Propensity Score Model 구축에서 상관성을 고려한 변수선택

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
박 성 훈

Propensity Score Model 구축에서 상관성을 고려한 변수선택

지도 남 정 모 교수

이 논문을 석사 학위논문으로 제출함

2009년 12월 일

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

박 성 훈

차 례

표 차례

그림 차례

국문요약

Propensity score model 구축에서 상관성을 고려한 변수선택

Propensity score는 관찰연구에서 matching, 층화, 회귀보정 등의 방법으로 그 쓰임새가 증가하고 있다. Propensity score model(PSM)을 구축할 때 공변량 선택이라는 문제가 발생할 수 있다. 기존 연구에서는 propensity score를 구하는데 측정가능한 모든 공변량을 이용하는 것이 추천되어 왔다. 이러한 과적합 PSM을 이용할 경우 다중공선성 문제가 발생할 수 있으며, 충분한 matching number를 확보할 수 없는 문제가 발생한다.

본 연구에서는 PSM에 포함되는 공변량들의 상관성을 고려하여 propensity score matching을 수행하는 모의실험을 수행 하였다. 측정되어진 공변량들을 처리변수, 결과변수와의 상관관계를 구별하고, 각각의 범주에 따라 상관관계를 구했다. 공변량간의 상관관계와 상관성의 강도에 따라 PSM의 오즈비, MSE, matching number를 측정하여 결과에 영향을 주는 공변량을 확인해 보았다. 그 결과, 공변량들간에 상관관계가 존재하고 강도가 강할수록 matching number가 작아졌다. PSM에 포함된 공변량과 강한 상관성이 있는 공변량은 제거되더라도 공변량의 균형성이 크게 깨어지지 않는 것을 확인할 수 있었다.

본 연구를 통해 제안된 변수선택방법을 이용하면 PSM의 불필요한 과적합을 하지 않고도 matching number를 증가시킬 수 있다고 판단된다.

핵심되는 말 : propensity score, matching, simulation, 변수선택

제 1장 서론

1.1 연구 배경 및 목표

임상의학분야에서 사용되어지는 관찰연구 중에서 두 군을 비교하는 연구는 대부분 비무작위 시험을 기준으로 되어 왔다. 이러한 비무작위 시험은 선택편의(selection bias)를 통제할 수 없는 문제점을 가지고 있다. 선택편의는 치료집단과 대조집단 간의 이질성으로 인해 발생되는 것으로, 치료와 결과의 인과관계에 대한 올바르지 못한 추론을 하거나 치료효과를 과소 혹은 과대 추정하는 오류를 발생시키게 된다. 따라서 치료군에 대응되는 대조군을 선별할 때 공변량(covariate)들의 불균형(unbalanced)을 통제할 수 있는 대상이 선정 되어야 할 필요성이 있다. 두 집단에서 공변량들을 균등화하면서 구조적인 군 간의 차이가 발생하지 않도록 하기 위한 방법으로 두 군간 matching을 수행하거나 비슷한 개체끼리 층화시킨 후에 처리효과를 추정한다(노성유 2008).

처리변수와 결과변수가 이분형 자료인 경우 로지스틱회귀모형을 이용한 PSM (Propensity score model)을 구축하여 두 군간 개체들을 대응시키는 방법이 많이 사용되어지고 있다. 기존 저자들의 경우 측정할 수 있는 모든 공변량들을 PSM에 포함하므로써 측정되지 못한 잠재적 혼란변수를 통제하는 것을 제안해왔다(Sherry 2004).

Alan(2006)의 연구에서는 연관성(association)을 고려한 3개의 범주에서 공변량을 정의하여 모의실험을 수행하였다. 본 연구에서는 변수선택을 통해 Alan(2006)의 연구와 Austin(2007)의 연구실험설정을 확장하여 각 범주에서 2개의 공변량을 생성하여 모의실험을 수행하였고, 나누어진 각 범주별로 상관성(correlation)의 조합을 고려하여 PSM에 포함할 공변량의 특성과 상관성을 제시하고자 한다.

1.2 연구 내용 및 방법

본 연구는 모의실험을 위한 데이터를 생성하고 몬테카를로(Monte Carlo) 방법을 적용하여 통계량 평균값을 구한다. 데이터는 특정 분포에서 임의적으로 공변량들을 생성하고, 생성된 공변량들을 이용하여 처리변수, 결과변수간의 연관성과 공변량 간의 상관성 여부에 고려하여 처리변수와 결과변수를 베르누이시행을 통해 생성한다. PSM은 로지스틱회귀를 이용하며, 포함되는 변수의 특성에 따라 각기 다른 PSM을 구축한다. 각 모형으로 PSmatching을 수행한 표본에서 오즈비(odds ratio), MSE(mean squared error), matched number, 표준화차이계수(standardized differences) 값을 이용하여 통계량을 구하고 구해진 통계량들의 평균을 구하고 이 값들을 통하여 각 모형에 포함된 변수들의 상관성을 비교한다.

1.3 논문의 구성

1장에서 연구 배경 및 목적에 대해 소개하고 연구 내용 및 방법에 대해서 제시한다. 2장에서는 이론적 방법으로 propensity score의 개념에 대해서 설명하고, propensity score matching 알고리즘의 종류 설명한다. PSM의 평가측도인 표준화차이계수, 오즈비, MSE를 소개한다. 3 장에서는 공변량의 생성알고리즘, PSM 정의, 공변량들의 연관성과 상관성을 설정한 내용을 소개한다. 4장에서는 모의실험의 결과를 비교하고 구해진 통계량을 통해 각각의 PSM을 비교 평가한다. 5장에서는 결론 및 고찰에 대해서 논의하며, 제안 점을 서술한다.

제 2장 이론적 배경

2.1 Propensity score

2.1.1 인과효과

임상에서의 인과관계(Causal relationship)는 다음의 조건을 만족했을 때 성립된다.

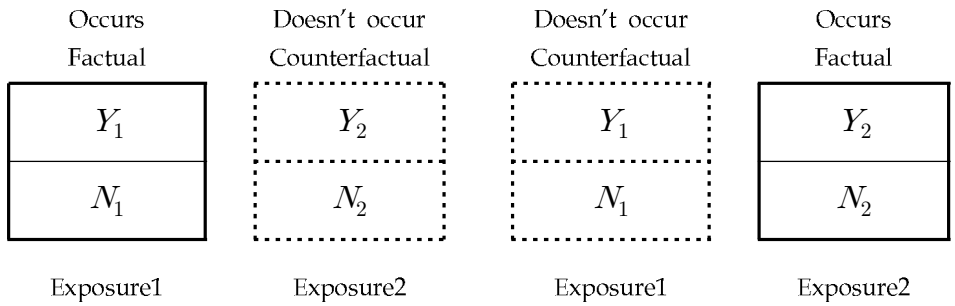
1. 두 사건 A와 B 사이에 공간적, 시간적, 인접성이 존재한다.
2. 한 사건 A가 다른 사건 B에 선행한다.
3. 전자 A가 일어나지 않았더라면 후자 B는 발생할 것 같지 않았을 경우이다.

	Occurs Factual	Doesn't occur Counterfactual	
No. of new cases	Y_1	Y_2	$P_1 = Y_1/N_1$
denominator	N_1	N_2	$P_2 = Y_2/N_2$
	Exposure1	Exposure2	

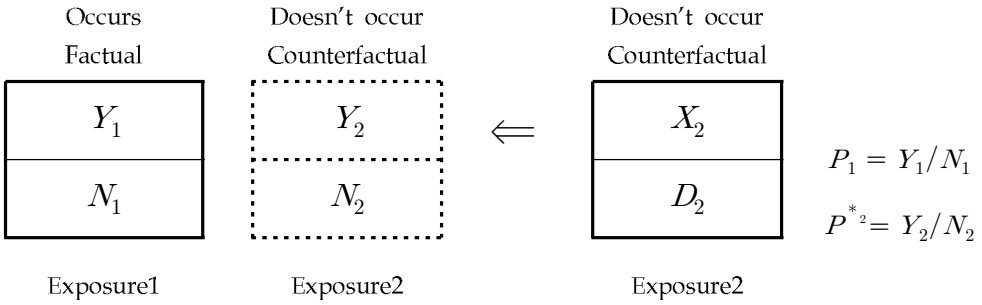
$$RR_{causal} = \frac{P_1}{P_2} = \frac{Y_1/N_1}{Y_2/N_2} \qquad RD_{causal} = P_1 - P_2$$

P_1과 P_2간의 격차는 하나의 대상 인구집단에서 단일한 병인 기간 동안, 노출 상태의 차이 때문에 나타나는 것으로 인과적 효과를 나타낸다.

그러나 현실에서의 임상연구대상은 하나의 처리만을 가정하기 때문에 단일한 대상이 서로 다른 노출 상태를 동시에 경험할 수는 없다.



반사실적상태(Counterfactual condition)는 실제로는 A와 B가 일어났지만, 만일 A가 발생하지 않았다면 B도 일어나지 않았을 것이라는 조건이다. 관찰되지 않은 대상인구집단의 반사실적상태 대신 대체집단을 사용하여 인과대비를 측정한다.



$$RR_{ass} = \frac{P_1}{P^{*}{}_2} = \frac{Y_1/N_1}{X_2/D_2}$$

2.1.2 강한 무관성가정(Strongly ignorable)

관찰연구에서 측정되어진 공변량에 의해 얻어진 효과의 추정치와 실제효과가 다른 편향(bias)가 발생할 수 있다. 또한 관측되지 않은 공변량에 관해서는 다룰 수 없는 제한점이 있기 때문에 두 가지 가정이 필요하게 된다.

조건부 독립성의 가정(conditional independence assumption)은 개인의 공변량이 주어졌을 때, 처리할당(Z)은 성과변수들(Y_0, Y_1)과 독립인 가정이며, 여기서 X는 공변량 벡터이다.

$$(Y_0, Y_1) \perp Z \mid X$$

공통 영역의 가정(common support assumption)은 두 처리집단의 할당 확률의 분포는 공통영역에 존재한다는 가정이다.

$$0 < \Pr(Z=1 \mid X) < 1$$

위의 가정들이 충족된다면, 처리할당을 측정할 수 있는 변수들이 충분히 존재한 경우, 이들을 통제하는 것으로 선택편향이 없는 인과효과 추정치를 산출해 낼 수 있음을 의미한다.

2.1.3 Propensity score

Propensity score는 관찰된 공변량들이 주어졌을 때, 특정 시험군에 할당되도록 영향을 주는 독립변수에 대한 조건부 확률로써 정의된다.

$$e(x) = pr(z = 1|x)$$

z는 처리수준을 나타내며 x는 측정되어진 공변량이다.

강한 무관성의 가정이 성립할 경우, propensity score를 이용하여 두 군의 개체들을 matching을 하여, 각 군의 측정 가능한 특성들은 동일한 분포를 가지게 된다. 이러한 것은 무작위 실험과 같이 선택편의가 없는 효과를 추정하게 해준다.

Rusenbaum과 Rubin(1983)은 propensity score의 특성을 제시 하였다. Propensity score는 균형점수(balancing score)이며, propensity score의 임의의 값에서, 처리집단간의 평균 차이는 평균처리효과의 불편추정량이다.

이러한 사실은 propensity score가 몇 가지 보정방법을 통해 관찰연구 자료에서 편의를 줄이는데 사용할 수 있다는 것을 뜻한다. Propensity score의 어떤 특정한 값에서 공변량들이 주어졌을 때 처리할당의 개념을 무시할 수 있는 상황이라면, matching하거나 층화, 혹은 공분산 보정 등을 통해 처리효과의 불편추정량을 구하는데 사용할 수 있다. 이렇게 구해진 propensity score의 값들을 가진 집단별 평균 차이는 평균처리 효과의 불편추정치가 된다(노성유 2008).

2.2 Propensity Score matching 알고리즘

2.2.1 국소최적알고리즘(Local Optimal Algorithms)

국소최적알고리즘은 주로 그리디(Greedy) 알고리즘이라고 불리기도 한다. 비교가 되는 처리군 A와 처리군 B 양쪽 군의 개체들을 무작위로 배열하고 처리군 A의 가장 위에 정렬된 개체부터 순차적으로 처리군 B 개체의 propensity score 차이를 구하고 이값들의 절대값이 가장 작은 것을 선택한다. 이러한 과정을 처리군 A의 모든 개체들에 수행한다. 이 알고리즘은 모든 처리군 A의 개체들에 수행 시 최적의 처리군 B와 matching되는 장점이 있지만, propensity score의 절대값 차이의 총합은 최소가 되지 못한다(Marcelo 2007).

국소최적알고리즘을 기반으로 하여 matching 방법에 따라 많이 쓰이는 3가지 형태가 있다. 최근접이웃방법(Nearest available neighbor matching)은 선택된 처리군 A의 개체와 처리군 B의 개체들을 대응시켜 거리가 가장 가까운 개체를 찾는 방법으로 적어도 한 개의 처리군 A의 개체가 한 개의 처리군 B의 개체와 대응되게 된다. 범위설정방법(Caliper matching)은 처리군 A를 중심으로 일정한 propensity score의 범위를 설정하여 이 범위에 들어오는 모든 처리군 B의 개체들 중에서 가장 가까운 개체를 선택하는 방법이다. 이 방법은 하나의 개체에 적어도 한 개의 개체가 선택되지 않을 수도 있다. 앞서 두 가지 방법을 혼합하여 수정한 방법으로 1:N 대응을 위한 반경설정방법(radius matching)이 있다. 이 방법은 대응이 된 개체를 반복적으로 대응시킬 때 사용한다(Dehejia 1999).

2.2.2 마할라노비스 행렬 matching

이 방법은 propensity score와 공변량을 이용하여 마할라노비스의 거리를 구하여 matching하는 방법이다. 두 군의 개체들을 무작위로 정렬하고 처리군 A의 각 개체와 처리군 B의 모든 개체들 간의 거리를 계산하는 방법으로 다음과 같다.

$$D_{ij} = \sqrt{(a_i - b_j)^T S^{-1} (a_i - b_j)}$$

여기서

D_{ij} : 처리군 A의 i번째 개체와 처리군 B의 j번째 개체의 마할라노비스 거리

a_i : 처리군 A의 i번째 개체의 다항벡터

b_j : 처리군 B의 j번째 개체의 다항벡터

S^{-1} : 비교된 모든 개체들로부터 선택된 분산-공분산 행렬이다.

D_{ij} 가 가장 작은 처리군 A의 i번째 개체와 처리군 B의 j번째 개체가 matching 되며, 선택된 개체는 각 군에서 제외된다. 이 방법의 단점은 계산된 마할라노비스 거리의 차원의 수가 증가할수록 개체 간의 평균거리도 함께 증가하기 때문에 모형 안에 많은 공변량들이 존재하면 가까운 개체들 간의 matching이 힘들다(노성유 2008).

2.2.3 대역최적알고리즘(Global Optimal Algorithms)

위의 다른 알고리즘은 최적의 matching을 하였으나, propensity score의 거리를 최소화시키지 못했다. Rosenbaum(1989)은 네트워크 방식의 광범위연결 방식인 그래픽 이론의 아이디어를 빌려왔다. 처리군 A와 처리군 B가 matching되는 것을 도식화하는 방식으로 재계산하여 잠재적 처리군 B와의 연결된 마디를 설정한다. 두 집단 개체들 간의 마디 값을 다시 재해석하여 총 거리가 최소가 되도록 찾아나가는 매칭방법이다.

2.3 평가변수

2.3.1 표준화차이계수

PSmatching에 의해서 대응된 표본에서 각 군 간의 변수들이 균등하게 분배된 정도를 확인하는 값으로 표준화차이계수를 사용한다(D'Agostino 1998).

이분형 변수의 표준화차이계수는

$$d = \frac{100(P_{treatA} - P_{treatB})}{\sqrt{(P_{treatA}(1-P_{treatA}) + P_{treatB}(1-P_{treatB}))/2}}$$

로 정의되며, 여기서 P_{treatA}와 P_{treatB}는 처리군과 대조군에서의 비율이다.

연속형 변수의 표준화차이계수는

$$d = \frac{100(\overline{x}_{treatA} - \overline{x}_{treatB})}{\sqrt{(S^2_{treatA} + S^2_{treatB})/2}}$$

로 정의되며, $\overline{x}_{treatA}$, $\overline{x}_{treatB}$와 S^2_{treatA} S^2_{treatB}는 처리군 A,B의 각각의 평균값과 분산을 의미한다.

표준화차이계수 d가 10% 이상의 값을 가지게 되면 대응된 표본의 두 군 간의 변수가 균등하게 배분되었다고 볼 수 없다 (Normand 2001).

2.3.2 Matched pair 데이터에서의 odds ratio

Matched pair 된 데이터에서 처리효과를 추정하기 위해서 다음의 범주형 테이블을 설정할 수 있다. 표 1의 테이블에서 a는 처리군 A와 처리군 B 양쪽에서 모두 사건이 발생한 빈도 값이며, b는 처리군 B에만, c는 처리군 A에만 사건이 발생한 빈도 값이다. d는 양쪽 모두에서 발생하지 않은 빈도 값이다.

표 1. Matched pair 데이터 빈도표

Matched pair		Treatment A subject	
		Outcome=1	Outcome=0
Treatment B subject	Outcome=1	a	b
	Outcome=0	c	d

Matched pair 된 데이터의 오즈비는 조건부 최대우도 추정량으로 c/b로 구해진다. 이 값은 개체특정적인 $2 \times 2 \times n$ 테이블에서 공통오즈비의 Mantel-Haenszel 추정값과 동일하게 된다(Agresti 2004, Austin 2007).

2.3.3 추정치에 관한 MSE

MSE의 추정치는 다음과 같이 계산된다.

$$\widehat{MSE} = \frac{1}{N}\sum_{n=1}^{N}(\exp(\beta_{treat}) - \hat{\gamma}(n))^2$$

여기서 , β_{treat}는 참 오즈비를 뜻하는 값이며, $\hat{\gamma}(n)$는 전체 N번의 모의실험에서 n번째 생성된 matched pair 데이터에서 추정된 오즈비이다(Alan 2006).

제 3장 모의실험

3.1 모형의 설정

처리변수와 결과변수에 대한 연관성 설정을 3개로 분류하였다. 처리변수와 결과변수에 동시에 영향을 주는 공변량(V1), 처리변수에만 영향을 주는 공변량(V2), 결과변수에만 영향을 주는 공변량(V3)으로 나누었다(Alan 2006). 상관성의 영향을 평가하기 위해 각각의 연관성을 가진 변수들을 2개씩 설정하여 총 6개의 변수를 설정하였다.

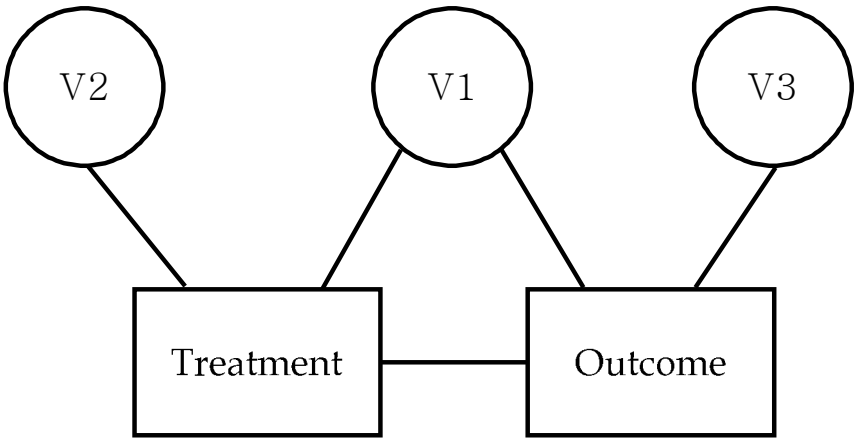


그림 2 처리-결과변수의 관계에 영향을 미치는 공변량의 가상적 모형 (Alan 2006)

3.2 PSM 설정

PS를 추정하기 위해 사용된 PSM은 로지스틱회귀모형을 기본으로 사용하였으며, PSM으로는 모든 변수를 사용한 모형과 연관성에 따른 모형으로 설정하였다. PSM1은 모든 변수를 넣은 모형이며, PSM2는 처리변수에 영향을 주는 변수들을 포함한 모형이다. PSM3은 결과 변수에 영향을 주는 변수들을 포함한 모형이며, PSM4는 처리변수와 결과변수에 동시에 영향을 주는 변수들을 제거한 모형이다. PSM5는 처리변수와 결과변수에 영향을 주는 두 변수 중에서 한 변수를 제외한 모형이다.

표 2. Propensity Score Model 설정

모형	모형에 포함된 공변량
PSM1	V11, V12, V21, V22, V31, V32
PSM2	V11, V12, V21, V22
PSM3	V11, V12, V31, V32
PSM4	V21, V22, V31, V32
PSM5	V11, V21, V22, V31, V32

3.3 공변량간의 상관성 설정

공변량간의 상관성은 첫 번째로 동일한 특성범주에 포함된 공변량들간의 관계에만 상관성을 설정하였고, 두 번째로 다른 범주에 있는 공변량 간에 상관성을 설정하였다. 공변량들의 상관성의 정도는 0.9, 0.7, 0.5, 0.3으로 설정하였다.

각 공변량들은 균일분포 [0,1]에서 독립적으로 랜덤하게 생성하였다.

$$V_{lm} \sim U(0,1) \quad (l = 1,2,3\ ,\ m = 1,2)$$

공변량 간의 상관관계가 있도록 아래와 같이 생성한다.

$$V_{2m} = a\, V_{1m} + U(0,1)$$

여기서 a값은 각 공변량 별로 상관관계의 강도를 조정할 수 있도록 설정된다.

3.4 공변량의 생성절차

데이터 생성은 Austin(2007)의 논문에서 제시하는 데이터 생성과정을 수행하였다. 처리변수에 연관성을 가지도록 공변량들의 계수를 다르게 주어 처리변수와의 연관성의 강도를 설정하며, 무연관인 공변량들은 계수값을 0을 주어 식에서 제외시킨다.

$$logit(P_{i,treatment}) = \beta_{0,treatment} + \beta_1 v_{11} + \beta_2 v_{12} + \beta_3 v_{21} + \beta_4 v_{22}$$

$$Z_i \sim Bernoulli(P_{i,treatment})$$

위의 식에서 처리변수에 할당할 확률을 구하고 N개의 개체에 대해서 Bernoulli 분포를 이용하여 처리변수(Z_i)를 생성한다.

결과변수의 생성은 처리변수의 생성과 동일하게 진행되며 T_i를 포함하여 확률을 구하고 Bernoulli 분포를 이용하여 생성시킨다.

$$logit(P_{i,outcome}) = \alpha_{0,outcome} + \beta_{treat} T_i + \alpha_1 v_{11} + \alpha_2 v_{12} + \alpha_3 v_{31} + \alpha_4 v_{32}$$

$$Y_i \sim Bernoulli(P_{i,outcome})$$

β_{treat}의 값은 처리변수와 결과변수의 로그 오즈비로 나타난다.

3.5 모의실험설계

표본은 10,000개의 개체를 생성시키며, 생성된 표본에서 PSmatching을 사용하여 대응된 표본의 숫자와 SE, odds ratio, 개별 변수의 표준화차이계수를 구한다. 이러한 표본을 100개를 생성하여 평균값을 구한다.

PSmatching에서 사용하는 Caliper의 값은 처리변수 A군의 propensity score의 표준편차의 4분의 1에 해당하는 값이 적당하다고 제안하였다(Ralph B 1998).

표본 데이터 생성시 처리변수와 결과변수를 생성시킬 때 임의로 정해주는 β_i와 α_i는 log(5)로 설정하였고, $\beta_{0,treatment}$의 값은 처리군 A군과 처리군 B군의 발생이 비슷한 비율이 되도록 모의실험을 통하여 정하였으며, 동일한 방법으로 $\alpha_{0,outcome}$의 값은 대조군에서 발생 비율이 25%가 되도록 하였다. 이 실험에서 $\beta_{,treat}$의 값은 오즈비가 1이 되도록 하였으며, 공변량들간의 상관성은 0.3, 0.5, 0.7, 0.9가 되도록 a값을 설정하였다.

몬테카를로 모의실험을 수행 시 사용한 프로그램으로 SAS 9.2를 사용하였으며, %PSMatching SAS macro를 실행하여 표본의 처리군 A, B군을 1:1 대응시켰다.

제 4장 모의실험 결과

4.1 상관성에 따른 모형 비교

다음 표는 상관성 설정을 달리한 PSM의 모의실험 결과이다.

표 3. 상관성 설정에 따른 모형 비교 (상관성 강도 0.7)

PS model		Odds ratio		MSE[a]		Matched number	
상관성	Model	Mean	SD[b]	Mean	SD[b]	Mean	SD[b]
Uncorrelated	PSM1	1.020	0.081	0.006	0.008	3351.33	38.08
	PSM2	1.021	0.033	0.001	0.001	3351.70	38.42
	PSM3	1.007	0.052	0.003	0.003	3851.83	40.01
	PSM4	1.392	0.050	0.156	0.039	3851.40	46.40
V1	PSM1	1.030	0.069	0.005	0.006	2912.23	44.81
	PSM2	1.015	0.045	0.002	0.003	2921.83	49.47
	PSM3	1.027	0.059	0.004	0.004	3216.37	36.77
	PSM4	2.144	0.154	1.330	0.360	3967.73	49.55
V2	PSM1	1.001	0.071	0.004	0.009	2931.47	43.22
	PSM2	1.015	0.060	0.003	0.004	2916.43	48.24
	PSM3	0.994	0.038	0.001	0.002	3967.23	47.10
	PSM4	1.431	0.086	0.193	0.066	3227.10	45.87
V3	PSM1	1.009	0.066	0.004	0.004	3348.43	36.69
	PSM2	0.999	0.053	0.003	0.004	3341.67	35.17
	PSM3	0.995	0.059	0.003	0.004	3865.13	51.1
	PSM4	1.367	0.113	0.146	0.081	3863.97	43.37
V1V2	PSM1	1.009	0.070	0.004	0.011	2658.17	51.54
	PSM2	1.013	0.060	0.003	0.003	2670.43	30.43
	PSM3	0.992	0.041	0.002	0.002	3400.00	37.30
	PSM4	2.073	0.088	1.159	0.188	3377.07	44.10

[a] Mean square error

[b] Standard deviation

표 3의 결과를 보면, PSM4모형이 각 상관성에서 오즈비, MSE, matched number가 크게 나타나는 것을 확인할 수 있으며, V1에 상관성이 있는 PSM4의 경우에 오즈비의 값이 크게 차이가 난다. PSM3과 PSM4의 결과를 비교해보면 PSM3은 처리변수에만 영향을 주는 V2를 제거한 모형임에도 큰 차이를 보이지 않으나 처리-결과변수에 동시에 영향을 주는 공변량인 V1을 제거할 경우에는 결과에 큰 차이를 줄 수 있다는 것을 확인할 수 있었다.

상관성이 무상관 관계에 있는 PSM의 결과들과 V3에만 상관성이 있는 PSM의 결과가 유사하게 나타나는 것을 확인할 수 있다. 이것은 결과변수에만 영향을 주는 공변량들의 상관성 여부는 PSM의 결과에 큰 영향을 주지 않는다는 것을 판단할 수 있다.

각각의 상관성에서 PSM1과 PSM2를 비교해 보면, 무상관인 경우보다 V1과 V2에서 matched number가 낮아지는 것을 확인할 수 있으며, V1V2에서 matched number가 가장 적은 수를 가지는 것을 볼 수 있다. 이 결과로 처리변수에 영향을 주면서 공변량들간에 상관성이 있는 공변량들이 PSM에 많이 포함 될수록 matched number가 낮아지는 것을 확인할 수 있다.

표 4의 결과를 보면, 모형에 포함되지 않는 공변량들의 불균등성을 확인할 수 있는데, 특히 상관성을 가지는 공변량들이 모형에 모두 포함되지 않을 경우 불균등성이 높게 나타난다. 두 공변량들간에 상관성이 있을 때 한 공변량이 모형에 포함될 경우 이 불균등성이 낮아진다.

표 4-1. 상관성 설정에 따른 모형별 표준화차이계수

PS Model		V11		V12		V21		V22		V31		V32	
상관성	Model	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]
Uncorrelated	PSM1	0.956	0.934	0.498	0.839	0.511	1.003	0.511	0.873	0.369	1.403	-0.188	1.135
	PSM2	0.471	0.788	0.901	0.985	0.581	0.757	0.500	1.115	0.359	2.293	-0.125	2.014
	PSM3	0.663	0.620	0.603	0.645	43.075	2.134	44.137	2.933	-0.073	0.820	-0.088	0.901
	PSM4	44.228	2.325	44.102	2.421	0.601	0.439	0.705	0.490	0.065	1.047	0.147	0.760
V1	PSM1	1.148	0.823	1.347	0.931	0.499	1.457	0.569	1.073	0.242	1.115	0.094	1.346
	PSM2	1.161	0.785	1.398	1.106	0.552	1.502	0.514	1.130	-0.183	2.812	0.017	2.450
	PSM3	1.173	0.595	1.465	0.502	44.523	2.557	43.525	2.540	-0.053	1.223	0.241	1.101
	PSM4	86.082	1.787	90.672	2.448	0.447	0.511	0.720	0.529	-0.239	0.973	0.111	0.836
V2	PSM1	0.458	0.985	0.734	0.925	1.140	0.881	1.351	1.068	-0.168	1.137	0.034	0.955
	PSM2	0.615	1.455	0.299	1.453	1.169	1.116	1.448	0.960	0.103	1.781	-1.185	2.368
	PSM3	0.682	0.547	0.509	0.582	86.654	2.741	90.871	2.471	0.063	0.823	0.079	0.811
	PSM4	43.526	3.293	43.655	2.916	1.209	0.519	1.442	0.400	-0.025	1.094	-0.101	1.294

[a] Standard deviation

표 4-2. 상관성 설정에 따른 모형별 표준화차이계수

PS Model		V11		V12		V21		V22		V31		V32	
상관성	Model	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]
V3	PSM1	0.386	0.876	0.641	0.895	0.642	1.020	0.725	1.193	-0.145	0.727	-0.135	0.853
	PSM2	0.757	1.192	0.639	1.139	0.651	1.005	0.403	1.010	0.663	2.098	0.176	2.560
	PSM3	0.718	0.513	0.562	0.476	44.103	1.634	45.206	2.564	0.125	0.693	0.018	0.741
	PSM4	44.313	2.086	43.501	2.167	0.644	0.697	0.672	0.651	0.044	0.891	0.252	0.912
V1V2	PSM1	1.025	1.222	1.288	1.169	1.030	1.289	1.164	0.876	0.015	1.297	-0.067	1.502
	PSM2	1.405	0.934	1.541	1.167	0.642	1.166	0.862	0.953	-0.017	3.227	0.034	2.832
	PSM3	1.375	0.616	1.182	0.460	86.005	2.615	90.532	2.969	-0.195	1.030	0.454	0.910
	PSM4	86.756	2.625	91.777	2.577	1.115	0.615	1.326	0.392	0.082	1.128	0.195	1.055

[a] Standard deviation

4.2 상관성의 강도에 따른 비교

표 5는 V12와 V22간에 상관성의 강도를 변화시키면서 PSM 별로 모의실험을 수행한 결과이다.

표 5. 상관성의 강도에 따른 모형 비교

PS model		Odds ratio		MSE[a]		Matched number	
상관성	Model	Mean	SD[b]	Mean	SD[b]	Mean	SD[b]
Uncorrelated	PSM1	1.007	0.044	0.002	0.002	3354.030	33.120
	PSM2	1.004	0.057	0.003	0.005	3334.330	45.030
	PSM3	1.023	0.049	0.003	0.004	3861.870	38.790
	PSM5	1.201	0.063	0.044	0.023	3581.930	38.800
0.3	PSM1	1.004	0.051	0.003	0.003	3228.500	45.130
	PSM2	1.008	0.050	0.002	0.004	3228.270	46.370
	PSM3	1.024	0.049	0.003	0.003	3685.570	43.160
	PSM5	1.194	0.052	0.040	0.020	3447.870	33.490
0.5	PSM1	1.010	0.054	0.003	0.003	3134.430	47.920
	PSM2	1.000	0.065	0.004	0.006	3133.370	39.510
	PSM3	1.028	0.055	0.004	0.004	3524.830	38.560
	PSM5	1.180	0.065	0.036	0.025	3306.730	45.180
0.7	PSM1	1.024	0.060	0.004	0.005	2923.700	36.290
	PSM2	1.010	0.056	0.003	0.004	2931.230	42.580
	PSM3	1.027	0.052	0.003	0.005	3241.300	45.280
	PSM5	1.121	0.055	0.018	0.016	3042.970	45.040
0.9	PSM1	1.006	0.056	0.003	0.003	2437.900	36.270
	PSM2	1.018	0.074	0.006	0.009	2446.130	30.470
	PSM3	1.040	0.070	0.006	0.010	2660.200	32.760
	PSM5	1.072	0.071	0.010	0.010	2465.230	45.330

[a] Mean square error

[b] Standard deviation

처리-결과변수에 연관성을 가진 V12와 처리변수에만 연관성을 가진 V22의 상관성의 강도를 0.3, 0.5, 0.7, 0.9로 변화시켰으며, PSM1, PSM2, PSM3, PSM5에 대해 모의실험을 수행하였다. PSM4는 V1범주의 공변량이 없는 모형이므로 이번 분석에서는 제외시켰다.

각 조건에서 PSM1과 PSM2의 오즈비와 MSE, matching number가 두 모형에서 비슷하게 나타나고 있으며, PSM3은 matching number가 가장 크게 나타나고 있다. PSM5는 오즈비와 MSE가 다른 모형에 비해 크게 나타나는 것을 확인할 수 있다. 이것은 4.1의 결과에서 밝혔듯이 처리-결과변수에 영향을 미치는 공변량이 PSM의 결과에 영향을 많이 준다는 것을 다시 한 번 확인할 수 있다.

상관성의 강도가 강할수록 PSM의 MSE 값들이 작아지는 것을 확인할 수 있고, 상관성의 강도가 0.5이하에서는 matching number가 크게 차이가 나지 않았지만 0.7 이상일 때부터는 작아지는 것을 확인할 수 있다.

표 6에서 PSM5의 V12 공변량의 표준화차이계수의 값은 상관성의 강도가 높아질수록 작아지고 있으며, V22는 커지는 것을 확인할 수 있다.

표 6-1. 상관성의 강도에 따른 모형 별 표준화차이계수

PS Model		V11		V12		V21		V22		V31		V32	
상관성	Model	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]
Uncorrelated	PSM1	0.191	1.017	0.636	0.719	0.802	0.991	0.764	1.162	0.098	0.995	-0.160	1.063
	PSM2	0.616	1.023	0.528	0.861	0.629	0.981	0.642	0.983	0.122	2.205	-0.156	1.893
	PSM3	0.644	0.579	0.646	0.575	43.891	1.911	43.527	2.079	0.102	0.970	0.252	0.908
	PSM5	0.836	0.674	46.239	2.553	0.656	0.792	0.381	0.776	0.063	0.816	-0.088	1.092
0.3	PSM1	0.729	1.016	0.686	0.984	0.579	0.926	0.654	0.868	-0.147	1.054	0.098	1.278
	PSM2	0.515	1.051	0.739	1.097	0.490	1.045	0.972	0.995	-0.130	2.404	0.103	2.407
	PSM3	0.539	0.754	0.826	0.629	44.049	2.486	42.501	2.430	-0.022	1.080	0.448	1.074
	PSM5	0.533	1.090	43.312	2.638	1.207	1.085	0.377	0.833	0.187	1.166	0.348	1.241
0.5	PSM1	0.614	1.280	0.932	0.981	0.948	0.980	0.602	0.891	-0.011	1.162	0.039	1.237
	PSM2	0.525	1.065	0.832	0.967	0.548	1.259	1.043	0.719	0.553	1.946	0.408	2.355
	PSM3	0.534	0.992	0.945	0.651	44.269	2.174	40.110	1.929	-0.243	0.972	0.113	1.220
	PSM5	1.199	1.101	38.169	1.991	0.702	1.230	0.469	0.854	-0.215	1.085	-0.141	1.164

[a] Standard deviation

표 6-2. 상관성의 강도에 따른 모형 별 표준화차이계수

PS Model		V11		V12		V21		V22		V31		V32	
상관성	Model	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]	Mean	SD[a]
0.7	PSM1	0.651	1.203	1.069	1.040	0.825	1.270	1.116	0.949	0.089	1.297	0.088	1.306
	PSM2	0.969	1.345	0.852	1.004	0.147	1.175	1.467	0.811	-0.631	1.898	0.335	2.444
	PSM3	0.408	0.906	1.332	0.526	43.962	2.302	34.215	1.703	0.117	1.080	0.156	1.170
	PSM5	0.536	1.252	27.957	1.705	0.554	1.101	1.288	0.755	-0.052	1.242	0.153	1.343
0.9	PSM1	1.028	1.300	1.442	0.782	0.344	1.372	1.800	0.891	-0.465	1.395	0.296	1.303
	PSM2	0.955	1.587	1.553	0.893	0.272	1.504	1.855	0.860	-0.456	3.324	0.254	3.518
	PSM3	-0.091	1.228	2.010	0.494	42.774	2.392	25.160	1.337	-0.190	1.239	0.369	1.553
	PSM5	0.494	1.388	13.730	1.248	-0.170	1.314	2.173	0.721	-0.054	1.438	0.207	1.390

[a] Standard deviation

제 5장 결론 및 고찰

PSM을 이용하여 matching number를 산출할 때, 모형에 들어가는 공변량으로는 처리변수에만 영향을 주는 공변량과 결과변수에만 영향을 주는 공변량, 그리고 처리-결과 두 변수에 동시에 영향을 주는 공변량으로 나누어 볼 수 있다.

본 연구에서는 공변량들 중에서 처리-결과변수에 영향을 주는 공변량이 PSM 결과에 가장 큰 영향을 주고, 그다음으로 처리변수에만 연관성이 있는 공변량이 영향을 주었다. 결과변수에만 영향을 주는 공변량의 경우 PSM 결과에 영향력이 거의 없다고 판단된다.

PSM에 포함되는 공변량들 간에 상관성이 존재하고, 이러한 상관성이 존재하는 공변량들이 처리변수에 연관성을 가지고 있으면 matching number가 줄어드는 것을 확인할 수 있었다. 공변량들간에 상관성의 강도가 높아질수록 MSE가 낮아지는 것을 확인할 수 있었으며, matching number가 줄어드는 것을 확인할 수 있었다. 상관성이 높은 공변량들 중에서 하나의 공변량이 PSM에서 빠지더라도, 빠진 공변량의 균형성의 정도가 심하게 불균형하지 않게 되었다.

이러한 결과들을 종합해 보면, PSmatching을 수행하기 위해 모든 공변량을 사용하여 과적합한 모형을 사용하기 보다는 공변량들과 처리변수, 결과변수들 간의 관계를 확인하고 상관성을 고려하여 PSM을 구축하는 것이 효율적이라고 판단된다.

PSM 구축시 공변량 선택에 대한 제안사항을 다음과 같이 정리 할 수 있었다.

1. 처리변수와 결과변수에 연관성을 가진 공변량들을 선별하여, 처리-결과변수에 영향을 주는 공변량(V1), 처리변수에만 영향을 주는 공변량(V2), 결과변수에만 영향을 주는 공변량(V3)으로 나눈다.
2. 처리-결과에 연관성이 있는 공변량(V1)들 간에 상관성을 구하고, 상관성이 강한 공변량들을 묶어서 분류한다. 이 공변량들 중에서 균형성이 반드시 확보가 되어야 하며, 연구의 목적에 합당한 변수들을 남기고 상관성이 높은 공변량을 우선적으로 제거한다.

3. 제거되지 않은 V1 공변량들과 V2 공변량들의 상관관계를 구하고, V1에 상관관계가 높은 V2 공변량을 제거한다.

4. V2 간의 공변량들 중에서도 상관관계가 높은 공변량은 제거한다.

5. 결과변수에만 영향을 주는 공변량(V3)도 V2방법과 동일한 방법으로 V1에 상관성이 강한 공변량을 제거한다.

Austin(2007)의 연구에서 처리변수와 결과변수에 연관성의 강도는 log(5)를 강한 연관성으로 설정하였으며, log(2)를 약한 연관성으로 설정하였다.

기존 연구에서는 PSM에 포함되는 공변량을 처리변수에 연관성을 가진 공변량들만을 넣는 방법, 결과변수에 관계된 잠재적인 모든 공변량을 넣는 방법, 처리-결과변수 모두에 연관성을 가진 변수를 넣는 방법들이 제안되어 왔다(Austin 2007).

Alan(2006)의 연구는, 결과변수에만 상관성을 가지는 공변량은 PSM에 포함이 되어도 편차를 증가시키지 않으며, 처리변수에만 상관성이 있는 공변량은 처리효과의 분산을 증가시킨다고 하였다. Austin(2007)의 연구에서는 처리-결과변수에 상관된 공변량만을 포함한 PSM의 matching number가 가장 크게 가진다고 보고되었다.

본 연구에서는 기존 연구의 설정에 공변량들간의 상관성을 추가하여 수행하기 위해 공변량들을 연속형 변수로 설정하였으며, 공변량들간의 상관성 여부와 상관성의 강도가 달라지면 생성되는 데이터가 다르게 되는 문제점을 갖고 있다.

본 연구에서는 matching방법만을 이용하였으나 층화와 회귀보정 등의 방법에서도 공변량의 상관성을 고려한 연구가 이루어 져야 할 것으로 판단된다. 또한 생성오즈비를 변화시켜 보면서 결과를 분석해 보아야 할 필요성이 있다.

참고 문헌

노성유. 2008. Propensity Score matching 방법을 이용한 간경변증 위험 인자의 재평가. 석사학위논문. 연세대학교 대학원.

손건태. 2005. 전산통계개론. 자유아카데미. 제4판.

Agresti, A., Min, Y. 2004. Effects and Non-Effects of Paired Identical Observations in Comparing Proportion with Binary Matched-Pairs Data. *Stat Med*, 23: 65-75.

Austin, P. C., Grootendorst, P., Anderson, G. M. 2007. A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables Between Treated and Untreated Subjects: a Monte Carlo Study. *Stat Med*, 26: 734-753.

Binder, K., Heermann, D. W. 1997. *Monte Carlo Simulation in Statistical Physics*, 3rd ed. Springer.

Brokhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., Stürmer, T. 2006. Variable Selection for Propensity Score Models. *Am J Epidemiol*, 163: 1149-1156.

D'Agostino, R. B. 1998. Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group. *Stat Med*, 17: 2265-2281.

Jones, O., Maillardet, R., Robinson, A. 2009. *Introduction to Scientific Programming and Simulation Using R.* Chapman & Hall Book.

Maldonado, G., Greenland, S. 2002. Estimating Causal effects. *Int J Epidemiol,* 31: 422-429.

Marcelo, C. P. 2007. "Local and Global Optimal Propensity Score Matching". *SAS Global Forum 2007,* 185.

Ming, K., Rosenbaum, P. R. 2001. A Note on Optimal Matching with Variable Controls Using the Assignment Algorithm. *J Am Stat Assoc,* 10: 3; 455-463.

Rosenbaum, P. R. 1989. Optimal Matching for Observational Studies. *J Am Stat Assoc,* 84: 408; 1024-1032.

Rosenbaum, P. R., Rubin, D. B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika,* 70: 41-55.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., Mor, V. 2004. Principles for Modeling Propensity Scores in Medical Research: a Systematic Literature Review. *Pharmacoepidemiol Drug Saf,* 13: 841-853.

ABSTRACT

Variable selection for Propensity score Models considering the Correlations between Covariates

Park, Seong Hun
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

The propensity score is increasingly being used to match and stratify the data in observational studies. However, there maybe a problem of covariate selection included in the propensity score model(PSM). In the previous researches, including all the covariates that can be observed has been recommended. In this point of view, there are problems that appear multicollinearity and do not obtain the matching number needed using overfitted propensity score model.

In this thesis, we studied the method of variable selection for PSM considering the correlations between covariates.

All the covariates were classified according to the relation with treatment and outcome and generated considering the correlations each other. We examined the odds ratio and MSE (Mean Squared Error) of PSM and the matching number of simulated data.

The matching number decreased as the correlation of covariates was stronger.

In conclusion, we found that our procedure of variable selection for PSM had

the advantages that increase the matching number without using all the covariates

Key words : propensity score, matching, simulation, variable selection