

R을 이용한 성향점수분석 실습

문건웅

2024-1-27

필요한 패키지 설치

```
install.packages(c("MatchIt", "optmatch", "cobalt", "lmtest", "sandwich", "boot", "devtools"))  
install.packages("https://cran.r-project.org/src/contrib/Archive/Zelig/Zelig_5.1.7.tar.gz",  
                 repos=NULL, type="source")  
devtools::install_github("cardiomoon/webrPSM")
```

패키지 불러오기

```
library(MatchIt)  
library(cobalt)  
library(lmtest)      # coeftest  
library(sandwich)    # vcovCL  
library(boot)        # boot()  
library(webrPSM)
```

분석의 목표

직업훈련프로그램(75년 - 78년 실시)의 임금인상(78년 측정, re78) 효과

데이터 확인

```
str(lalonde)
```

```
'data.frame':    614 obs. of  9 variables:
 $ treat    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ age      : int  37 22 30 27 33 22 23 32 22 33 ...
 $ educ     : int  11 9 12 11 8 9 12 11 16 12 ...
 $ race     : Factor w/ 3 levels "black","hispan",...: 1 2 1 1 1 1 1 1 1 3 ...
 $ married  : int  1 0 0 0 0 0 0 0 0 1 ...
 $ nodegree: int  1 1 0 1 1 1 0 1 0 0 ...
 $ re74     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ re75     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ re78     : num  9930 3596 24909 7506 290 ...
```

첫번째 매칭 :Nearest Matching(1:1)

```
out1 <- matchit(treat ~ age + educ + race + married + nodegree + re74 + re75, data = lalonde,  
               method = "nearest", distance = "glm")  
out1
```

A matchit object

- method: 1:1 nearest neighbor matching without replacement
- distance: Propensity score
 - estimated with logistic regression
- number of obs.: 614 (original), 370 (matched)
- target estimand: ATT
- covariates: age, educ, race, married, nodegree, re74, re75

공변량균형성 점검

```
summary(out1)
```

Call:

```
matchit(formula = treat ~ age + educ + race + married + nodegree +  
        re74 + re75, data = lalonde, method = "nearest", distance = "glm")
```

Summary of Balance for All Data:

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max
distance	0.5774	0.1822	1.7941	0.9211	0.3774	0.6444
age	25.8162	28.0303	-0.3094	0.4400	0.0813	0.1577
educ	10.3459	10.2354	0.0550	0.4959	0.0347	0.1114
raceblack	0.8432	0.2028	1.7615	.	0.6404	0.6404
racehispan	0.0595	0.1422	-0.3498	.	0.0827	0.0827
racewhite	0.0973	0.6550	-1.8819	.	0.5577	0.5577
married	0.1892	0.5128	-0.8263	.	0.3236	0.3236
nodegree	0.7081	0.5967	0.2450	.	0.1114	0.1114
re74	2095.5737	5619.2365	-0.7211	0.5181	0.2248	0.4470
re75	1532.0553	2466.4844	-0.2903	0.9563	0.1342	0.2876

Summary of Balance for Matched Data:

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
distance	0.5774	0.3629	0.9739	0.7566	0.1321	0.4216	0.9740
age	25.8162	25.3027	0.0718	0.4568	0.0847	0.2541	1.3938
educ	10.3459	10.6054	-0.1290	0.5721	0.0239	0.0757	1.2474
raceblack	0.8432	0.4703	1.0259	.	0.3730	0.3730	1.0259
racehispan	0.0595	0.2162	-0.6629	.	0.1568	0.1568	1.0743
racewhite	0.0973	0.3135	-0.7296	.	0.2162	0.2162	0.8390
married	0.1892	0.2108	-0.0552	.	0.0216	0.0216	0.8281
nodegree	0.7081	0.6378	0.1546	.	0.0703	0.0703	1.0106
re74	2095.5737	2342.1076	-0.0505	1.3289	0.0469	0.2757	0.7965
re75	1532.0553	1614.7451	-0.0257	1.4956	0.0452	0.2054	0.7381

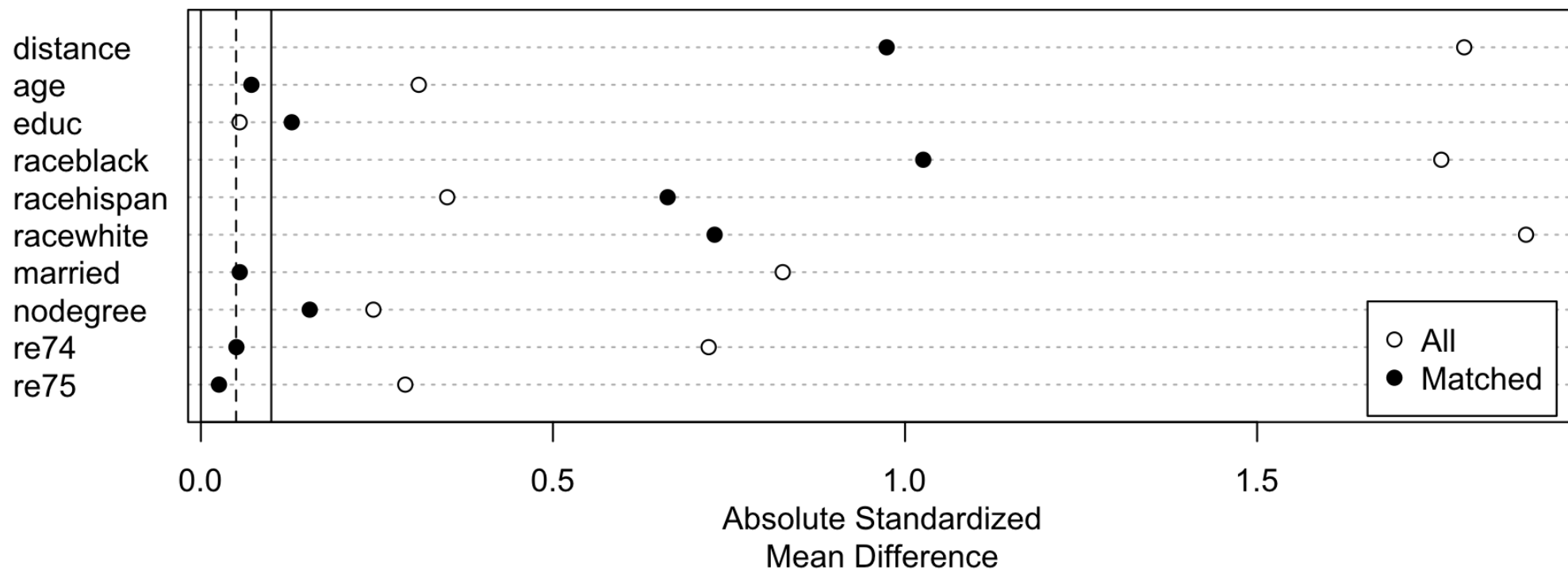
Percent Balance Improvement:

	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max
distance	45.7	-239.6	65.0	34.6
age	76.8	4.6	-4.2	-61.1
educ	-134.8	20.4	31.2	32.1
raceblack	41.8	.	41.8	41.8
racehispan	-89.5	.	-89.5	-89.5
racewhite	61.2	.	61.2	61.2
married	93.3	.	93.3	93.3
nodegree	36.9	.	36.9	36.9
re74	93.0	56.8	79.1	38.3
re75	91.2	-800.7	66.3	28.6

Sample Sizes:

	Control	Treated
All	429	185
Matched	185	185
Unmatched	244	0
Discarded	0	0

```
plot(summary(out1))
```



PSMTable(out1)

Variables	Before Propensity Score Matching				After Propensity Score Matching			
	treat=0	treat=1	p	standardized difference	treat=0	treat=1	p	standardized difference
Covariates	N=429	N=185			N=185	N=185		
age	28.0 ± 10.8	25.8 ± 7.2	0.003	-0.309	25.3 ± 10.6	25.8 ± 7.2	0.585	0.072
educ	10.2 ± 2.9	10.3 ± 2.0	0.585	0.055	10.6 ± 2.7	10.3 ± 2.0	0.290	-0.129
race			< 0.001				< 0.001	
- black	87 (20.3%)	156 (84.3%)		1.762	87 (47.0%)	156 (84.3%)		1.026
- hispan	61 (14.2%)	11 (5.9%)		-0.350	40 (21.6%)	11 (5.9%)		-0.663
- white	281 (65.5%)	18 (9.7%)		-1.882	58 (31.4%)	18 (9.7%)		-0.730
married:1	220 (51.3%)	35 (18.9%)	< 0.001	-0.826	39 (21.1%)	35 (18.9%)	0.697	-0.055
nodegree:1	256 (59.7%)	131 (70.8%)	0.011	0.245	118 (63.8%)	131 (70.8%)	0.184	0.155
re74	5619.2 ± 6788.8	2095.6 ± 4886.6	< 0.001	-0.721	2342.1 ± 4239.0	2095.6 ± 4886.6	0.605	-0.050
re75	2466.5 ± 3292.0	1532.1 ± 3219.3	0.001	-0.290	1614.7 ± 2632.4	1532.1 ± 3219.3	0.787	-0.026

```
PSMTable(out1,grouplabel=c("Control","Treated"))
```

Variables	Before Propensity Score Matching				After Propensity Score Matching			
	Control	Treated	p	standardized difference	Control	Treated	p	standardized difference
Covariates	N=429	N=185			N=185	N=185		
age	28.0 ± 10.8	25.8 ± 7.2	0.003	-0.309	25.3 ± 10.6	25.8 ± 7.2	0.585	0.072
educ	10.2 ± 2.9	10.3 ± 2.0	0.585	0.055	10.6 ± 2.7	10.3 ± 2.0	0.290	-0.129
race			< 0.001				< 0.001	
- black	87 (20.3%)	156 (84.3%)		1.762	87 (47.0%)	156 (84.3%)		1.026
- hispan	61 (14.2%)	11 (5.9%)		-0.350	40 (21.6%)	11 (5.9%)		-0.663
- white	281 (65.5%)	18 (9.7%)		-1.882	58 (31.4%)	18 (9.7%)		-0.730
married:1	220 (51.3%)	35 (18.9%)	< 0.001	-0.826	39 (21.1%)	35 (18.9%)	0.697	-0.055
nodegree:1	256 (59.7%)	131 (70.8%)	0.011	0.245	118 (63.8%)	131 (70.8%)	0.184	0.155
re74	5619.2 ± 6788.8	2095.6 ± 4886.6	< 0.001	-0.721	2342.1 ± 4239.0	2095.6 ± 4886.6	0.605	-0.050
re75	2466.5 ± 3292.0	1532.1 ± 3219.3	0.001	-0.290	1614.7 ± 2632.4	1532.1 ± 3219.3	0.787	-0.026

매칭 방법의 선택(I)

- 가장 좋은 매칭 방법은 자기가 가지고 있는 데이터의 고유한 특징과 목표로 하는 추정량에 따라 달라진다.
- 목표로 하는 추정량이 ATT, ATE, ATC 인지에 따라 다른 매칭 방법이 사용될 수 있으며 경우에 따라 특정 매칭방법이 다른 매칭 방법에 비해 보다 효과적일 수 있다.
- 치료효과를 추정하지 않았다면 여러 가지 매칭 방법을 시도해 볼 수 있다.
- 매칭방법을 선택하는 기준은 균형성과 매칭 후 유효표본크기(effective sample size, ESS)이다.
- 한가지 매칭 방법을 시도해본 후 균형성이 좋지 않거나 유효표본크기가 너무 적은 경우 만족할 수 있을 때까지 다른 방법을 시도해 보아야 한다.
- 균형성 점검은 광범위하게 해 보아야 하고 공변량 균형성 달성 기준에 도달했다고 하더라도 가능한 한 보다 완벽한 균형성을 보이는 다른 방법을 더 찾아보아야 한다.

매칭 방법의 선택(II)

- ATE 를 추정하고자 할 때는 전체매칭이나 서브클래스 매칭을 사용할 수 있으며 ATT나 ATC를 추정할 때는 어떤 매칭 방법도 사용 가능하다.
- 정확매칭(exact)이나 준정확매칭(cem)은 전체 공변량의 완벽한 균형을 목표로 하므로 가장 강력한 방법이다. 정확매칭을 수행할 수 있는 경우 정확매칭을 하여야 하고 연속형 공변량이 있는 경우 준정확매칭을 시도해 볼 수 있다. 이 경우 표본 손실에 유의하여야 한다. 통제집단의 풀(pool)이 처치집단에 비해 충분히 크지 않으면 일부 처치군이 손실될 수 있으며 이런 경우 추정량이 변하고 정밀도가 극도로 감소될 수 있다. 이 방법들은 아주 낙관적인 환경에서만 사용 가능하다. 다른 매칭 방법을 사용하면서 일부 공변량에 대해 정확 또는 준정확매칭을 사용할 수 있다(matchit 패키지의 exact 인수 사용).
- ATE 를 추정하고자 할 때는 전체매칭이나 서브클래스 매칭을 사용할 수 있다. 전체매칭은 균형성을 최적화하므로 보다 효과적일 수 있고 전체매칭을 하면서 일부 공변량에 대해 정확매칭을 하거나 마할라노비스거리를 사용하여 매칭할 수도 있다. 데이터셋이 큰 경우 전체 매칭은 어려울 수 있으며 이런 경우 서브클래스매칭이 보다 빠른 해결방법이 될 수 있다. 서브클래스매칭을 할 경우 서브클래스의 수를 변경해보아야 한다. 표본 수가 큰 경우 서브클래스의 수가 많을 수록 보다 수행이 좋아진다. 서브클래스의 수를 여러 개 사용해보지 않고 디폴트인 6이나 5로 설정하지는 말아야 한다.

매칭 방법의 선택(III)

- ATT 를 추정할 때는 다양한 방법을 시도해 볼 수 있다. 유전매칭은 직접 공변량 균형성을 최적화시키므로 좋은 균형성을 얻을 수 있으나 데이터가 큰 경우 시간이 오래 걸릴 수 있다.
- 최적매칭과 반복표집 없는 그리디매칭은 매우 비슷한 수행 결과를 보여주지만 데이터가 큰 경우 그리디 매칭이 선호된다. 두 가지 매칭 모두 일부 변수에 대한 정확매칭이 가능하고 마할라노비스 매칭을 사용할 수 있으며 일대다 ($k:1$)매칭을 할 수 있다. 반복표집을 하는 니어리스트 매칭과 전체매칭, 서브클래스 매칭은 통제집단에 대해 균등하지 않은 가중치를 사용하는데 균형성은 좋아지지만 모든 표본이 유지되더라도 유효표본수는 줄어 들 수 있다.
- 좋은 균형성과 높은 유효표본수를 얻기 위해 여러 방법들과 인수들을 사용하여 시도해 보아야 한다. 데이터셋의 특징에 따라 가장 좋은 방법이 달라지기 때문에 권장할 수 있는 최적의 방법은 없다.
- 처치집단을 유지하는 것이 중요하지 않을 경우 caliper 매칭을 사용하면 불균형성과 편종을 제거할 수 있으나 불완전한 매칭으로 편종을 일으킬 수 있다.

공변량 균형성 점검

- 완벽한 균형
 - 처치집단과 통제집단 사이의 공변량의 평균차이 0
 - 처치집단과 통제집단 사이의 공변량의 분산비 1
- 공변량 균형성 달성 기준
 - 처치집단과 통제집단 사이의 표준화시킨 공변량 및 성향점수의 **평균차이**가 0.10(또는 0.05) 또는 0.25보다 작아야 한다.
 - 처치집단과 통제집단 사이의 표준화시킨 공변량 및 성향점수의 **분산비**가 0.5 - 2.0 사이에 놓여야한다.

두번째 매칭 : Full-Matching with probit link function

```
out2 <- matchit(treat ~ age + educ + race + married +  
               nodegree + re74 + re75, data = lalonde,  
               method = "full",  
               link="probit",  
               distance = "glm")  
out2
```

A matchit object

- method: Optimal full matching
- distance: Propensity score
 - estimated with probit regression
- number of obs.: 614 (original), 614 (matched)
- target estimand: ATT
- covariates: age, educ, race, married, nodegree, re74, re75

공변량균형성 점검

```
summary(out2)
```

Call:

```
matchit(formula = treat ~ age + educ + race + married + nodegree +  
        re74 + re75, data = lalonde, method = "full", distance = "glm",  
        link = "probit")
```

Summary of Balance for All Data:

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max
distance	0.5773	0.1817	1.8276	0.8777	0.3774	0.6413
age	25.8162	28.0303	-0.3094	0.4400	0.0813	0.1577
educ	10.3459	10.2354	0.0550	0.4959	0.0347	0.1114
raceblack	0.8432	0.2028	1.7615	.	0.6404	0.6404
racehispan	0.0595	0.1422	-0.3498	.	0.0827	0.0827
racewhite	0.0973	0.6550	-1.8819	.	0.5577	0.5577
married	0.1892	0.5128	-0.8263	.	0.3236	0.3236
nodegree	0.7081	0.5967	0.2450	.	0.1114	0.1114
re74	2095.5737	5619.2365	-0.7211	0.5181	0.2248	0.4470
re75	1532.0553	2466.4844	-0.2903	0.9563	0.1342	0.2876

Summary of Balance for Matched Data:

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
distance	0.5773	0.5765	0.0041	0.9941	0.0042	0.0541	0.0198
age	25.8162	25.7665	0.0069	0.4635	0.0839	0.2736	1.2815
educ	10.3459	10.4240	-0.0388	0.6151	0.0214	0.0643	1.2214
raceblack	0.8432	0.8389	0.0119	.	0.0043	0.0043	0.0162
racehispan	0.0595	0.0499	0.0403	.	0.0095	0.0095	0.4985
racewhite	0.0973	0.1112	-0.0467	.	0.0139	0.0139	0.3911
married	0.1892	0.1688	0.0519	.	0.0203	0.0203	0.4916
nodegree	0.7081	0.6816	0.0583	.	0.0265	0.0265	0.9593
re74	2095.5737	2075.6043	0.0041	1.3589	0.0315	0.2080	0.8492
re75	1532.0553	1592.3461	-0.0187	1.4855	0.0509	0.2122	0.8334

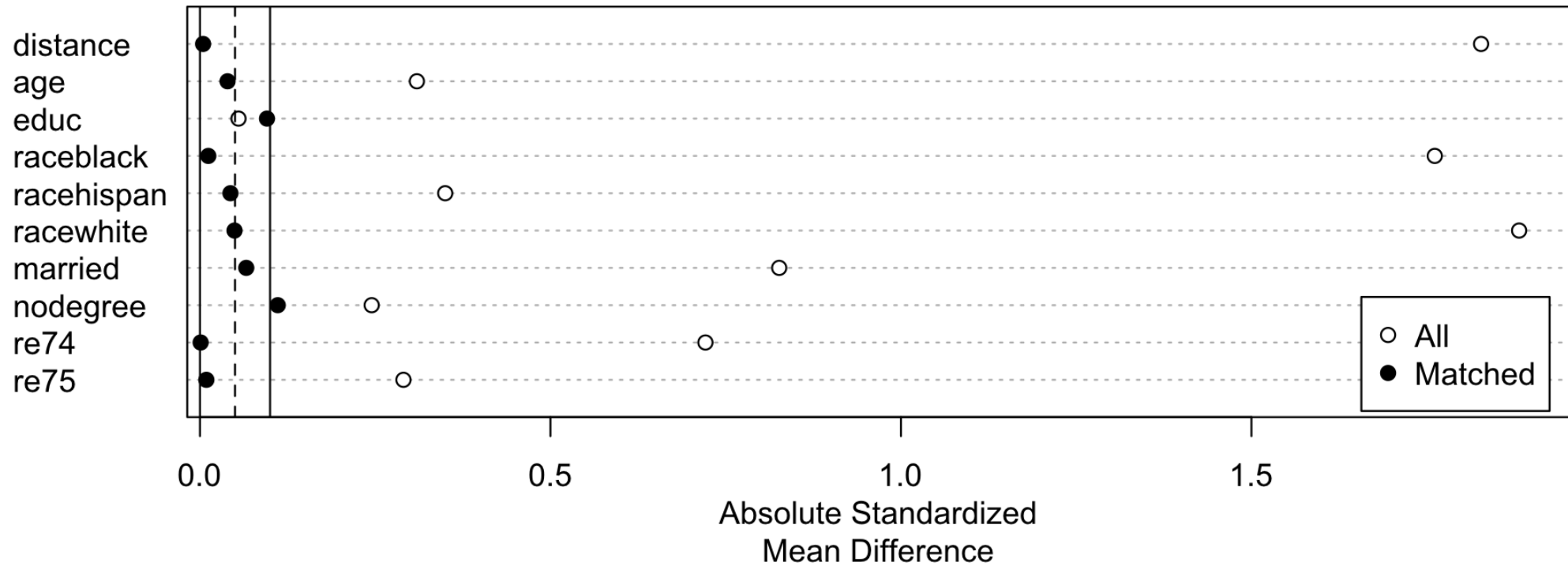
Percent Balance Improvement:

	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max
distance	99.8	95.5	98.9	91.6
age	97.8	6.3	-3.1	-73.5
educ	29.3	30.7	38.3	42.3
raceblack	99.3	.	99.3	99.3
racehispan	88.5	.	88.5	88.5
racewhite	97.5	.	97.5	97.5
married	93.7	.	93.7	93.7
nodegree	76.2	.	76.2	76.2
re74	99.4	53.4	86.0	53.5
re75	93.5	-785.5	62.1	26.2

Sample Sizes:

	Control	Treated
All	429.	185
Matched (ESS)	53.18	185
Matched	429.	185
Unmatched	0.	0

```
plot(summary(out2))
```



```
PSMTable(out2,grouplabel=c("Control","Treated"))
```

Variables	Before Propensity Score Matching				After Propensity Score Matching			
	Control	Treated	p	standardized difference	Control ^a	Treated ^a	p	standardized difference
Covariates	N=429	N=185			N=429	N=185		
age	28.0 ± 10.8	25.8 ± 7.2	0.003	-0.309	25.5 ± 10.2	25.8 ± 7.2	0.734	0.039
educ	10.2 ± 2.9	10.3 ± 2.0	0.585	0.055	10.5 ± 2.5	10.3 ± 2.0	0.360	-0.096
race			< 0.001				0.787	
- black	87 (20.3%)	156 (84.3%)		1.762	359.9 (83.9%)	156.0 (84.3%)		0.012
- hispan	61 (14.2%)	11 (5.9%)		-0.350	21.1 (4.9%)	11.0 (5.9%)		0.043
- white	281 (65.5%)	18 (9.7%)		-1.882	48.0 (11.2%)	18.0 (9.7%)		-0.049
married:1	220 (51.3%)	35 (18.9%)	< 0.001	-0.826	70.1 (16.3%)	35.0 (18.9%)	0.587	0.066
nodegree:1	256 (59.7%)	131 (70.8%)	0.011	0.245	282.1 (65.8%)	131.0 (70.8%)	0.465	0.111
re74	5619.2 ± 6788.8	2095.6 ± 4886.6	< 0.001	-0.721	2100.2 ± 4174.1	2095.6 ± 4886.6	0.990	-0.001
re75	2466.5 ± 3292.0	1532.1 ± 3219.3	0.001	-0.290	1561.4 ± 2530.3	1532.1 ± 3219.3	0.904	-0.009

^aValues are weighted mean ± weighted sd or weighted percentages;

매칭된 데이터

```
matchedData=match.data(out2)
head(matchedData)
```

	treat	age	educ	race	married	nodegree	re74	re75	re78	distance	weights	subclass
1	1	37	11	black	1	1	0	0	9930.0460	0.6356769	1	1
2	1	22	9	hispan	0	1	0	0	3595.8940	0.2298151	1	55
3	1	30	12	black	0	0	0	0	24909.4500	0.6813558	1	63
4	1	27	11	black	0	1	0	0	7506.1460	0.7690590	1	70
5	1	33	8	black	0	1	0	0	289.7899	0.6954138	1	79
6	1	22	9	black	0	1	0	0	4056.4940	0.6943658	1	86

치료효과 추정

Without Covariates Adjustment

```
fit=lm(re78~treat,  
      data=matchedData, weights=weights)
```

With Covariates Adjustment

```
fit=lm(re78~treat+age+educ+race+married+nodegree+re74+re75,  
      data=matchedData, weights=weights)
```

매칭 후 효과 추정

- MatchIt 패키지의 `match.data()` 함수를 이용하여 매칭 데이터 추출
- 효과 추정을 위한 모델링
 - 반응변수가 연속형 변수인지, 이분형 변수인지, 생존변수인지에 따라 `lm()`, `glm()`, `coxph()` 함수를 이용하여 가중치를 사용하여 모형적합시킨다.
- 표준오차 및 신뢰구간 추정
 1. cluster-robust standard error 추정
 2. 부트스트래핑에 의한 추정

1. cluster-robust standard error 추정

- 보통의 OLS 회귀에서 추정하는 표준오차를 보정한 **robust standard errors**를 사용하는 것을 추천한다. 일반적으로 성향점수를 추정할 때 추정치에 균일하지 않은 가중치(non-uniform weights)가 포함되어 있는 경우(예를 들어 전체 매칭이나 IPW를 사용하는 경우)에는 robust standard errors를 반드시 사용하여야 한다.
- 보통의 robust standard error는 효과추정을 할 때 표집에 따른 다양성(sampling variability)을 과대 또는 과소추정할 수 있으므로 클러스터 내 관측치들 간의 의존성을 설명할 수 있는 **cluster-robust standard error**를 사용할 수 있다. 여러 시뮬레이션 연구들에서 성향점수 매칭 후 cluster-robust standard error의 유용성이 입증되었기 때문에 대부분의 경우 cluster-robust standard error의 사용이 권장된다.

2. 부트스트래핑에 의한 추정

- 모의실험을 통하여 표본을 반복 표집하여 표준오차와 신뢰구간을 계산한다.
- 성향점수 추정, 매칭, 효과 추정의 전 과정을 반복하여 추정한다.
- 반복횟수가 많을 수록 좋으나 시간이 많이 걸릴 수 있고 부트스트랩 분석 도중 한번 이상의 에러가 날 수 있으므로 499회, 999회 반복 등이 사용된다.
- 신뢰구간을 계산하는 여러 방법 중 bias-corrected accelerated(BCa) 방법에 의한 신뢰구간이 가장 잘 수행되고 사용하기 쉽다.

1. Cluster-Robust Standard Error 추정

```
coeftest(fit,vcov.=vcovCL,cluster=~subclass)
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.7202e+03	3.2182e+03	0.8452	0.3983081	
treat	1.9857e+03	7.3004e+02	2.7200	0.0067168	**
age	-3.4296e+01	3.8349e+01	-0.8943	0.3715097	
educ	2.2799e+02	2.1570e+02	1.0570	0.2909347	
racehispan	9.6806e+02	1.4589e+03	0.6636	0.5072144	
racewhite	1.4676e+03	8.2932e+02	1.7696	0.0772985	.
married	9.2695e+02	1.1917e+03	0.7778	0.4369796	
nodegree	-1.4583e+03	1.2821e+03	-1.1374	0.2558320	
re74	-7.6333e-03	1.4612e-01	-0.0522	0.9583551	
re75	5.5009e-01	1.4054e-01	3.9140	0.0001011	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
coefci(fit,vcov.=vcovCL,cluster=~subclass)
```

	2.5 %	97.5 %
(Intercept)	-3600.0303069	9040.3431519
treat	551.9611570	3419.4276093
age	-109.6095494	41.0174586
educ	-195.6164929	651.6013100
racehispan	-1896.9811351	3833.1051653
racewhite	-161.1375420	3096.2705400
married	-1413.4815725	3267.3785568
nodegree	-3976.2218485	1059.7055481
re74	-0.2946017	0.2793350
re75	0.2740760	0.8261075

2. Bootstrapping

```
pair_ids <- levels(matchedData$subclass)
est_fun <- function(pairs, i) {

  #Compute number of times each pair is present
  numreps <- table(pairs[i])

  #For each pair p, copy corresponding md row indices numreps[p] times
  ids <- unlist(lapply(pair_ids[pair_ids %in% names(numreps)],
                     function(p) rep(which(matchedData$subclass == p),
                                     numreps[p]))))

  #Subset md with block bootstrapped ids
  md_boot <- matchedData[ids,]

  #Effect estimation
  fit_boot <- lm(re78~treat+age+educ+race+married+nodegree+re74+re75,
                data = md_boot, weights = weights)

  #Return the coefficient on treatment
  return(coef(fit_boot)["treat"])
}
```

```
set.seed(1)
(boot_est <- boot(pair_ids, est_fun, R = 499))
```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = pair_ids, statistic = est_fun, R = 499)

Bootstrap Statistics :

	original	bias	std. error
t1*	1985.694	-74.27486	760.4809

```
boot.ci(boot_est, type = "bca")
```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 499 bootstrap replicates

CALL :
boot.ci(boot.out = boot_est, type = "bca")

Intervals :

Level	BCa
95%	(589, 3467)

Calculations and Intervals on Original Scale
Some BCa intervals may be unstable

webrPSM 패키지를 이용한 성향점수 분석

```
library(webrPSM)  
reportPSM(out2)
```

We used propensity score matching to estimate the average marginal effect of the 'treat' on those who received it accounting for confounding by the included covariates. We first attempted 1:1 nearest neighbor propensity score matching without replacement with a propensity score estimated using logistic regression of the treatment on the covariates. But this matching yielded poor balance, so we tried full matching on the propensity score, which yielded adequate balance, as indicated in Table and Figure. After matching, all standardized mean differences for the covariates were below 0.1 and all standardized mean differences for squares and two-way interactions between covariates were below 0.15, indicating adequate balance. The propensity score was estimated using a probit regression of the 'treat' on the covariates, which yielded better balance than did a logistic regression. Full matching uses all treated and all control units, so no units were discarded by the matching.

webrPSM 패키지를 이용한 치료효과 추정

```
estimateEffect(out2,dep="re78",multiple=TRUE)
```

Estimate Effect

	Estimate	Std. Error	t value	Pr(> t)	2.5 %	97.5 %
re78	1985.694	730.0442	2.719965	0.006716776	551.9612	3419.428

Interpretation

To estimate the 'treat' effect and its standard error, we fit a linear regression model with re78 as the outcome and the 'treat' and the covariates age, educ, race, married, nodegree, re74 and re75 as additive predictors and included the matching weights in the estimation. The coefficient on the 'treat' was taken to be the estimate of the 'treat' effect. The `lm()` function was used to estimate the effect, and a cluster-robust variance as implemented in the `vcovCL()` function in the sandwich package was used to estimate its standard error with matching stratum membership as the clustering variable. For the outcome 're78', the estimated effect was 1985.69 (SE = 730.04, p = 0.0067), indicating that the average effect of the 'treat' for those who received it is to increase 're78'.

```
estimateEffect2(out2,dep="re78",multiple=TRUE)
```

```
$est
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP
```

```
Call:
```

```
boot(data = pair_ids, statistic = est_fun, R = 499, multiple = multiple,  
      mode = mode)
```

```
Bootstrap Statistics :
```

	original	bias	std. error
t1*	1985.694	-74.27486	760.4809

```
$ci
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
```

```
Based on 499 bootstrap replicates
```

```
CALL :
```

```
boot.ci(boot.out = boot_est, type = "bca")
```

```
Intervals :
```

Level	BCa
95%	(589, 3467)

```
Calculations and Intervals on Original Scale
```

```
Some BCa intervals may be unstable
```

샤이니 앱을 이용한 ps matching

<https://cardiomoon.shinyapps.io/psmatching>

Web-R.org

웹에서 하는 R 통계분석 PSM 4.3.4

자신의 컴퓨터에 R을 설치할 필요 없이 R을 이용한 통계분석을 할 수 있습니다. 그룹변수와 행변수를 선택하여 쉽게 표를 만들 수 있으며 그래프를 통한 자료 탐색과 여러가지 통계분석이 가능합니다. **자신의 데이터**를 xls or csv 형식으로 업로드하여 분석을 할 수 있을 뿐 아니라 그 결과를 pdf, docx, powerpoint 파일로 다운로드할 수 있습니다. 또한 Plot을 원하는 크기로 저장할 수 있습니다. **표기 보일 때까지 잠시만** 기다려주세요.

Select Language
☐ English ☒ 한국어(Korean)

Web-R.org DataSelect dataWrangling Propensity Score PPTxList Citation

샘플 데이터를 선택 하시거나 자료를 업로드 하세요. 현재 지원하는 파일 형식은 csv, xls (Microsoft Excel), dbf (dbase 3+), sav (SPSS), dta (STATA), sas7bdat (SAS) 파일 등이나 예러가 있을 경우 csv 형식으로 바꾸어 업로드할 것을 권장합니다.

파일 업로드

Browse... No file selected

데이터선택

☒ lalonde

☐ exData

☐ swData

☐ simData

☐ simData2

☐ AOD

☐ dat

☐ colon

☐ GBSQ2

☐ 데이터 도움말 보기

☐ 전처리하기 전처리 초기화

데이터 전처리하기

여기에 R명령어를 입력하면 데이터를 전처리할 수 있습니다. R 명령어를 입력/수정하기 전에 체크박스를 해제하고 입력후 다시 체크박스를 선택하시기 바랍니다.

데이터이름

lalonde

Data files in Server

treat age educ race married nodegree re74 re75 re78

참고문헌

1. MatchIt 패키지 vignette

<https://cran.r-project.org/web/packages/MatchIt/index.html>

2. 백영민, 박인서(2021). R 기반 성향점수분석: 루빈 인과모형 기반 인과추론(한나래)

