Final Project report

과목:	최신통계적방법
담당교수:	**이권상 교수님**
학과:	**통계학과**
학번:	**2020711995, 2020711985**
이름:	**하성연, 박소담**

1. Introduction

Final project를 위해 선정한 논문은 'propensity score-matching methods for nonexperimental causal studies'로 propensity score-matching 기법을 비실험적인 데이터, 즉 관측 데이터에 적용해 인과관계를 추론하는 일련의 과정을 다루고 있다. Final project를 준비한 과정은 다음과 같다. 우선 논문을 읽고 요약을 통해 리뷰를 할 내용을 정리하였다. 그리고 해당 논문에서 적용된 기법을 실제 데이터에 적용하기 위해 필요한 데이터를 수집하였다. 그 후 연구설계 및 분석을 하였고 그에 대한 결과와 한계점에 대해 파악할 수 있었다.

2. Paper review

해당 논문은 NSW라는 일자리 프로그램을 treatment로 하여 NSW의 효과를 보고자 했다. 그에 따라 treatment 집단은 NSW에 참여한 사람들이며 control 집단은 CPS / PSID의 두 집단이다. 또한 처리 효과, 즉 일자리 프로그램의 효과를 알아보기 위해 3가지 매칭 기법을 사용했다. 3가지 방법에 따라 매칭된 control 집단과 treatment 집단간의 propensity score의 유사성이 다르다는 것을 확인할 수 있었다. 그 중 nearest neighbor 매칭이 두 집단간의 propensity score가 가장 유사했다. 이를 통해 매칭 알고리즘의 선택도 상당히 중요하다는 것을 알 수 있었다.

사실상 treatment 집단이 control 집단과 나이, 인종, 교육수준 등의 측면에서 상당히 다르다. 하지만 해당 논문은 propensity score 매칭이 실험 데이터가 아닌 관측 데이터를 통해 정확한 처리효과의 추정치를 산출했다는 점에 의의가 있다. 덧붙여 다른 연구에서 propensity score 매칭을 사용할 때 관측할 수 있는 covariate에 의해 treatment 집단이 control 집단과 매칭이 되는지, 아니면 매칭 프로세스가 관측할 수 없는 변수에 의존하는 것이 아닌지 살펴보는 것도 중요함을 언급하고 있다.

3. Data & Covariate

연구의 주제는 흡연과 폐질환과 인과관계가 있는지 그리고 있다면 처리효과는 어떻게 되는지를 알아보고자 하였다. 데이터는 국민건강보험공단의 건강검진 코호트 DB를 사용했다. 해당 데이터는 개인의 기본적인 정보와 관련된 테이블, 의료 기록에 대한 테이블, 그리고 검진결과에 대한 테이블로 총 3개의 테이블로 구성되어 있다.

각 테이블은 person_id라는 개인고유식별번호를 통해 연결이 가능하다. 구체적으로 각 테이블에서 사용한 covariate에 대해 알아보자. 개인정보 테이블에서는 처음에 성별을 covariate로 간주하

였다. 하지만 추후 매칭 결과, 흡연율이 여성에 비해 남성이 현저히 높으며 여성의 경우 treatment 집단, 즉 흡연을 한 여성 자체가 정의되지 않아 매칭이 불가능한 경우도 있었다. 그래서 남성 집단으로 연구 대상을 좁히게 되었다. 또한 연령이 높을수록 질병에 걸릴 확률이 높아지기에 나이 변수를 선택하였고, 어느 지역에 거주하는지, 그리고 소득수준에 따라 의료의 질에 차이가 있기 때문에 소득분위를 변수로 선정하였다.

다음으로 의료 기록 테이블에서는 main_sick과 sub_sick, 즉 주상병과 부상병을 사용했으며, 이 변수들을 사용해 암 또는 폐질환에 대한 과거력이 있는 대상을 제외하고 treatment의 결과인 폐질환의 유무를 보고자 하였다.

마지막으로 검진 테이블에서는 검진년도, 즉 2009~2013년까지의 검진결과를 활용하였고, 가족력, 과거력을 사용하였다. 또한 treatment 집단과 control 집단을 정의하기 위해 흡연 상태 변수를 활용하였다. current smoker를 treatment 집단으로 정의하고 그 이외의 none smoker와 ex_smoker를 control 집단으로 고려하였다. 그리고 추가로 운동의 빈도도 covariate로 선정하였다.

4. Study design

실제로 실험을 할 수 없었기에 2009년 ~ 2013년의 기록 중 2009년을 기준년도로 설정하고 해당 시점에서 current smoker인 사람을 treatment 그룹으로 선정하였다. 그리고 해당 시점으로부터 진료 테이블에서 treatment 집단과 control 집단의 5년 후 주상병 또는 부상병에서의 폐질환 유무를 통해 outcome을 확인하였다.

5. Matching

Matching은. Without replacement, With replacement, ranked-base Mahalanobis distance – caliper 1000 matching, full matching, 총 4가지의 방법을 사용하였다. 논문에서는 총 4가지의 방법이 나오지만, 논문에서 without replacement와 with replacement 차이에 대해 명시하였으며, 수업에서 사용한 방법도 적용하고 싶어서 ranked-base Mahalanobis distance – caliper 1000 matching과 full matching을 추가적으로 진행해보았다.

Matching을 진행하기 전 지역을 effect modifier라고 판단하였다. 그 이유는 권역별로 의료시설 및 생활수준에 차이가 있다고 판단하였다. 우리나라의 의료시설은 대형병원 쏠림현상이 일어나고 있고, 각 병원마다 전문으로 하는 과목도 다르다. 권역별로 종사하는 직업군의 분포도 다를 것이

다. 따라서 6개의 권역을 나누어서 matching을 진행하였다. [1]

Matching을 진행한 방법은 다음과 같다.

1) 전체 데이터 셋에서 propensity score를 도출한다.
2) 도출한 Propensity score를 전체 데이터 셋에서 변수로 추가한다.
3) 데이터 셋을 6개의 권역으로 나눈다.
4) 각 권역별로 Matching을 진행한 후, Covariates balance를 살펴본다.
5) Covariates balance가 적절하다면, matching된 outcome들을 하나의 데이터 셋으로 만든다.
6) McNemar's Test[2]로 treatment 그룹과 control 그룹간 차이가 있는지 확인한다.

6. Conclusion

1) Without Replacement

Without matching 방법은 propensity score의 절대값을 이용하여 distance를 구한 후[3], treatment와 가장 가까운 거리의 control을 matching하는 것이다. 그러나 이미 앞에서 matching이 된 control 개체는 이후의 matching에서는 중복으로 사용할 수 없다.

Plot 5-1을 살펴보면, matching 이전에는 'BMI'와 'AGE'의 covariates가 큰 차이를 보였다. Matching이 된 이후 Covariates difference가 0.1 이내로 들어왔음을 확인할 수 있다. 모든 covariates의 balance가 조정되어 matching 되었다고 할 수 있다.

McNemar's Test를 진행해보았을 때, p-value는 약 0.48로 treatment와 control 집단간의 유의미한 차이가 있다고 할 수 없다.

	Before	**After**
BMI	-0.212	-0.067
PAST1	-0.041	-0.011
FMLY1	0.031	0.003
MOV30_WEK_NEW1	-0.190	-0.022

[1] 서울, 경기 및 인천, 충청, 호남 및 제주, 경북 및 강원, 경남

[2] Outcome은 binary여서 McNemar's Test 진행

[3] $\left|\hat{\lambda}(x_i) - \hat{\lambda}(x_j)\right|, i \neq j$

AGE	-0.265	0.051
CTRB_PT2	0.067	0.017
CTRB_PT3	-0.134	-0.072

Plot 5-1 : Covariates Difference for 'Without Replacement'

McNemar's Chi-squared test with continuity correction data: tab
McNemar's chi-squared = 0.4962, df = 1, **p-value = 0.4812**

Output 5-1 : McNemar's Test for 'Without Replacement'

2) With Replacement

With replacement 방법은 without replacement 방법과 유사하다. 이 방법도 propensity score의 절대값을 이용하여 distance를 구한 후, 가장 가까운 거리의 treatment와 control을 matching 한다. Without replacement 방법과의 차이는 control 개체가 앞에서 이미 다른 treatment와 matching이 되어도 이후 matching에서 중복 사용이 가능하다는 것이다.

Plot 5-2를 살펴보면, Covariates difference들은 matching이 된 이후 0.1 이내로 들어왔음을 확인할 수 있다. 따라서 이후 McNemar's Test를 진행하였다. P-value는 약 0.58로 treatment와 control 집단 간 유의미한 차이가 있다고 할 수 없다.

	Before	**After**
BMI	-0.212	0.000
PAST1	-0.041	-0.021
FMLY1	0.031	0.012
MOV30_WEK_NEW1	-0.190	-0.020
AGE	-0.265	0.014
CTRB_PT2	0.067	-0.004
CTRB_PT3	-0.134	-0.004

Plot 5-2 : Covariates Difference for 'With Replacement'

McNemar's Chi-squared test with continuity correction data: tab
McNemar's chi-squared = 0.29803, df = 1, **p-value = 0.5851**

Output 5-2 : McNemar's Test for 'Without Replacement'

3) Ranked-base Mahalanobis distance – Caliper 1000 matching

이 방법은 앞에서 이용한 distance와 다르다. 먼저 모든 개체들의 propensity score의 rank를 매

긴 후, Mahalanobis distance[4]를 이용하여 구한다. 또한 실제 covariates 들의 차이가 일정 수준 이상 차이가 난다면 caliper(=penalty)[5]를 부여한다. 이러한 방법을 이용하여 distance를 수정한 후, matching을 진행해보았다.

Matching을 진행한 이후, 모든 covariates difference가 0.1 이하인 것을 확인하였다. Covariates balance가 조정되었음을 확인하고 McNemar's Test를 진행해본 결과, p-value는 약 0.11로 앞에서 진행한 Mcnemar's Test의 p-value들보다 상당히 감소하였음을 확인할 수 있다. 그러나 담배가 폐질환에 유의미한 영향을 준다고 할 수 없다.

	Before	**After**
BMI	-0.212	-0.061
PAST1	-0.041	0.067
FMLY1	0.031	0.081
MOV30_WEK_NEW1	-0.190	0.010
AGE	-0.265	-0.002
CTRB_PT2	0.067	0.005
CTRB_PT3	-0.134	-0.044

Plot 5-3 : Covariates Difference for 'Ranked-base Mahalanobis distance – Caliper 1000 matching'

McNemar's Chi-squared test with continuity correction data: tab
McNemar's chi-squared = 0.4962, df = 1, **p-value = 0.4812**

Output 5-3 : McNemar's Test for 'Ranked-base Mahalanobis distance – Caliper 1000 matching'

4) Full Matching

Full Matching은 paired matching과는 다르게 모든 control을 treatment와 matching하는 것이다. 그러나 모든 control을 matching한다면 efficiency size는 줄어들게 되어 실제 paired matching보다 정보가 줄어들게 된다.

이 방법은 앞에서 한 matching이 잘 적합 되었으므로 covariates difference만 살펴보았다. Plot 5-4를 살펴보면, 앞에서 진행한 방법들과 마찬가지로 covariates balance가 적합 되었음을 알 수 있다.

---

[4] $(x_i - x_j)^T \hat{\Sigma}^{-1} (x_i - x_j)$

[5] 실제 논문에서 사용한 caliper의 개념은 propensity score의 distance의 허용범위를 의미

	Before	After
BMI	-0.212	-0.009
PAST1	-0.041	-0.052
FMLY1	0.031	-0.022
MOV30_WEK_NEW1	-0.190	0.001
AGE	-0.265	-0.001
CTRB_PT2	0.067	0.001
CTRB_PT3	-0.134	0.001

Plot 5-4 : Covariates Difference for 'Full Matching'

5) Sensitivity Analysis

앞에서의 결과들을 볼 때, Conclusion이 기각되었지만 Sensitivity Analysis까지 적용해보는 것으로 연구를 마치려고 한다. Sensitivity Analysis를 위해 사용된 Matching 방법은 p-value 값이 가장 작았던 'Ranked-base Mahalanobis distance – Caliper 1000 matching' 이다. Sensitivity Analysis의 Gamma 값은 1부터 2까지 0.1씩 더한 시퀀스를 사용하였다. 결과적으로 free of hidden bias일 때도, p-value는 약 0.96으로 conclusion은 채택되지 않음을 확인할 수 있다.

Gamma	lower	upper
1.0	0.9562	0.9562
1.1	0.7680	0.9964
1.2	0.4375	0.9998
1.3	0.1651	1.0000
1.4	0.0421	1.0000
1.5	0.0076	1.0000
1.6	0.0010	1.0000
1.7	0.0001	1.0000
1.8	0.0000	1.0000
1.9	0.0000	1.0000
2.0	0.0000	1.0000

Plot 5-5 : Sensitivity Analysis

7. Limitation

일반적으로 알고 있는 상식은 담배는 폐질환에 유의미한 영향을 준다는 것이다. 그러나 Nonexperimental study 결과, 담배가 폐질환에 영향을 주지 않는다고 할 수 있다. 따라서 연구의 한계점에 대해 논의하고자 한다.

1) Confounder

질환이 발병된다는 것은 복합적인 원인들이 작용하게 된다. 특히 폐질환인 경우 가족 중

흡연자 유무, 특정 직업군(화학물질에 노출되거나 직업성 분진에 노출되는 직업) 등, 고려해야한다. 그러나 실제 데이터에서는 위와 같은 정보는 구할 수 없었다.

2) 정보의 손실

처음 정제한 데이터셋은 약 8000개의 개체를 가지로 있었다. 그러나 data cleansing과 matching 이후 데이터셋은 약 2000개 정도의 dataset이 matching 되었다. Confounder도 간단한 factor형으로 수정되면서 데이터가 많이 손실되었다.

3) Binary outcome

질환이 발병하는 이유는 몸 내부에 촉발물질이 생기거나, 수치가 비정상적인 것이 원인이 되는 것이다. 그러나 그 수치는 절대적인 것이 아니며, 촉발물질의 수치가 비정상적인 것은 질환에 노출될 확률이 높은 것이다. 따라서 일정 촉발물질에 대한 수치와 관련하여 연구를 진행한다면 더욱 정확한 연구결과를 얻을 수 있을 것이다.

4) Design

연구 설계 진행과정에서 'former smoker'를 control 그룹에 포함하여 연구를 진행하였다. 사실 former smoker는 예전에 담배에 대해 노출된 사람이다. 그러나 현재는 담배에 노출되지 않는 사람들이다. 따라서 더욱 정확한 연구를 위해서 'former smoker'애 대해 제외 혹은'pack-year'와 관련된 자료를 얻어 분석하면 정확해질 것이다. 또한 2009년과 2013년 사이의 사망자들을 제외하였다. 그러나 실제로 2013년 이전 폐질환을 진단받거나 폐질환으로 인해 사망할 수 있기 때문에 drop out 데이터도 고려해야한다.