성향점수 분석의 확장

세 군 이상의 범주형 변수, 연속형 변수 대상

문건웅

2024-1-27

필요한 패키지 설치

```
install.packages(c("twang","twangContinuous"))
devtools::install_github("cardiomoon/webrPSM")
```

세 군 이상의 범주형 변수

twang package

- Toolkit for Weighting and Analysis of Nonequivalent Groups(twang)
- 성향점수를 추정하기 위해 그레디언트 부스트 모형(gradient boosted models)을 사용한다.
- 그레디언트 부스트 소개
  - https://bkshin.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-15-Gradient-Boost
- 공변량 사이의 상호작용과 비선형성을 자동으로 통합해주는 유연한 기계학습 방법
- twang 패키지는 가장 좋은 균형성을 달성하기 위해 그레디언트 부스트 모형을 반복적으로 선택하여 처치군과 대조군 사이의 유사성을 최적화한다.
- 사용자가 균형성을 달성하고자 하는 공변량만 선택해주면 알고리즘이 최적의 균형성을 확보하기 위해 비선형성과 상호작용을 통합해준다.

Typical workflow with twang

1. 추정하고자 하는 통계량 결정 : ATE, ATT

2. 균형성을 달성해야 하는 관측된 공변량 결정

3. ps(), mnps() 함수로 성향점수 모형 적합

    - 알고리즘이 수렴하는지 평가
    - 성향점수 가중치 적용 전후 공변량 균형성 평가
    - 필요하다면 그레디언트 부스트 인수를 조정하여 다시 알고리즘 수행

4. 처치효과 추정

    - 성향점수 가중치를 추출하여 가중치가 적용된 모형에 적합시킨다.

세 군 이상의 범주형 변수 예제

데이터 : AOD

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3710547/
- treat: 대상자의 치료방법. community, metcbt5, scy
  - community : ten “exemplary” community-based care treatment programs from from the Adolescent Treatment Model (ATM) program(1998-9)
  - MET/CBT-5 : Motivational Enhancement Therapy plus Cognitive Behavior Therapy(2003)
  - SCY : Strengthening Communities for Youth(2001-2)
- suf12: outcome 변수: substance use frequency at 12 month follow-up
- 공변량
  - illact : illicit activities scale
  - crimjust : criminal justice involvement
  - subdep : substance use dependence scale
  - white : 1 if non-Hispanic white, 0 otherwise

성향점수 추정

```
library(twang)
data(AOD)
set.seed(1)
out <- mnps (treat ~ illact + crimjust + subprob + subdep + white,data=AOD,
             estimand = "ATE",  # "ATE" or "ATT"
             #treatATT = "community", # which treatment condition is considered 'the treated'
             verbose = FALSE,   # 화면에 자세한 정보 출력
             stop.method=c("es.max"),  # Default "es.max", can select c("es.mean","ks.max")
             n.trees = 3000,    # Default 10000
             interaction.depth=3,  # Default
             shrinkage=0.01,     # Default
             n.minobsinnode=10)  # Default
```

그레디언트 부스팅을 조절하는 인수들

- stop.method :
  - es : absolute standardized mean difference (=standardized effect size)
  - ks : Kolmogorov-Smirnov statistic
  - 각각 mean과 max 선택 가능
- n.trees : number of gbm iterations passed on to gbm. Default: 10000
- interaction.depth :
  - 그레디언트 부스팅의 트리 깊이를 결정
  - 상호작용에 포함되는 변수의 최대갯수
  - 이 값이 커지면 모형의 복잡도가 증가한다
  - Default: 3
- shrinkage :
  - learning rate(0-1)
  - 그레디언트 부스트가 반복될 때 더해지는 값
  - Learnig rate가 적으면 반복횟수가 많아져야 하지만 과적합을 어느 정도 막을 수 있다
  - Default 0.01.
- n.minobsinnode :
  - 종말노드에 포함되는 관측치의 최소갯수
  - 값이 적으면 모형의 복잡도가 증가하고 값이 클 경우 과적합을 어느 정도 막을 수 있다.
  - Default 10.

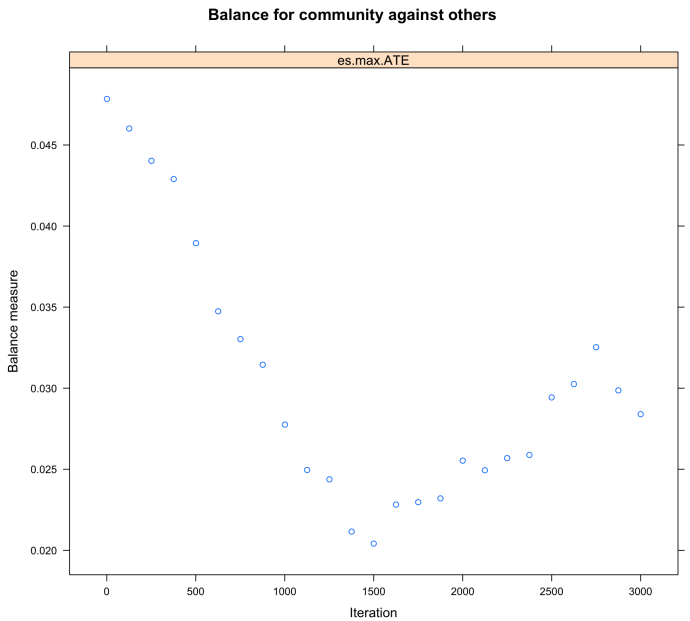
결과 요약

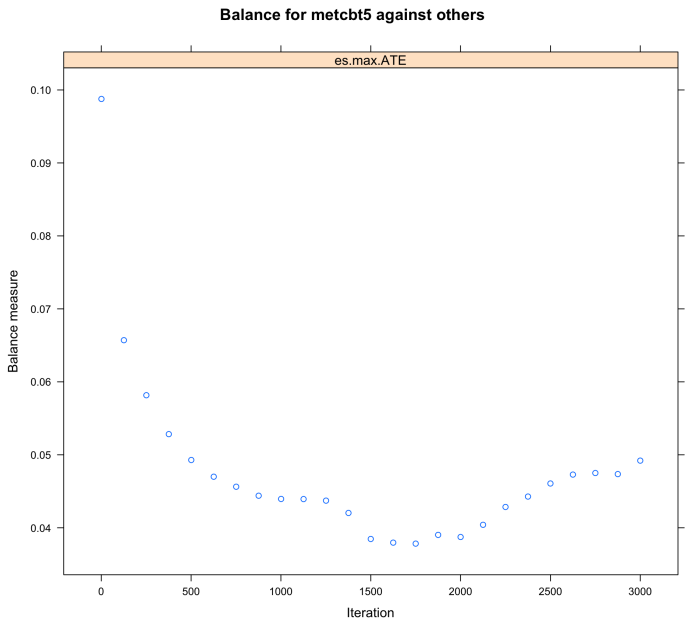
```
summary(out)
```

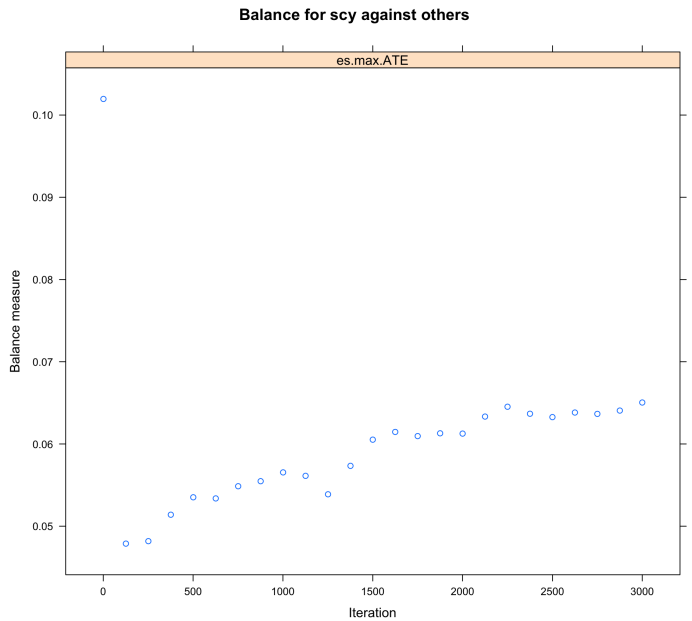
```
Summary of pairwise comparisons:
  max.std.eff.sz      min.p     max.ks min.ks.pval stop.method
1     0.20266446 0.04161562 0.13000000  0.06809222         unw
2     0.07617111 0.43588977 0.08190495  0.54700613      es.max

Sample sizes and effective sample sizes:
  treatment   n ESS:es.max
1 community 200   185.8713
2   metcbt5 200   183.2522
3       scy 200   197.1929
```

알고리즘의 반복횟수가 충분한가?

```
result=plot(out,plots=1,print=FALSE)
result[[1]]
result[[2]]
result[[3]]
```







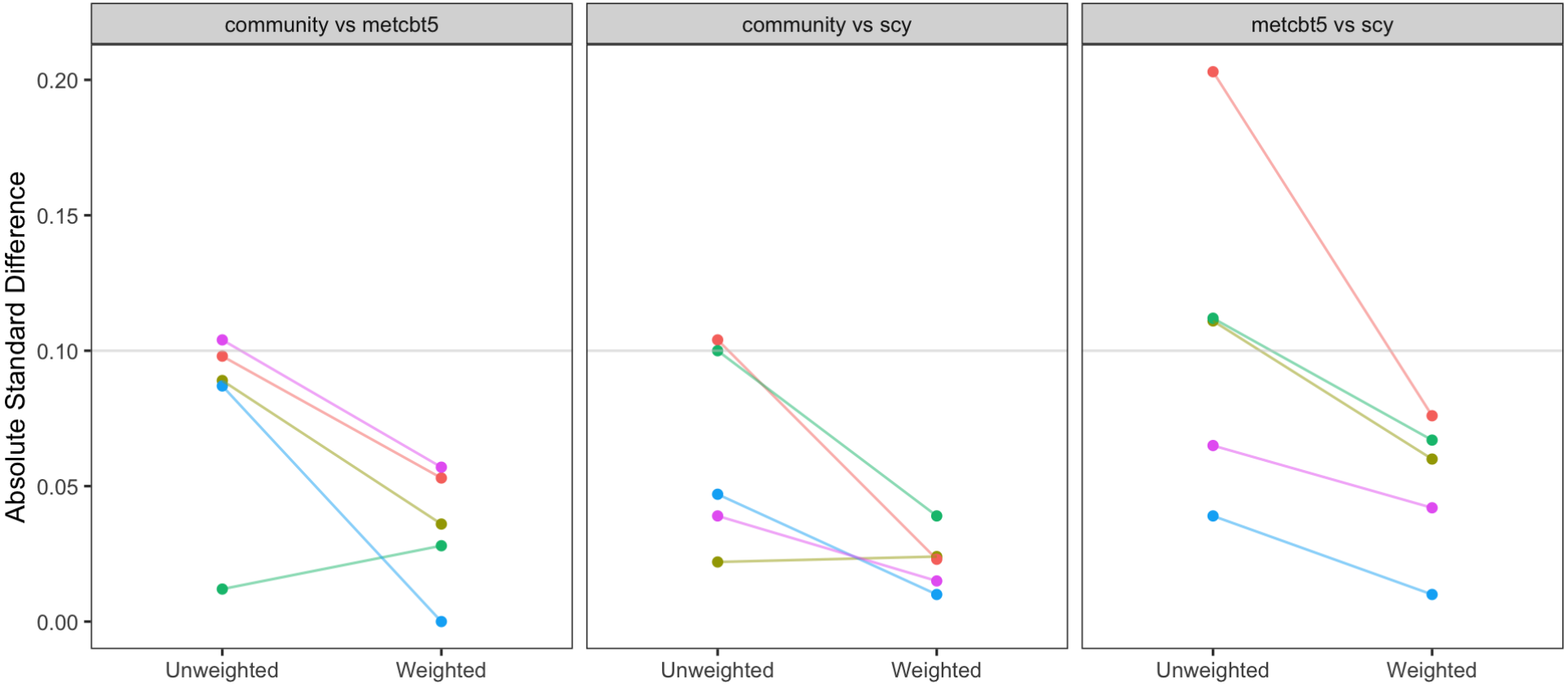
성향점수 가중치 적용 전후 공변량 균형성 평가

```
bal.table(out)
```

	tmt1	tmt2	var	mean1	mean2	pop.sd	std.eff.sz	p	ks	ks.pval	stop.method
1	community	metcbt5	illact	0.097	0.007	1.014	0.089	0.385	0.100	0.270	unw
2	community	metcbt5	crimjust	-0.065	0.037	1.041	0.098	0.328	0.105	0.220	unw
3	community	metcbt5	subprob	-0.060	0.026	0.985	0.087	0.390	0.090	0.393	unw
4	community	metcbt5	subdep	0.046	0.058	1.031	0.012	0.910	0.055	0.923	unw
5	community	metcbt5	white	0.160	0.200	0.383	0.104	0.298	0.040	0.997	unw
6	community	scy	illact	0.097	0.120	1.014	0.022	0.823	0.060	0.864	unw
7	community	scy	crimjust	-0.065	-0.174	1.041	0.104	0.295	0.080	0.544	unw
8	community	scy	subprob	-0.060	-0.013	0.985	0.047	0.631	0.090	0.393	unw
9	community	scy	subdep	0.046	-0.058	1.031	0.100	0.312	0.085	0.465	unw
10	community	scy	white	0.160	0.175	0.383	0.039	0.688	0.015	1.000	unw
11	metcbt5	scy	illact	0.007	0.120	1.014	0.111	0.259	0.110	0.178	unw
12	metcbt5	scy	crimjust	0.037	-0.174	1.041	0.203	0.042	0.130	0.068	unw
13	metcbt5	scy	subprob	0.026	-0.013	0.985	0.039	0.696	0.065	0.792	unw
14	metcbt5	scy	subdep	0.058	-0.058	1.031	0.112	0.251	0.090	0.393	unw
15	metcbt5	scy	white	0.200	0.175	0.383	0.065	0.523	0.025	1.000	unw

```
16 community metcbt5   illact  0.084  0.048  1.014      0.036 0.711 0.064   0.849      es.max
17 community metcbt5 crimjust -0.088 -0.033  1.041      0.053 0.586 0.058   0.916      es.max
18 community metcbt5  subprob -0.005 -0.005  0.985      0.000 1.000 0.058   0.920      es.max
19 community metcbt5   subdep  0.009  0.037  1.031      0.028 0.785 0.047   0.986      es.max
20 community metcbt5    white  0.171  0.193  0.383      0.057 0.590 0.022   1.000      es.max
21 community     scy   illact  0.084  0.108  1.014      0.024 0.811 0.063   0.839      es.max
22 community     scy crimjust -0.088 -0.112  1.041      0.023 0.814 0.050   0.971      es.max
23 community     scy  subprob -0.005 -0.015  0.985      0.010 0.918 0.066   0.795      es.max
24 community     scy   subdep  0.009 -0.031  1.031      0.039 0.692 0.048   0.981      es.max
25 community     scy    white  0.171  0.177  0.383      0.015 0.884 0.006   1.000      es.max
26   metcbt5     scy   illact  0.048  0.108  1.014      0.060 0.545 0.082   0.547      es.max
27   metcbt5     scy crimjust -0.033 -0.112  1.041      0.076 0.436 0.067   0.787      es.max
28   metcbt5     scy  subprob -0.005 -0.015  0.985      0.010 0.921 0.043   0.995      es.max
29   metcbt5     scy   subdep  0.037 -0.031  1.031      0.067 0.497 0.065   0.824      es.max
30   metcbt5     scy    white  0.193  0.177  0.383      0.042 0.684 0.016   1.000      es.max
```

```
library(webrPSM)
balancePlotTwang(out)
```



Estimating Treatment Effect

```
library(survey)
AOD$wts <- twang::get.weights(out)
design.ps <- svydesign(ids=~1,weights=~wts,data=AOD)
glm1 <- svyglm(suf12 ~ as.factor(treat), design=design.ps)
summary(glm1)
```

```
Call:
svyglm(formula = suf12 ~ as.factor(treat), design = design.ps)

Survey design:
svydesign(ids = ~1, weights = ~wts, data = AOD)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.09609    0.06777  -1.418    0.157
as.factor(treat)metcbt5  0.15759    0.10390   1.517    0.130
as.factor(treat)scy      0.07686    0.09912   0.775    0.438

(Dispersion parameter for gaussian family taken to be 0.9879139)

Number of Fisher Scoring iterations: 2
```

예제 1: 웹R을 이용한 성향점수 분석

AOD 데이터를 이용하여 treat를 그룹변수로 하고 공변량으로 illact, crimjust, subprob, subdep, white를 선택하여 성향점수를 추정하고 이를 이용하여 연속형 변수인 suf12를 outcome 변수로 하여 treat변수의 효과(average treatment effect, ATE)를 추정하라.

예제 1: 웹R을 이용한 성향점수 분석

AOD 데이터를 이용하여 treat를 그룹변수로 하고 공변량으로 illact, crimjust, subprob, subdep, white를 선택하여 성향점수를 추정하고 이를 이용하여 연속형 변수인 suf12를 outcome 변수로 하여 treat변수의 효과(average treatment effect, ATE)를 추정하라.

```
$`Ref : community`
                     Estimate Std. Error    t value  Pr(>|t|)       lower       upper
(Intercept)       -0.09609053 0.06776711 -1.4179524 0.1567264 -0.22891162   0.03673055
metcbt5-community  0.15759222 0.10389998  1.5167686 0.1298544 -0.04604799   0.36123243
scy-community      0.07686456 0.09911842  0.7754821 0.4383619 -0.11740397   0.27113308

$`Ref : metcbt5`
                     Estimate Std. Error    t value  Pr(>|t|)       lower       upper
(Intercept)        0.06150169 0.07875801  0.7808943 0.4351742 -0.09286118   0.21586456
community-metcbt5 -0.15759222 0.10389998 -1.5167686 0.1298544 -0.36123243   0.04604799
scy-metcbt5       -0.08072766 0.10693411 -0.7549290 0.4505893 -0.29031467   0.12885935

$`Ref : scy`
                   Estimate Std. Error    t value  Pr(>|t|)      lower     upper
(Intercept)    -0.01922597 0.07233312 -0.2657977 0.7904868 -0.1609963 0.1225443
community-scy -0.07686456 0.09911842 -0.7754821 0.4383619 -0.2711331 0.1174040
metcbt5-scy    0.08072766 0.10693411  0.7549290 0.4505893 -0.1288593 0.2903147
```

```
$`causal effect of each tx relative to the average potential outcome of all tx.`
               Estimate Std. Error    t value  Pr(>|t|)       lower      upper
(Intercept) -0.01793827 0.04219964 -0.4250812 0.6709306 -0.10064956 0.06477302
community   -0.07815226 0.05754653 -1.3580708 0.1749542 -0.19094346 0.03463894
metcbt5      0.07943996 0.06203562  1.2805541 0.2008476 -0.04214985 0.20102977
scy         -0.00128770 0.05937033         NA        NA -0.11765355 0.11507815
```

예제 2

survival package의 대장암 데이터(colon)를 이용하여 rx를 그룹변수로 하고 공변량으로 sex, age, obstruct, perfor, adhere, nodes, differ, extent, surg 를 선택하여 성향점수를 추정하고 이를 이용하여 상태변수인 status 와 시간변수인 time 을 사용하여 생존변수를 만들고 Observation군(rx=Obs)과 비교한 Lev, Lev+5FU 군의 Hazard Ratio를 추정하라.

예제 2

survival package의 대장암 데이터(colon)를 이용하여 rx를 그룹변수로 하고 공변량으로 sex, age, obstruct, perfor, adhere, nodes, differ, extent, surg 를 선택하여 성향점수를 추정하고 이를 이용하여 상태변수인 status 와 시간변수인 time 을 사용하여 생존변수를 만들고 Observation군(rx=Obs)과 비교한 Lev, Lev+5FU 군의 Hazard Ratio를 추정하라.

```
Call:
coxph(formula = survival::Surv(time, , status) ~ rx, data = data1,
    weights = w)

  n= 1858, number of events= 920

               coef exp(coef) se(coef) robust se      z Pr(>|z|)
rxLev     -0.01543   0.98469  0.04629   0.07866 -0.196    0.844
rxLev+5FU -0.40310   0.66824  0.04931   0.08471 -4.758 1.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
rxLev        0.9847      1.016     0.844    1.1488
rxLev+5FU    0.6682      1.496     0.566    0.7889

Concordance= 0.541  (se = 0.009 )
Likelihood ratio test= 85.94  on 2 df,   p=<2e-16
Wald test            = 27.23  on 2 df,   p=1e-06
Score (logrank) test = 82.55  on 2 df,   p=<2e-16,   Robust = 28.94  p=5e-07
```

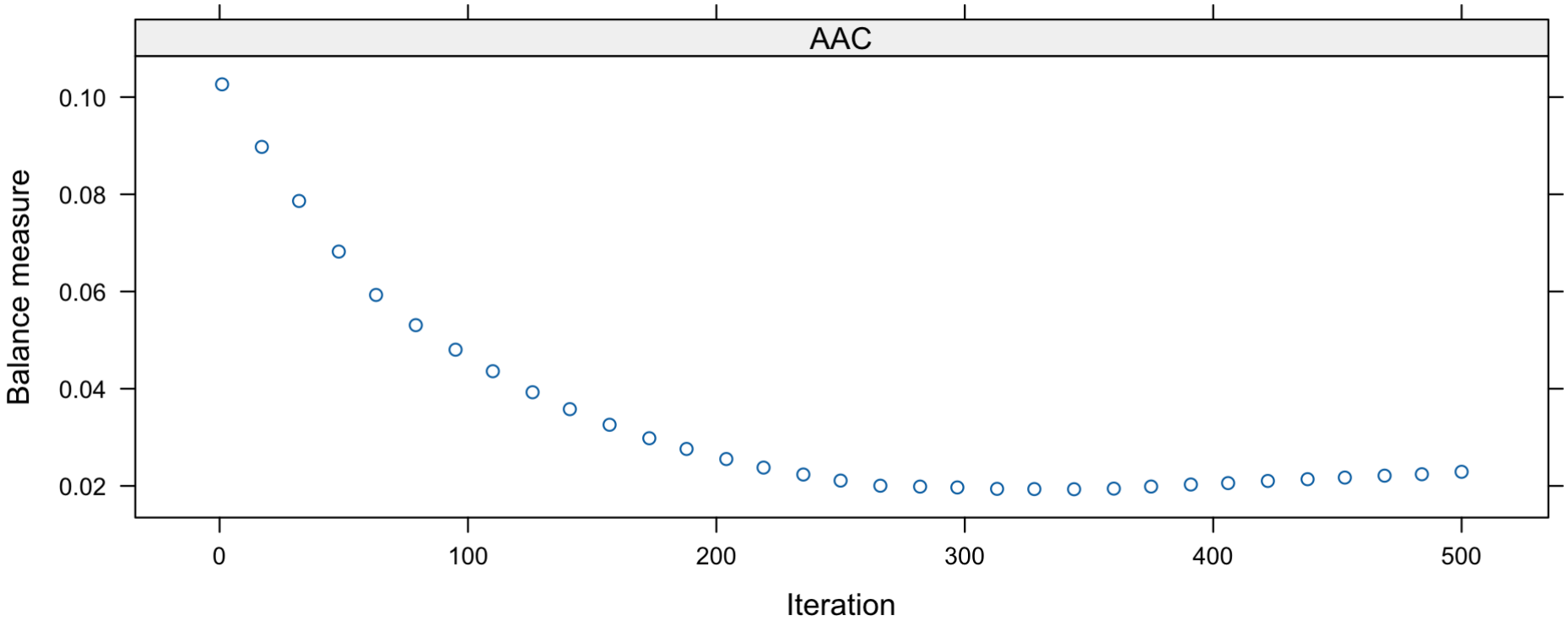
연속형 변수 대상 성향점수 분석

연속형 변수 대상 성향점수 분석

```
# install.packages("twangContinuous")
library(twangContinuous)
set.seed(1234)
data(dat)
test.mod <- ps.cont(tss_0 ~ sfs8p_0 + sati_0 + sp_sm_0 + recov_0 + subsgrps_n + treat,
                    data=dat,
                    n.trees = 500,
                    shrinkage = 0.01,
                    interaction.depth = 3,
                    verbose = FALSE)
```

반복횟수는 충분한가?

```
plot(test.mod, plots="optimize")   # Average absolute correlation
```



```
summary(test.mod)
```

```
       n      ess   max.wcor mean.wcor   rms.wcor iter
unw 4000 4000.000 0.21633979 0.1034782 0.11887909   NA
AAC 4000 3559.661 0.04680874 0.0192826 0.02492979  347
```

- ess: effective sample size
- max.wcor: maximum absolute correlation
- mean.wcor: mean or average absolute correlation
- rms.wcor: root mean square of the absolute correlation

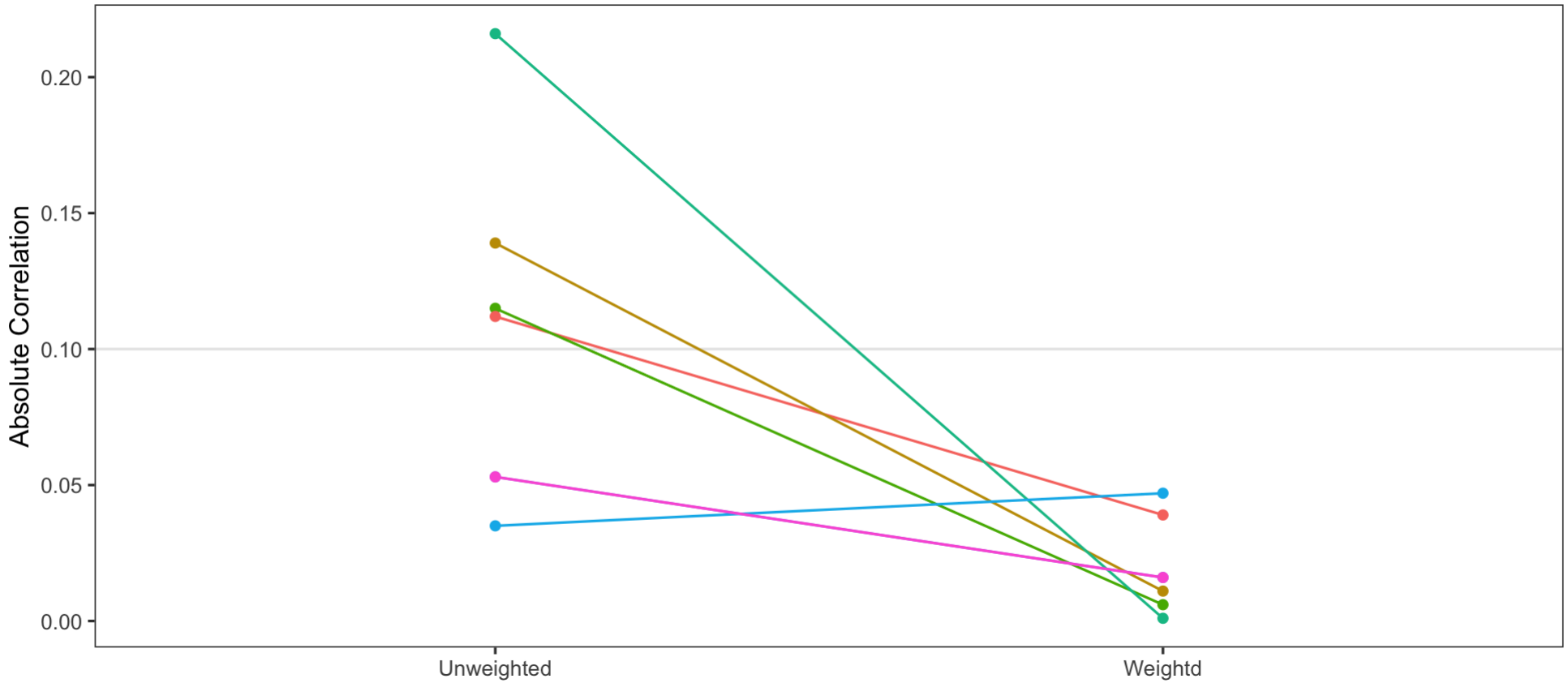
균형성 점검

```
bal.table(test.mod, digits=3)
```

```
              unw   wcor
sfs8p_0     0.115 -0.006
sati_0      0.139 -0.011
sp_sm_0     0.216  0.001
recov_0    -0.112 -0.039
subsgrps_n  0.035 -0.047
treatA      0.053  0.016
treatB     -0.053 -0.016
```

균형성점검

```
balancePlot(test.mod)
```



처치효과 추정

```
library(survey)
dat$wts <- get.weights(test.mod)
design.ps <- svydesign(ids=~1, weights=~wts, data=dat)
outcome.model <- svyglm(sfs8p_3 ~ tss_0, design = design.ps, family = gaussian())
summary(outcome.model)
```

```
Call:
svyglm(formula = sfs8p_3 ~ tss_0, design = design.ps, family = gaussian())

Survey design:
svydesign(ids = ~1, weights = ~wts, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.595461   0.232069  28.420   <2e-16 ***
tss_0       0.002616   0.062416   0.042    0.967
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 111.6769)

Number of Fisher Scoring iterations: 2
```

예제 3

twangContinuous 패키지의 dat 데이터에서 연속형 변수인 tss_0을 처치변수로 하고 sfs8p_0,sati_0,sp_sm_0,recov_0,subsgrps_n,treat를 공변량으로 하여 성향점수를 구하고 sfs8p_3을 outcome variable로 하는 처치효과를 구하라.

예제 3

twangContinuous 패키지의 dat 데이터에서 연속형 변수인 tss_0을 처치변수로 하고 sfs8p_0,sati_0,sp_sm_0,recov_0,subsgrps_n,treat를 공변량으로 하여 성향점수를 구하고 sfs8p_3을 outcome variable로 하는 처치효과를 구하라.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.595461   0.232069  28.420   <2e-16 ***
tss_0       0.002616   0.062416   0.042    0.967

                 2.5 %   97.5 %
(Intercept)  6.1406144 7.050308
tss_0       -0.1197171 0.124950
```

The results indicate that for each additional trauma symptom, substance use frequency increases by 0.003, which is **not** statistically significant, t=0.042,p=.967.

예제 4

simData2 데이터에서 treat를 처치변수로 하고 x1, x2를 공변량으로 하여 성향점수를 구하고 연속형 변수인 y를 outcome variable로 하는 처치효과를 구하라.

예제 4

simData2 데이터에서 treat를 처치변수로 하고 x1, x2를 공변량으로 하여 성향점수를 구하고 연속형 변수인 y를 outcome variable로 하는 처치효과를 구하라.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.86983    0.42292   2.057   0.0400 *
treat        0.20837    0.06417   3.247   0.0012 **

                2.5 %    97.5 %
(Intercept) 0.03991163 1.6997577
treat       0.08244833 0.3342978
```

강의가 끝났습니다. 수고 많으셨습니다!