.579	0.784	0.223
X9	0.572	-0.196	-0.797

```
## 3.3 The variable means
knitr::kable(
  pca6.3$center,
  caption = 'The variable means set 3',
  digits = 3
)
```

Table 28: The variable means set 3

X7	0.034
X8	0.000
X9	0.020

6.1.2 y ~ tr + 3PCs on unmatched data

```
#2 Linear regression on unmatched data
lm6.1.2 <- lm(y ~ tr + PC1 + PC2 + PC3, data = pca6data)

# summary the output
knitr::kable(
  summary(lm6.1.2)$coefficients,
  caption = 'Linear regression between y and treatment and 3 PCs on unmatched data',
  digits = 3
)
```

Table 29: 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(pca6data$y, factor(pca6data$tr, levels = c(0, 1)))$estimate
```

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

6.1.3 Cohen's d for each covariate by tr

```
cohen.d(pca6data$PC1,factor(pca6data$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(pca6data$PC2,factor(pca6data$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(pca6data$PC3,factor(pca6data$tr, levels = c(0, 1)))$estimate
```

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

Section 6.2 Regression on matched data

6.2.1 `y ~ tr` on matched data

```
#1 propensity score matching
matcheddata6 <- match.data(
  matchit(tr ~ PC1 + PC2 + PC3,
          data = pca6data,
          method = "nearest",
          ratio = 1))

#2 combine the PCs data (including PC1 - PC3) and original data (including X1 - X9)
matcheddata6 <- cbind(matcheddata6, dat[rownames(matcheddata6),-c(which(colnames(dat) == "y"), which(co
#2 linear regression - y and treatment
lm6.2.1 <- lm(y ~ tr, data = matcheddata6)
knitr::kable(
  summary(lm6.2.1)$coefficients,
  caption = 'Linear regression between y and treatment on matched data',
  digits = 3
)
```

Table 30: Linear regression between y and treatment on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.349	0.319	16.767	0
tr	8.202	0.451	18.177	0

```
#3 Cohen's d
cohen.d(matcheddata6$y, factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

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

6.2.2 y ~ tr + 3PC on matched data

```
#1 linear regression - y  treatment and covariates
lm6.2.2 <- lm(y ~ tr + PC1 + PC2 + PC3,
            data = matcheddata6)
knitr::kable(
  summary(lm6.2.2)$coefficients,
  caption = 'Linear regression between y ,treatment and 3 PCs on matched data',
  digits = 3
)
```

Table 31: Linear regression between y ,treatment and 3 PCs on matched data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.363	0.187	-1.941	0.053
tr	3.099	0.220	14.105	0.000
PC1	-3.623	0.104	-34.715	0.000
PC2	3.437	0.100	34.425	0.000
PC3	3.558	0.090	39.429	0.000

```
#2 Cohen's d
```

```
cohen.d(matcheddata6$y, factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

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

6.2.3 Cohen's d for each covariate by tr

```
#1 Cohen's d for each PCs by tr
cohen.d(matcheddata6$PC1,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata6$PC2,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

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

```
cohen.d(matcheddata6$PC3,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

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

```
#2 Cohen's d for each covariate by tr
cohen.d(matcheddata6$X1,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.1665416
```

```
cohen.d(matcheddata6$X2,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.3514834
```

```
cohen.d(matcheddata6$X3,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.1648639
```

```
cohen.d(matcheddata6$X4,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2886547
```

```
cohen.d(matcheddata6$X5,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2422791
```

```
cohen.d(matcheddata6$X6,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.2069675
```

```
cohen.d(matcheddata6$X7,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.5109331
```

```
cohen.d(matcheddata6$X8,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

```
##        0
## 0.433462
```

```
cohen.d(matcheddata6$X9,factor(matcheddata6$tr, levels = c(0, 1)))$estimate
```

```
##         0
## 0.3930671
```

Summary Table of the above information

Statistics	No matching	Nine Covariates	Seven Covariates(X_1 - X_7)	Five Covariates (X_1 - X_4, X_7)	Three Covariates (X_1, X_4, X_7)	3 PCs(1 decomp)
y ~ tr + covariates, tr coefficient (pre-match)	3.05	3.05	3.81	5.58	7.14	3.2
cohen's d, x1 ~ tr	-0.59	0.26	0.28	0.22	0.2	0.16
cohen's d, x2 ~ tr	-0.63	0.33	0.27	0.23	0.76	0.39
cohen's d, x3 ~ tr	-0.47	0.2	0.24	0.13	0.58	0.19
cohen's d, x4 ~ tr	-0.71	0.26	0.2	0.23	0.13	0.28

Statistics	No matching	Nine Covariates	Seven Covariates(X_1 - X_7)	Five Covariates (X_1 - X_4, X_7)	Three Covariates (X_1, X_4, X_7)	3 PCs(1 decomp)
cohen's d, x5 ~ tr	-0.78	0.26	0.28	0.74	0.63	0.23
cohen's d, x6 ~ tr	-0.75	0.15	0.12	0.68	0.64	0.25
cohen's d, x7 ~ tr	-1.01	0.45	0.43	0.34	0.27	0.49
cohen's d, x8 ~ tr	-1	0.47	0.68	0.61	0.58	0.41
cohen's d, x9 ~ tr	-1.01	0.41	0.64	0.52	0.48	0.37
y ~ tr, tr coef	16.19	8.26	8.6	9.24	9.83	8.15
y ~ tr + covariates, tr coef	3.05	3.03	4.09	6.16	7.86	3.16