Propensity Score-Matching Methods for Nonexperimental Causal Studies

202STG18 이재빈

CONTENTS

1. 논문 리뷰

1) 배경

2) 방법론

2. 코드 리뷰

1) Data

2) Code

3. 결론

1. 논문 리뷰

배경

✓ RH Dehija 저술 · 2002 · 5484회 인용

Treatment 에 따른
Outcome 의 효과
이를 잘 설명해주는
Subset 찾는 것 중요

PROPENSITY SCORE-MATCHING METHODS FOR NONEXPERIMENTAL CAUSAL STUDIES

Rajeev H. Dehejia and Sadek Wahba*

Abstract—This paper considers causal inference and sample selection bias in nonexperimental settings in which (i) few units in the nonexperimental comparison group are comparable to the treatment units, and (ii) selecting a subset of comparison units similar to the treatment units is difficult because units must be compared across a high-dimensional set of pre-treatment characteristics. We discuss the use of propensity score-matching methods, and implement them using data from the National Supported Work experiment. Following LaLonde (1986), we pair the experimental treated units with nonexperimental comparison units from the CPS and PSID, and compare the estimates of the treatment effect obtained using our methods to the benchmark results from the experiment. For both comparison groups, we show that the methods succeed in focusing attention on the small subset of the comparison units comparable to the treated units and, hence, in alleviating the bias due to systematic differences between the treated and comparison units.

I. Introduction

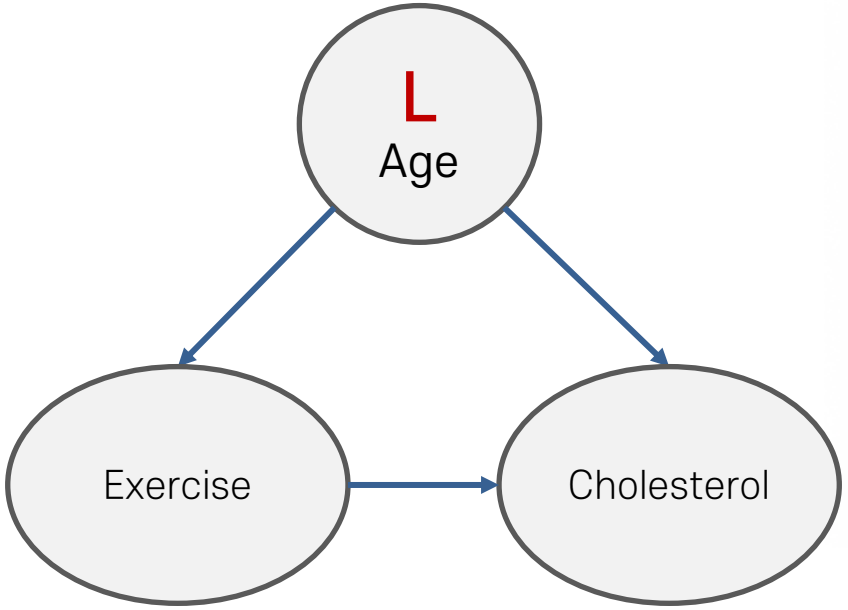
An important problem of causal inference is how to estimate treatment effects in observational studies, situations (like an experiment) in which a group of units is exposed to a well-defined treatment, but (unlike an experiment) no systematic methods of experimental design are used to maintain a control group. It is well recognized that the estimate of a causal effect obtained by comparing a treatment impact.[1] The first generation of matching methods paired observations based on either a single variable or weighting several variables. (See, inter alia, Bassi (1984), Cave and Bos (1995), Czajka et al. (1992), Cochran and Rubin (1973), Raynor (1983), Rosenbaum (1995), Rubin (1973, 1979), Westat (1981), and studies cited by Barnow (1987).)

The motivation for focusing on propensity score-matching methods is that, in many applications of interest, the dimensionality of the observable characteristics is high. With a small number of characteristics (for example, two binary variables), matching is straightforward (one would group units in four cells). However, when there are many variables, it is difficult to determine along which dimensions to match units or which weighting scheme to adopt. Propensity score-matching methods, as we demonstrate, are especially useful under such circumstances because they provide a natural weighting scheme that yields unbiased estimates of the treatment impact.

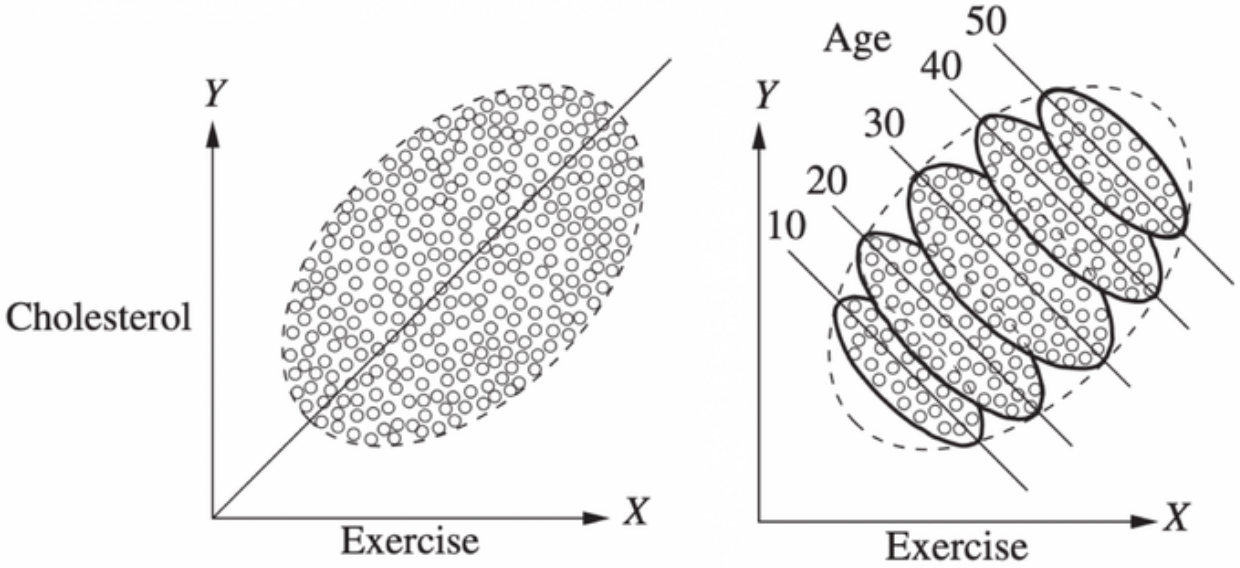
The key contribution of this paper is to discuss and apply propensity score-matching methods, which are new to the economics literature. (Previous papers include Dehejia and

배경

✓ Simpson's Paradox



Backdoor Path



Confounder 를 잘 찾아서, 통제하는 것 중요

배경

✓ Selection Bias

Observational Study (Nonexperimental Research)

Researcher 가 Causal Relationship 에 대한 구체적인 질문 할 수 없고,

Independent Variable 을 조작할 수 없음

Maintain a Control Group could be biased!

1. Self-Selection
2. Systematic Judgement by Researcher

배경

✓ Matching 가장 유사한 조건의 사람을 찾아서, 비교

***Matching**, def: any method that **strategically subsamples** dataset to **balance covariate distribution** in treated and control groups such that after matching both groups share an equal probability of treatment.*

Non-Random Treatment Assignment $\xrightarrow[\text{to Subsample}]{\text{Matching Methods}}$ ***Average Treatment Effect on the Treated + ~~Selection Bias~~***

방법론

✓ Matching Methods

A. The Role of Randomization

- The Problem of Unobservability

  We can estimate $E(Y_{i1}|T_i = 1)$, but not $E(Y_{i0}|T_i = 1)$

- We can estimate ATE, but it is biased estimator!

- Randomization can ensure Treatment & Control groups do not

  Systematically differ + Remove Selection Bias (Independence & Exchangeability)

- But it is Unethical & Costly

방법론

✓ Matching Methods

B. Exact Matching of Covariates

1. Unobserved data 의 Treatment 효과 보기 어려움
2. Comparison Group 충분히 확보 되어야, Exact Match 찾을 수 있음
3. 모두 Matching 해 보아야 해서 경우의 수 급증

방법론

✓ Matching Methods

C. Propensity Score & Dimensionality Reduction

- The **Probability** of receiving **Treatment** conditional on **Covariates**

Proposition 2 (Rosenbaum and Rubin, 1983). Let $p(X_i)$ be the probability of a unit i having been assigned to treatment, defined as $p(X_i) \equiv \Pr(T_i = 1|X_i) = E(T_i|X_i)$. Then,

$$(Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i | X_i \quad \text{f} \quad (Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i | p(X_i).$$

Proposition 3. $\tau|_{T=1} = E_{p(X)}[(\tau|_{T=1,p(X)})|T_i = 1]$.

$$\hat{\tau}|_{T=1} = \frac{1}{|N|} \sum_{i \in N} \left(Y_i - \frac{1}{|J_i|} \sum_{j \in J_i} Y_j \right),$$

N : treatment group , J : comparison units

방법론

✓ Propensity Score

Covariate 를 조건으로 했을 때, 처리 받을 **확률** 추정

ATT = E[Y(1)-Y(0)|X, T=1]

$$\hat{e}(X) = P[\,T = 1|X]$$

Logistic, GAM / SVM, TreeModels, ⋯

Step 1 of 2:

A. Define closeness.
B. Match.
C. Evaluate resulting **covariate balance**, perhaps repeat iteratively.

Step 2 of 2:

D. Estimate Treatment Effect

Stuart, E. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Science, Vol. 25, No. 1, 1–21. DOI: 10.1214/09-STS313. Page 4.

ISSUE

1. Matching w/wo Replacement
2. Set Comparison Units
3. Matching Methods

방법론

✓ Propensity Score

1. Matching with / without Replacement
    1. With Replacement
        - Propensity Score 를 Minimize 하는 방향으로!
        - Bias Reduction
        - Single-nearest-neighborhood matching, Caliper matching, …
    2. Without Replacement
        - Few comparison units are similar to the treated units
        - Increase bias, but Improve Precision
        - High to low, Low to high, Random

방법론

✓ Propensity Score

2. Set Comparison Units
    1. Number of Comparison Units
        - Single : Smallest Propensity-Score distance
        - More : Improve Precision, but bias also increased

방법론

✓ Propensity Score

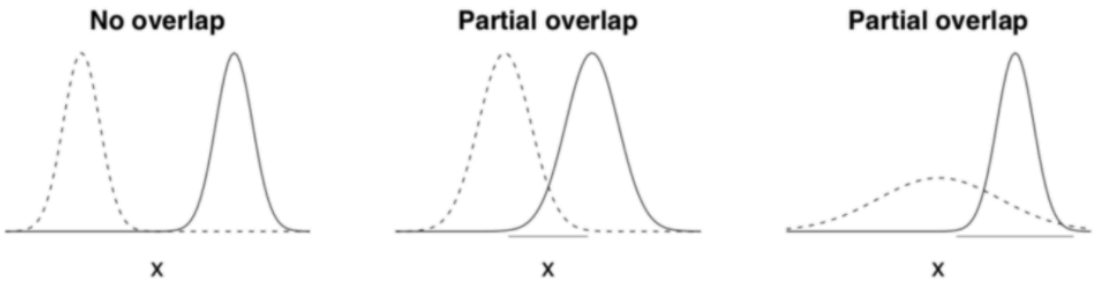
2. Set Comparison Units
   2. Matching Methods
      - Mahalanobis Matching : $D_{ij} = \sqrt{(x_i - y_j)^T S^{-1} (x_i - y_j)}$
      - Nearest-neighbor Matching
        - PS 절댓값 작은 순서대로 1:1 매칭 후, 다 짝지을 때 까지 반복
        - 1:k matching 을 진행할 경우 더 안정된 값을 얻을 수 있음
      - Caliper matching Matching
        - score radius 범위 안으로 제한 (.0001 ~ .01)
      - Kernel Matching : PS 가 가까울수록 큰 가중치 부여

방법론

✓ Propensity Score

3. Select Matching Methods

- Matching 방법은 Data 의 특성에 따라 선택
- Treatment Group / Comparison Unit 의 PS distribution 이 얼마나 overlap 되는지에 따라 선택



두 분포가 다르면, without replacement 매우 문제가 됨

2. 코드 리뷰

Data

✓ The National Supported Work Program (NSW)

	treat	age	educ	race	married	nodegree	re74	re75	re78
NSW176	1	28	11	1	0	1	1929.0290	6871.85600	0.0000
NSW177	1	22	12	0	1	0	492.2305	7055.70200	10092.8300
NSW178	1	33	11	1	0	1	0.0000	7867.91600	6281.4330
NSW179	1	22	12	0	0	0	6759.9940	8455.50400	12590.7100
NSW180	1	29	10	0	0	1	0.0000	8853.67400	5112.0140
NSW181	1	33	12	1	1	0	20279.9500	10941.35000	15952.6000
NSW182	1	25	14	1	1	0	35040.0700	11536.57000	36646.9500
NSW183	1	35	9	1	1	1	13602.4300	13830.64000	12803.9700
NSW184	1	35	8	1	1	1	13732.0700	17976.15000	3786.6280
NSW185	1	33	11	1	1	1	14660.7100	25142.24000	4181.9420
PSID1	0	30	12	0	1	0	20166.7300	18347.23000	25564.6700
PSID2	0	26	12	0	1	0	25862.3200	17806.55000	25564.6700
PSID3	0	25	16	0	1	0	25862.3200	15316.21000	25564.6700
PSID4	0	42	11	0	1	1	21787.0500	14265.29000	15491.0100
PSID5	0	25	9	1	1	1	14829.6900	13776.53000	0.0000
PSID6	0	37	9	1	1	1	13685.4800	12756.05000	17833.2000
PSID7	0	32	12	0	1	0	19067.5800	12625.35000	14146.2800
PSID8	0	20	12	1	0	0	7392.3140	12396.19000	17765.2300

(619, 9)

Treat : (0) 429 , (1) 185

Confounding Covariates

1. Age
2. Education
3. Race (Black / Hispanic)
4. Married
5. No Degree
6. Income 1974
7. Income 1975

	treat	n_people	mean_outcome	std_error
	<int>	<int>	<dbl>	<dbl>
1	0	429	6984.	352.
2	1	185	6349.	578.

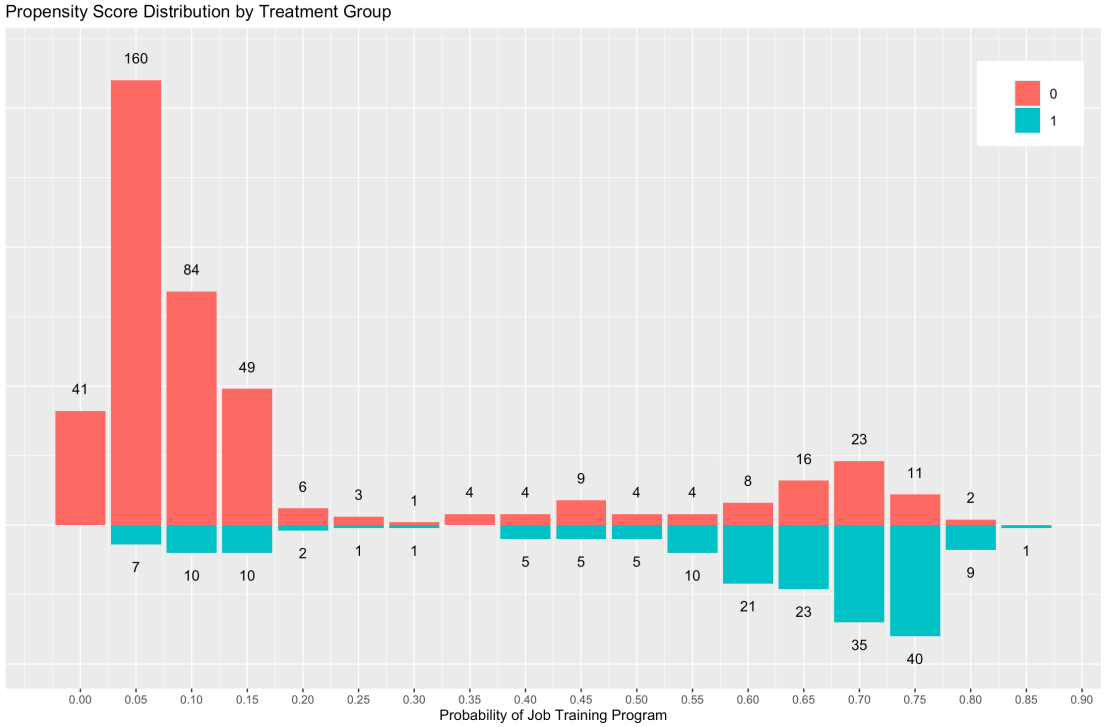
Job training 을 받은 사람들의 평균 임금이 $635 적다?

Code

1. Propensity Score Estimation , by Logistic Regression

treat 반응변수로 두고, Logistic Regression

```
Call:
glm(formula = treat ~ age + educ + race + married + nodegree +
    re74 + re75, family = binomial(), data = lalonde)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7774  -0.5044  -0.3120   0.7529   2.6484

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.266e+00  9.909e-01  -4.306 1.67e-05 ***
age          1.386e-02  1.336e-02   1.037  0.29953
educ         1.475e-01  6.501e-02   2.269  0.02326 *
race         2.786e+00  2.423e-01  11.499  < 2e-16 ***
married     -8.678e-01  2.892e-01  -3.001  0.00269 **
nodegree     7.193e-01  3.376e-01   2.130  0.03313 *
re74        -7.309e-05  2.869e-05  -2.547  0.01085 *
re75         6.033e-05  4.661e-05   1.294  0.19556
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 751.49  on 613  degrees of freedom
Residual deviance: 492.83  on 606  degrees of freedom
AIC: 508.83

Number of Fisher Scoring iterations: 5
```



Code

2. Executing a Matching Algorithm

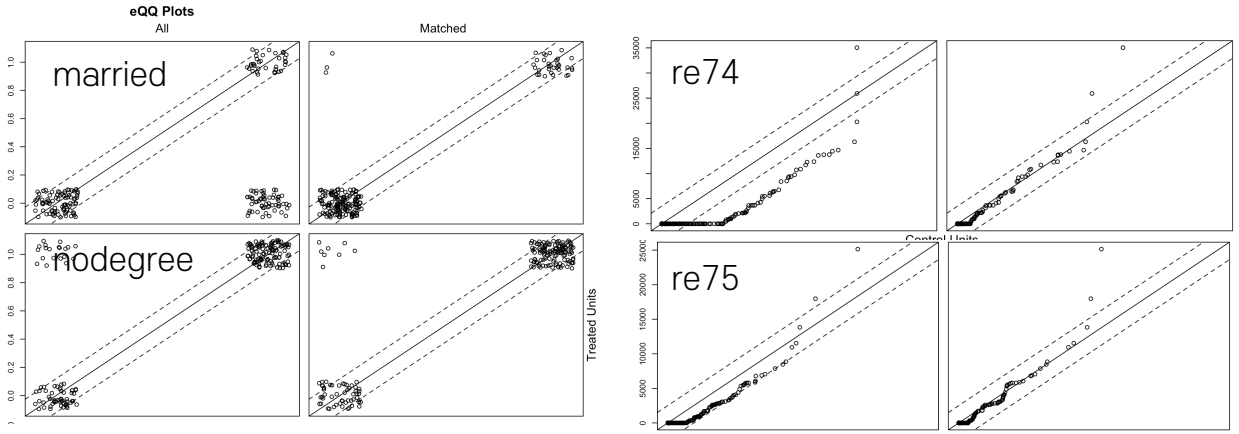
```r
m.out1 <- matchit(treat ~ age + educ + race + married
                  + nodegree + re74 + re75, data = lalonde,
                  distance = "glm", link = "linear.probit")
```

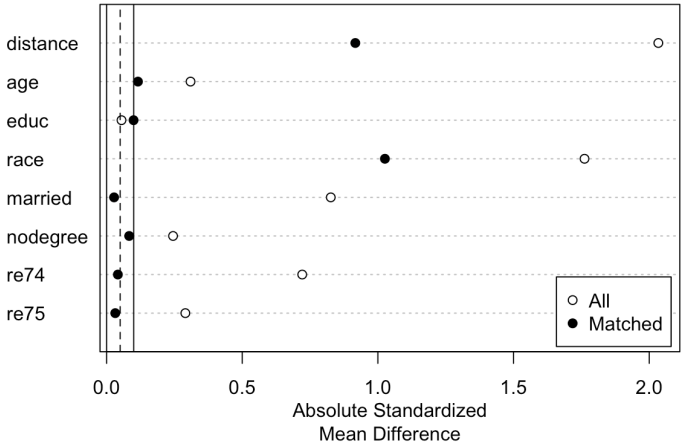
```
A matchit object
 - method: 1:1 nearest neighbor matching without replacement
 - distance: Propensity score
             - estimated with probit regression and linearized
 - number of obs.: 614 (original), 370 (matched)
 - target estimand: ATT
 - covariates: age, educ, race, married, nodegree, re74,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

Sample Sizes:

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

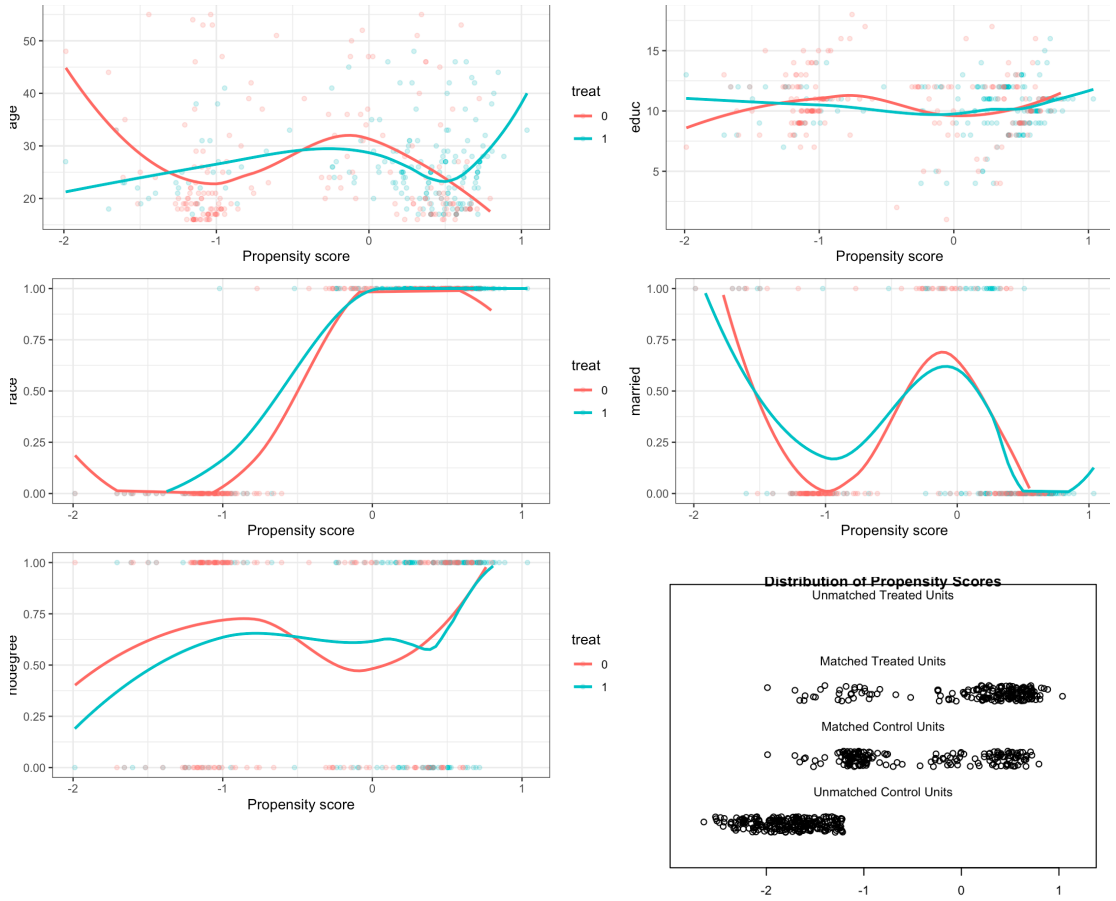
	Means Treated	Means Control
distance	0.16	-1.19
age	25.82	28.03
educ	10.35	10.24
race	0.84	0.20
married	0.19	0.51
nodegree	0.71	0.60
re74	2095.57	5619.24
re75	1532.06	2466.48





Code

3. Examining Covariate Balance in the Matched Samples



Code

4. Estimates Treatment Effects , by Linear Regression

Matching 후, Treatment 효과가 유의미하게 나타나는 양상을 확인할 수 있음

```
> with(dta.m1, t.test(re78 ~ treat))

        Welch Two Sample t-test

data:  re78 by treat
t = -1.6024, df = 338.6, p-value = 0.11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2566.0678   262.1413
sample estimates:
mean in group 0 mean in group 1
       5197.180        6349.144

> lm_treat1 <- lm(re78 ~ treat, data = dta.m1)
> summary(lm_treat1)

Call:
lm(formula = re78 ~ treat, data = dta.m1)

Residuals:
   Min     1Q Median     3Q    Max
 -6349  -5197  -2128   3006  53959

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5197.2      508.4  10.224   <2e-16 ***
treat         1152.0      718.9   1.602     0.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6914 on 368 degrees of freedom
Multiple R-squared:  0.006929,  Adjusted R-squared:  0.00423
F-statistic: 2.568 on 1 and 368 DF,  p-value: 0.1099
```

```
  treat   age  educ  race married nodegree  re74  re75
  <int> <dbl> <dbl> <dbl>   <dbl>    <dbl> <dbl> <dbl>
1     0  25.0  10.5 0.470   0.178    0.670 2299. 1635.
2     1  25.8  10.3 0.843   0.189    0.708 2096. 1532.
> lm_treat2 <- lm(re78 ~ treat + age + educ + race + married + nodegree +
+                 re74 + re75, data = dta.m1)
> summary(lm_treat2)

Call:
lm(formula = re78 ~ treat + age + educ + race + married + nodegree +
    re74 + re75, data = dta.m1)

Residuals:
   Min     1Q Median     3Q    Max
 -9620  -4907  -1607   3352  53586

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.928e+03  3.369e+03  -0.572   0.5674
treat        1.494e+03  7.742e+02   1.930   0.0544 .
age          8.071e+00  4.184e+01   0.193   0.8472
educ         5.869e+02  2.241e+02   2.619   0.0092 **
race        -6.400e+02  8.352e+02  -0.766   0.4440
married      5.134e+02  1.023e+03   0.502   0.6160
nodegree     7.569e+02  1.099e+03   0.689   0.4916
re74        -7.954e-03  1.012e-01  -0.079   0.9374
re75         2.785e-01  1.568e-01   1.776   0.0766 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6814 on 361 degrees of freedom
Multiple R-squared:  0.05392,   Adjusted R-squared:  0.03295
F-statistic: 2.572 on 8 and 361 DF,  p-value: 0.0097
```

Code

✓ Other Examples

1. Caliper

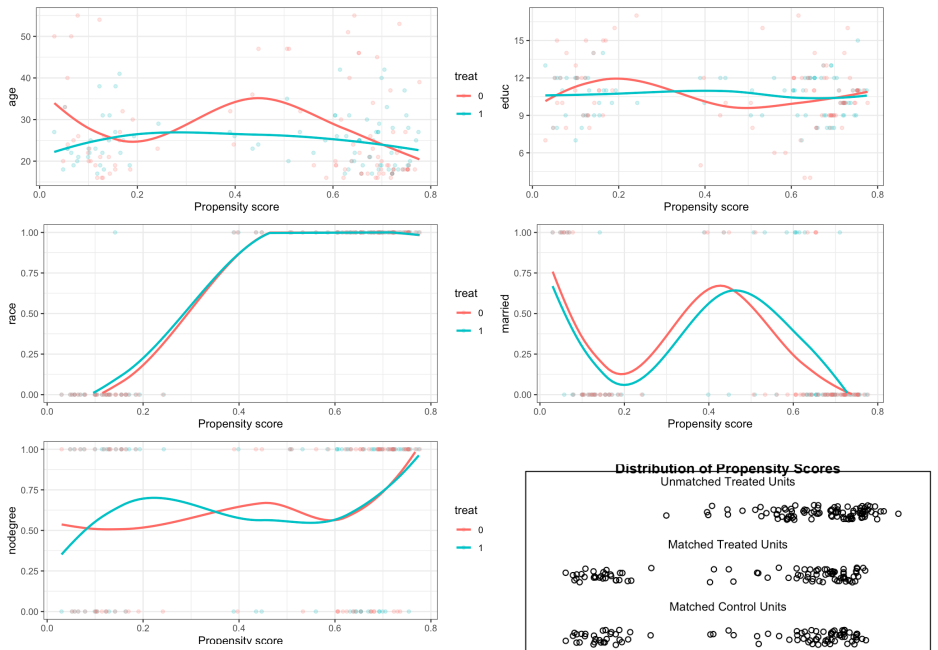
```
m.out2 <- matchit(treat ~ age + educ + race + married
                  + nodegree + re74 + re75, data = lalonde,
                  distance = "glm", caliper = .01)
```

	Means Treated	Means Control
distance	0.58	0.18
age	25.82	28.03
educ	10.35	10.24
race	0.84	0.20
married	0.19	0.51
nodegree	0.71	0.60
re74	2095.57	5619.24
re75	1532.06	2466.48

Sample Sizes:

	Control	Treated
All	429	185
Matched	90	90
Unmatched	339	95
Discarded	0	0

```
> with(dta.m2, t.test(re78 ~ treat))

        Welch Two Sample t-test

data:  re78 by treat
t = -2.2337, df = 174.84, p-value = 0.02677
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4035.9187  -249.4852
sample estimates:
mean in group 0 mean in group 1
       5000.955        7143.657
```



	treat	age	educ	race	married	nodegree	re74	re75
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	26.1	10.7	0.667	0.189	0.644	2297.	1453.
2	1	24.5	10.5	0.678	0.2	0.667	2323.	1682.

Code

✓ Other Examples

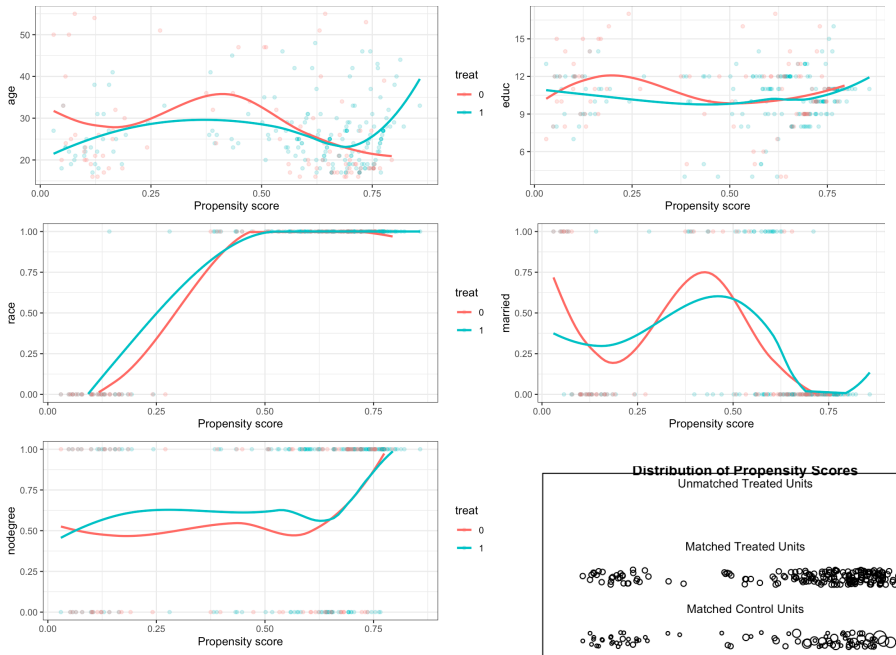
2. With Replacement

```
m.out3 <- matchit(treat ~ age + educ + race + married
                  + nodegree + re74 + re75, data = lalonde,
                  distance = "glm", replace = TRUE)
```

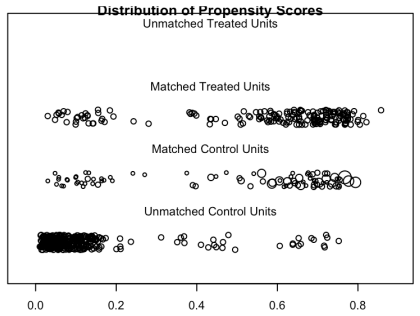
```
         Means Treated Means Control
distance          0.58          0.18
age              25.82         28.03
educ             10.35         10.24
race              0.84          0.20
married           0.19          0.51
nodegree          0.71          0.60
re74           2095.57       5619.24
re75           1532.06       2466.48
```

```
Sample Sizes:
               Control Treated
All             429.      185
Matched (ESS)    48.55    185
Matched          88.      185
Unmatched       341.        0
Discarded         0.        0
```

```
> with(dta.m3, t.test(re78 ~ treat))

        Welch Two Sample t-test

data:  re78 by treat
t = -1.1154, df = 212.31, p-value = 0.2659
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2711.674   751.858
sample estimates:
mean in group 0 mean in group 1
       5369.236        6349.144
```




```
  treat   age  educ  race married nodegree  re74  re75
  <int> <dbl> <dbl> <dbl>   <dbl>    <dbl> <dbl> <dbl>
1     0  26.1  10.8 0.670   0.205    0.591 2508. 1755.
2     1  25.8  10.3 0.843   0.189    0.708 2096. 1532.
```



Code

✓ Other Examples

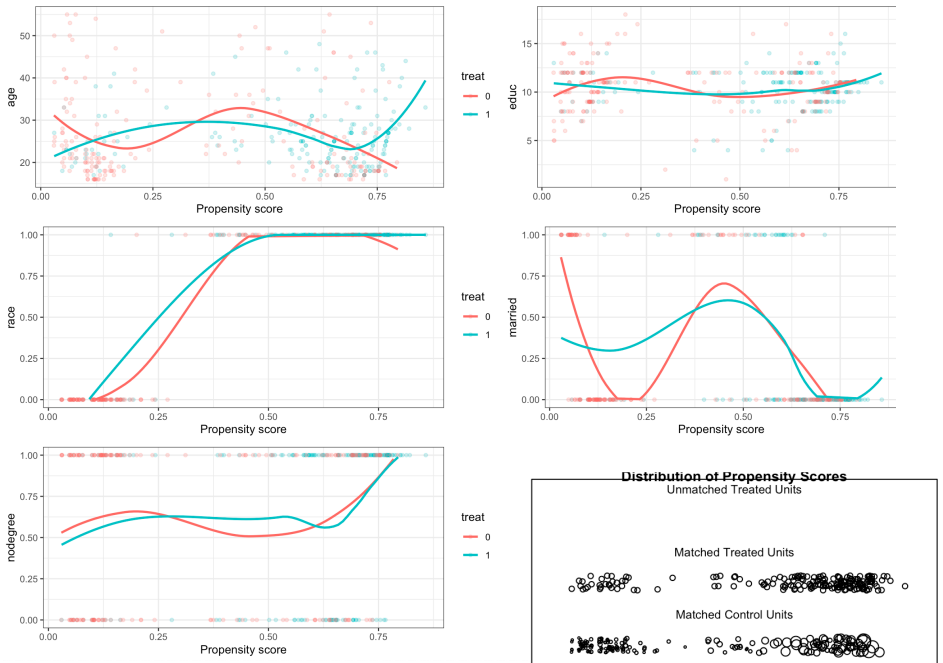
3. 5:1 matching with Replacement

```
m.out4 <- matchit(treat ~ age + educ + race + married
                  + nodegree + re74 + re75, data = lalonde,
                  distance = "glm", replace = TRUE, ratio=5)
```

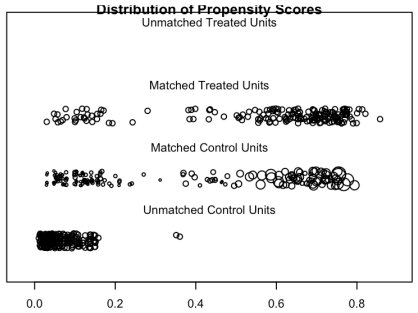
```
         Means Treated Means Control
distance          0.58          0.18
age              25.82         28.03
educ             10.35         10.24
race              0.84          0.20
married           0.19          0.51
nodegree          0.71          0.60
re74           2095.57       5619.24
re75           1532.06       2466.48
```

```
Sample Sizes:
               Control Treated
All             429.      185
Matched (ESS)    78.85    185
Matched         188.      185
Unmatched       241.        0
Discarded         0.        0
```

```
> with(dta.m4, t.test(re78 ~ treat))

        Welch Two Sample t-test

data:  re78 by treat
t = -1.0725, df = 346.24, p-value = 0.2842
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2215.510   651.924
sample estimates:
mean in group 0 mean in group 1
       5567.351        6349.144
```




```
treat   age  educ  race married nodegree  re74  re75
<int> <dbl> <dbl> <dbl>   <dbl>    <dbl> <dbl> <dbl>
    0  26.1  10.5 0.452   0.277    0.633 3055. 1877.
    1  25.8  10.3 0.843   0.189    0.708 2096. 1532.
```



Code

✓ Other Examples

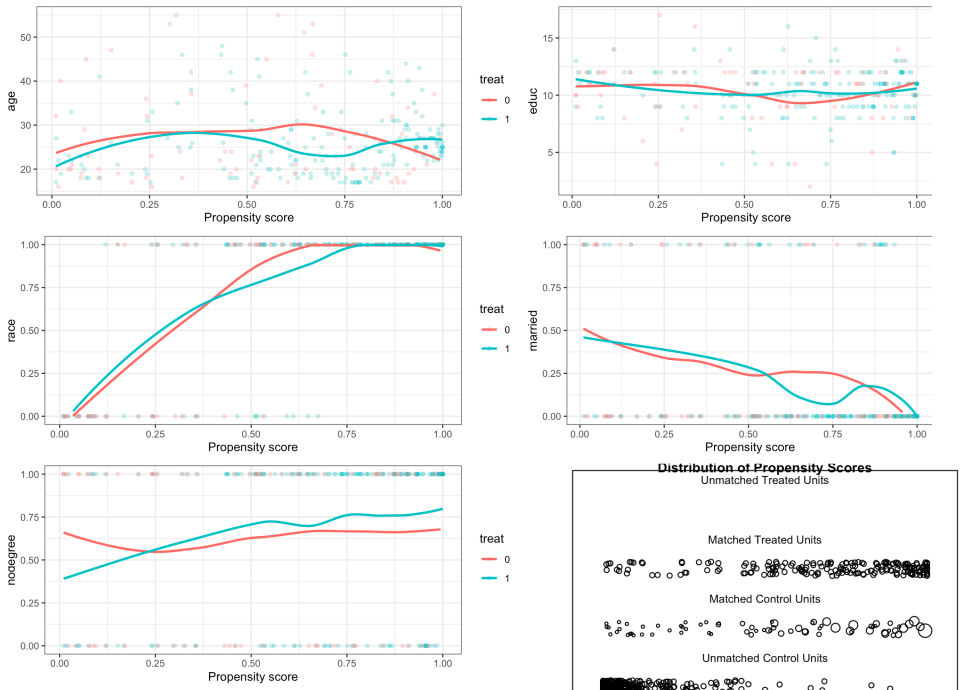
4. Random Forest

```
m.out5 <- matchit(treat ~ age + educ + race + married
                  + nodegree + re74 + re75, data = lalonde,
                  distance = "randomforest", replace = TRUE)
```

```
         Means Treated Means Control
distance          0.69          0.16
age              25.82         28.03
educ             10.35         10.24
race              0.84          0.20
married           0.19          0.51
nodegree          0.71          0.60
re74           2095.57       5619.24
re75           1532.06       2466.48
```

```
Sample Sizes:
               Control Treated
All             429.      185
Matched (ESS)    33.92    185
Matched          72.      185
Unmatched       357.        0
Discarded         0.        0
```

```
> with(dta.m5, t.test(re78 ~ treat))

        Welch Two Sample t-test

data:  re78 by treat
t = -0.68227, df = 157.51, p-value = 0.4961
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2530.627  1231.194
sample estimates:
mean in group 0 mean in group 1
       5699.427        6349.144
```



	treat	age	educ	race	married	nodegree	re74	re75
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	27.5	10.3	0.647	0.309	0.603	2599.	1740.
2	1	25.8	10.3	0.843	0.189	0.708	2096.	1532.

3. 결론

결론

✓ Advantages

① Easy Implementation

- Simple Modeling & Many Available R packages

② Remove Selection Bias

③ Solve Matching Problems for High Dimensions

✓ Disadvantages

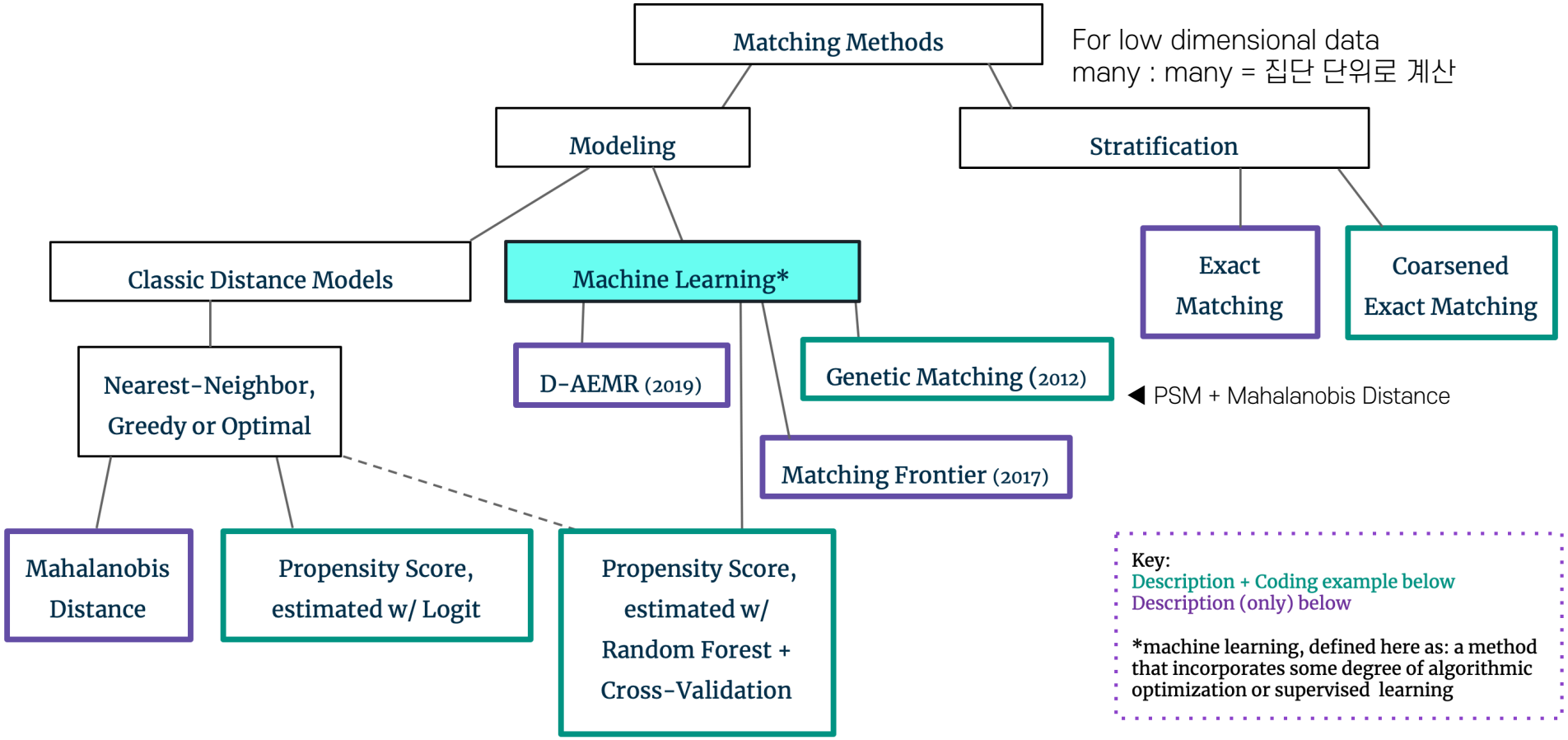
① Bad Models = Bad Matches

② Discard Samples

③ Matched Pairs maybe Dissimilar across Covariates

결론

✓ The Matching Family Trees



Reference

PAPERS

- Propensity Score-Matching Methods for Nonexperimental Causal Studies (Rajeev H., 2002)
- Matching Methods for Causal Inference: A review and a look forward (EA Stuart, 2010)
- Propensity Score Matching Method의 소개 (이동규, 2016)

POSTS

- MATCHING METHODS FOR CAUSAL INFERENCE: A MACHINE LEARNING UPDATE
- https://causalinference.gitlab.io/kdd-tutorial/methods.html#matching-and-stratification
- 인과 관계 분석 시리즈 (1): 성향 점수 매칭 (Propensity Score Matching)과 층화 (Stratification)
- Causal Inference : Primer (2019. 잔디콘)
- R Tutorial 8: Propensity Score Matching

감사합니다