

# 경영경제데이터분석

PSM

(Propensity Score Matching)

최 현 홍

(hongchoi@khu.ac.kr)

# Contents

- Introduction
- What is PSM?
- Conducting PSM Analysis
- Quick Review with Another Dataset
- Comparing PSM with Other Methods
- Recap
- Appendix: R Statistical Software

# **Introduction**

# Introduction

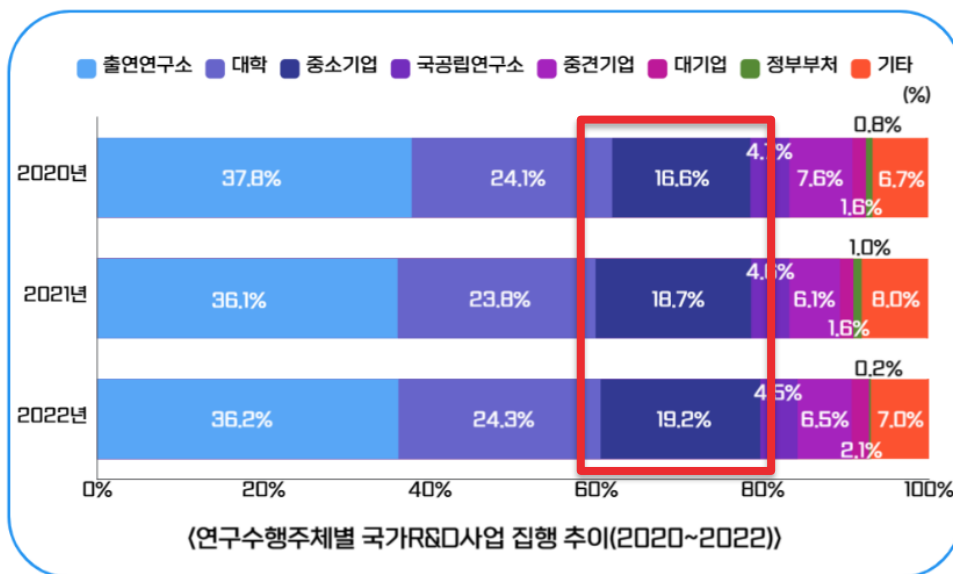
## 정부 R&D 지원과 기업 혁신 (1)

- 정부는 기업의 혁신활동을 증진시키고 혁신 성과물을 창출하기 위해 다양한 정책 수단을 활용함
  - 왜? 혁신 성과물을 바탕으로 기업이 성장하게 되면 고용 창출, 수출 증대, 경제 성장 등의 긍정적 효과들이 발생함
- 그런데, 특히 중소기업의 경우 대기업 및 중견기업 대비 자원의 제약, 정보의 비대칭성 등의 어려움을 겪기 때문에 혁신투자에 소극적일 수밖에 없음
  - 물론 대기업 대상 지원도 존재함
- 이에 따라 정부는 중소기업 대상 다양한 형태의 지원을 제공함
  - 보조금, 세제지원(R&D 조세지원 등), 금융지원(기술보증, 신용보증 등), 인증(벤처인증, 이노비즈인증 등), 관련 정보 지원(DB구축, 보고서 발간), 공공구매 등

# Introduction

## 정부 R&D 지원과 기업 혁신 (2)

- 정부가 할 수 있는 여러 지원 중 R&D 지원은 혁신 성과물 창출의 원천이 되는 R&D 활동을 지원하는 방식
  - 우리나라를 비롯해 전 세계적으로 널리 이루어지고 있는 정책
  - 초기 수익 창출이 어려운 기술 스타트업의 혁신기술개발 및 생존에도 큰 도움



ninewatt

Beeble

# Introduction

## 정부 R&D 지원과 기업 혁신 (3)

- 그런데, 일각에서는 기업 대상 R&D 지원이 기업 혁신 성과 창출에 큰 효과가 없다고 주장하기도 함
- 아래와 같은 논쟁 상황을 생각해보자
  - 정부: R&D 보조금을 지원 받은 기업과 그렇지 못한 기업의 성과를 비교했더니 지원 기업의 성과가 더 좋습니다!
  - ???: 선정 과정에서 더 유망한 기업을 지원했으니 당연한 것 아님?
- 즉, R&D 지원이라는 처치(treatment)가 실제로 기업들의 혁신 성과를 창출하는 데 기여하는지에 대한 엄밀한 인과성 추론이 필요함
  - 공동의 자원인 세금을 사용하는 일이기 때문에 해당 인과성을 밝히는 문제는 매우 중요함 (정책 효과성)

# Introduction

## 정부 R&D 지원과 기업 혁신 관련 연구

- 정부 R&D 지원이 제조기업의 혁신활동 및 혁신성장에 미치는 효과
  - 오승환 & 장필성 (2020)
    - 분석 대상 인과 관계: 정부 R&D 지원 → 기업 제품혁신 성과
      - ✓ 분석 대상 처치: 정부 R&D 지원 수혜 여부
      - ✓ 2012~2017년 정부 R&D 지원 여부가 국내 제조기업의 제품혁신성  
과에 영향을 미쳤는지 여부를 밝히고자 함
    - 고려 혼동 변수: 기업의 여러 특성 (후술)
    - 분석 방법: **PSM**
    - 활용 자료: 기업혁신조사 (KIS) - <https://www.stepi.re.kr/kis>

# Introduction

## 정부 R&D 지원과 기업 혁신 관련 연구: 혼동 변수

- 고려한 혼동변수 목록
  - 업력
  - 매출액
  - 종사자 수
  - 석사이상인력 비중
  - 연구인력 비중
  - ...

처치(정부 R&D 수혜)에 영향을  
미칠만한 요인은?



# Introduction

## 정부 R&D 지원과 기업 혁신 관련 연구: 매칭 이전

- PSM을 이용한 매칭 이전 기초통계량 비교 수혜기업의 주요 지표가 더 우수  
중소기업의 경우 특히 편차가 큼

<표 6> 정부 R&D 수혜기업과 비수혜기업 간 기초통계량 비교

구분	대기업					중소기업				
	R&D 수혜기업 (35 obs.)		R&D 비수혜기업 (56 obs.)		유의도	R&D 수혜기업 (533 obs.)		R&D 비수혜기업 (2,867 obs.)		유의도
연속변수	평균	표준편차	평균	표준편차		평균	표준편차	평균	표준편차	
업력 (년)	27.2	13.6	25.9	13.2		19.3	11.1	17.0	10.5	***
매출액(억원)	8379.3	14370.1	4102.2	10906.9		465.9	839.0	247.3	727.6	***
종업원수 명)	936.0	1199.2	624.6	1770.7		115.7	137.8	58.4	90.5	***
석사비중 (%)	10.8	14.7	7.7	15.7		8.0	12.3	3.4	9.1	***
연구인력비중 (%)	15.8	20.0	7.4	11.2	**	14.7	14.4	6.1	10.0	***

통상적 유의도 표현:

\*, \*\*, \*\*\* 각각 10%, 5%, 1% 유의수준에서 유의

t-검정 결과

t-검정 결과

# Introduction

## 정부 R&D 지원과 기업 혁신 관련 연구: 매칭 이후

- PSM을 이용한 매칭 이후 기초통계량 비교 **비록 완벽하진 않지만... 매칭 이전보단 나은 결과**

<표 7> 정부 R&D 수혜기업과 매칭기업 간 기초통계량 비교

구분	대기업					중소기업				
	R&D 수혜기업 (35 obs.)		R&D 비수혜기업 <b>35</b> obs.)		유의 도	R&D 수혜기업 (533 obs.)		R&D 비수혜기업 <b>533</b> obs.)		유의 도
연속변수	평균	표준편차	평균	표준편차		평균	표준편차	평균	표준편차	
업력 (년)	27.2	13.6	28.9	14.8		19.3	11.1	18.5	10.9	
매출액(억원)	8379.3	14370.1	4600.0	14466.3		465.9	839.0	439.3	992.4	
종업원수 명)	936.0	1199.2	381.5	158.0		115.7	137.8	102.8	136.6	
식사비중 (%)	10.8	14.7	7.9	13.2		8.0	12.3	6.0	11.8	***
연구인력비중 (%)	15.8	20.0	8.0	9.9	*	14.7	14.4	11.9	13.7	***

**이러한 결과를 인지하는 것은 왜곡된 해석을 막는 좋은 방법 중 하나**

# Introduction

## 정부 R&D 지원과 기업 혁신 관련 연구: 국내/세계 최초 혁신 성과 비교

- 정부 R&D 수혜 여부에 따른 **국내최초** 혹은 **세계최초** 제품혁신 비중

<표 14> 정부 R&D 수혜 여부에 따른 제품혁신 수준 : 국내최초, 세계최초

구분		표본기업수	국내 최초	세계 최초
비수혜기업	전체	2,932	116 *** (4.0%)	11 (0.4%)
	매칭기업	568	44 ** (6.92%)	4 (0.63%)
수혜기업		568	59 (10.4%)	5 (0.9%)

<R&D 지원 → 국내최초 제품혁신>

국내최초 제품혁신성과 비중 (차이 유의):

수혜기업 (10.4%) vs 전체기업 (4.0%)

수혜기업 (10.4%) vs 매칭기업 (6.92%)

<R&D 지원 → 세계최초 제품혁신>

세계최초 제품혁신성과 비중의 경우

차이가 유의하지 않음

(유의도 차이는 카이제곱 검정 결과 - 혁신 유무 이므로)

# Introduction

## 정부 R&D 지원과 기업 혁신 관련 연구: 시장/자사 최초 혁신 성과 비교

- 정부 R&D 수혜 여부에 따른 시장최초 혹은 자사최초 제품혁신 비중

구분		표본기업수	시장 최초	자사 최초
비수혜기업	전체	2,932	148 *** (5.0%)	444 *** (15.1%)
	매칭기업	568	57 ** (8.96%)	136 * (21.4%)
수혜기업		568	72 (12.7%)	146 (25.7%)

<R&D 지원 → 시장최초 제품혁신>  
시장최초 제품혁신성과 비중 (차이 유의):  
수혜기업 (12.7%) vs 전체기업 (5.0%)  
수혜기업 (12.7%) vs 매칭기업 (8.96%)

(유의도 차이는 카이제곱 검정 결과)

<R&D 지원 → 자사최초 제품혁신>  
자사최초 제품혁신성과 비중 (차이 유의):  
수혜기업 (25.7%) vs 전체기업 (15.1%)  
수혜기업 (25.7%) vs 전체기업 (21.4%)

앞선 매칭 결과를 고려할 때,  
결과를 해석할 때 주의해야 할 점은?

**What is PSM?**

## 성향점수 매칭(Propensity Score Matching, PSM)

- PSM은 인과성 분석을 수행함에 있어 무작위통제실험이 어려운 경우 자주 활용되는 매칭 방법 중 하나
  - 처치군과 대조군 사이의 차이를 통제하기 위한 통계적 기법
- **경향 점수 라고도 함**
  - 혼동 변수들을 바탕으로 추정된 성향 점수(propensity score)를 바탕으로 적절한 비교 대상을 짝지어주는(matching) 방법론
    - 성향점수: 어떠한 처치를 받을 확률을 나타내는 지표
    - 즉, 처치군에 적절히 대응하는 대조군을 생성해줌 (혹은 반대)
- 해당 방법론을 활용함에 있어 핵심은
  - 성향점수를 어떻게 추정하는가?
  - 표본 매칭을 어떤 방식으로 하는가?

## 인과성 추론에서 PSM의 의의

- 실무에서는 직접 무작위통제실험을 하기보다는 이미 존재하는 데이터를 분석하게 되는 경우가 (훨씬) 더 많을 것
  - 인과성을 밝힘에 있어 무작위통제실험은 가장 바람직한 대안 중 하나이긴 하지만, 실험과 관련된 시간, 비용, 윤리 이슈 등이 존재
- PSM은 이미 확보된 데이터를 바탕으로 추가적인 처리를 거쳐 실험에 준하는 분석 환경을 구축할 수 있도록 해주는 방법론
  - 준실험(quasi-experimental) 방법

## 관찰 연구 vs 실험 연구

PSM은 관찰연구에서 실험연구의 장점을 누릴 수 있게 해 줌

	관찰연구 (observation study)	실험연구 (experimental study)
정의	<ul style="list-style-type: none"> <li>연구자가 직접 개입하지 않고 자연 상태에서 발생하는 데이터를 관찰하여 분석</li> </ul>	<ul style="list-style-type: none"> <li>연구자가 특정 변수를 조작(처치)하고 이를 무작위로 할당하여 그 효과를 관찰 (무작위통제실험 등)</li> </ul>
장점	<ul style="list-style-type: none"> <li>데이터 수집이 비교적 용이함</li> <li>실험연구가 불가능할 경우에 유용하게 활용될 수 있음</li> <li>넓은 범위 및 집단에 대한 연구 용이</li> </ul>	<ul style="list-style-type: none"> <li>인과관계의 명확한 설정이 가능</li> <li>무작위할당을 통한 혼동 변수 영향 최소화</li> </ul>
단점	<ul style="list-style-type: none"> <li>혼동 변수의 위험</li> <li>선택 편향의 위험</li> <li>=인과관계의 확립 어려움</li> </ul>	<ul style="list-style-type: none"> <li>시간과 비용이 많이 소요될 수 있음</li> <li>윤리적 문제가 발생할 수 있음</li> <li>= 적용이 제한되는 경우가 많음</li> </ul>





## 번외) 온라인 무작위 통제실험

- 최근에는 온라인 환경에서의 실험이 용이해졌기 때문에, **기업이 자체적으로 무작위 통제실험을 수행**할 수도 있음
- E-커머스 기업인 HH마켓이 **새로 개발한 제품 추천 시스템**이 기존 시스템보다 더 효과적인지 여부를 검증하고자 한다면?
  - 자사 고객들을 2개의 그룹(A, B)으로 무작위 배정하고
  - A그룹 고객들에게는 기존 제품 추천 시스템 적용 (**대조군**)
  - B그룹 고객들에게는 신규 제품 추천 시스템 적용 (**처리군**)
  - 추후 A, B 그룹간 판매 성과 및 비교 및 통계적 검증
- 주의할 점
  - **윤리적/법적 문제나 고객 불만**이 발생할 여지가 없는지?

## PSM을 활용한 연구들 (1) 처치 여부에 따른 표본 특징이 유사할까 생각해보자

### 성향점수매칭을 이용한 코스닥시장 상장기업의 장기성과 분석

표한형, 홍성철 - 응용경제, 2013 - dbpia.co.kr

본 논문은 신규 상장이 가장 활발했던 시기인 2000-2002년까지 코스닥시장에 신규 상장한 기업들을 처리군으로 하고 상장요건을 갖추었지만 상장을 하지 않은 외감법인을 대조군으로 하여 ...

☆ Save Cite Cited by 5 Related articles All 3 versions

처치: 코스닥 상장

처치: 보조금 수혜

### [PDF] 성향점수 매칭을 이용한 정부 연구개발 보조금 효과분석

최석준, 김상신 - 한국산학기술학회논문지, 2009 - kais99.org

... 성향점수 매칭(propensity score matching)이 사용된다. 두 번째는 산업 내 기업수준의 패널 데이터를 사용한 연구로, 시간에 따라 변하지 않는 기업의 특성을 통제변수로 하여 정부 지원에 ...

☆ Save Cite Cited by 37 Related articles All 5 versions

### 외국어 고등학교 학교효과 분석: 성향점수 매칭모형을 활용하여

민병철, 박소영 - 한국교육, 2010 - scholarworks.sookmyung.ac.kr

... 이를 위해 한국교육고용패널의 중학교 3학년 코호트 자료와 2008학년도 대학수학능력시험 결과 자료를 성향점수 매칭모형(PSM)과 중다회귀분석(OLS)이 사용되었다. 분석결과 첫째, 전체 ...

☆ Save Cite Cited by 7 Related articles All 2 versions

처치: 외고 진학

처치: 바우처 지원

### AI 중소기업 바우처 지원이 기업성과에 미치는 영향: PSM-DID 결합모형을 활용한 정책효과 분석

최석원, 이주연 - 한국산업정보학회논문지, 2023 - dbpia.co.kr

... 이런 이유로 본 연구는 성향점수매칭(PSM)과 이중차분법(DID)을 활용하여 정부 인공지능 솔루션 바우처 지원 사업이 수혜기업의 경제적 성과에 미치는 정책효과를 살펴보고자 하였다. 실증...

☆ Save Cite Related articles

## PSM을 활용한 연구들 (2)    처치 여부에 따른 표본 특징이 유사할까 생각해보자

뇌졸중 환자와 건강군간의 생활습관과 삶의 질 비교: 성향점수매칭법을 활용하여

김민정 , 김진서 , 조영석 - 한국데이터정보과학회지, 2021 - dbpia.co.kr

본 연구는 뇌졸중을 앓은 대상자와 건강한 대상자의 생활습관과 삶의 질의 차이를 확인하기 위해 성향점수매칭법을 활용한 후향적 비교연구다. 연구대상은 국민건강영양조사 제7기 자료의 ...

☆ Save    Cite    Related articles

처치: 뇌졸중

처치: 임대 아파트

[PDF] 경향점수 매칭을 활용한 임대 및 분양 아파트 관리비 비교

이창로 , 박기호 - 대한지리학회지, 2017 - kgeography.or.kr

... 이후 인과관계적 추론을 위해 경향점수 매칭 기법을 활용하여 관리면적 등 특징이 서로 유사한 임대 아파트와 분양 아파트를 한 쌍씩 매칭하였다. 매칭된 축소 데이터에 기초하여분석한 결과, ...

☆ Save    Cite    Related articles    All 5 versions    ⌕

1인 가구 여부가 범죄두려움에 미치는 영향: 성향점수매칭을 이용한 차이 분석

최형근 , 박신의 , 민동기 , 장현석 - 한국범죄학, 2022 - papersearch.net

... 전국 범죄피해조사 2018 데이터를 활용하여 분석한 결과 성향점수매칭 이전에는 1인 가구와 다인 가구의 범죄두려움에는 유의미한 차이가 있었지만 매칭을 실시한 후에는 두 집단의 두려움...

☆ Save    Cite    Related articles    ⌕

처치: 1인 가구

### SEM

성향점수매칭과 구조적 모형을 이용한 육군 동원훈련 성과분석: 포병부대 주특기 훈련 중심으로: Performance Analysis of Army Mobilization Training Using Propensity ...

한봉규 - 2021 - dbpia.co.kr

... 방법론으로, 첫 번째 주제에는 성향점수매칭으로 교란요인을 제거한 후 회귀분석을 이용하여 적소 주특기와 훈련성과의 관계에 대해 연구하였고 두 번째 주제는 Ordered Probit, Bivariate ...

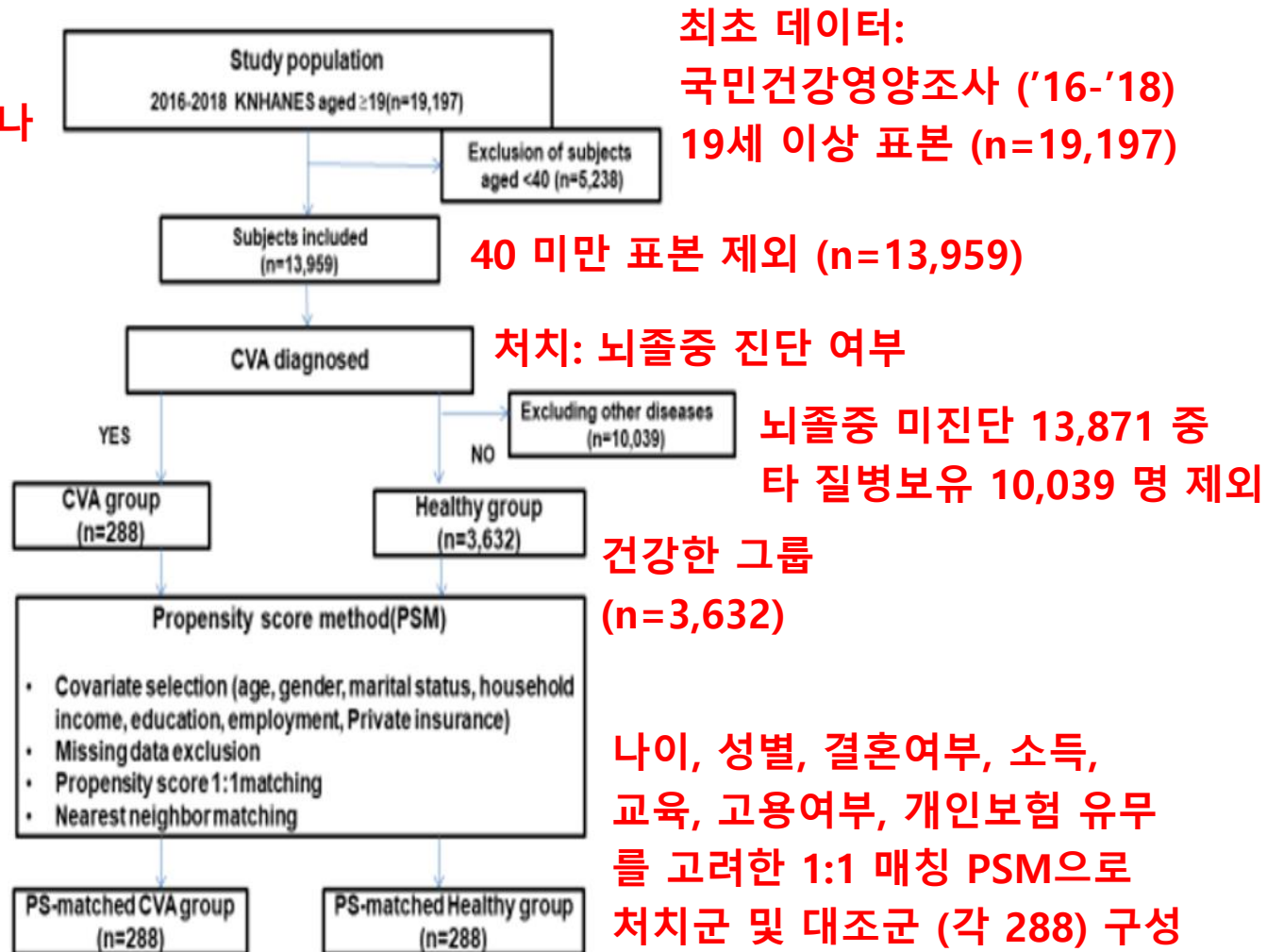
☆ Save    Cite    Related articles

처치: 동원훈련

## PSM을 활용한 표본 매칭 예시 (최민혁 & 최진혁, 2016)

의료 분야는 PSM이  
많이 활용되는 분야 중 하나

뇌졸중: 288



## PSM의 한계점 (및 이에 따른 주의사항)

- 데이터 손실 (data loss)
  - PSM은 표본들을 서로 매칭해주는 것이기 때문에, **매칭되지 못한 표본은 분석에서 제외되어** 표본 수가 감소할 수 있으며, 이에 따라 **결과 일반화에 제약이** 발생할 수 있음
- 관찰되지 못한 혼동 변수(unobserved confounding variable)
  - PSM에서는 고려된 혼동 변수들만 처치 여부에 영향을 미친다고 가정
  - **미처 고려되지 못한 혼동 변수로 인한 편향**을 완전히 배제할 수 없음
  - 잠재적 혼동 변수를 **최대한 식별**하여 반영하려 해야 함
- 불완전 매칭 (imperfect matching)
  - PSM을 통해 이루어진 매칭이 적절하지 않을 수 있음
  - **매칭이 적절히 이루어졌는지 추가적인 검증**을 반드시 거쳐야 함
- 적절히 설계된 무작위통제실험만큼의 인과관계를 확립하기는 어려움

# **Conducting PSM Analysis**

# PSM Analysis Using R

## PSM 분석 단계

- PSM 분석의 과정은 아래 3단계로 요약할 수 있음
  1. 성향점수 추정
  2. 매칭 및 매칭 결과 검증
  3. 인과성 추론

# PSM Analysis Using R

## R을 활용한 PSM 분석

- 본 수업에서는 통계 프로그램 **R**을 이용하여 주요 분석을 진행함
  - **R**에 대한 개략적인 설명 및 다운로드/설치 방법은 **본 자료 맨 뒤 부록 참조**





# PSM Analysis Using R

## R을 활용한 PSM 분석 – MatchIt 패키지

- R을 활용한 PSM 분석에는 주로 MatchIt 패키지가 활용됨

#MatchIt 패키지 설치

```
install.packages("MatchIt")
```

#MatchIt 패키지 불러오기

```
library(MatchIt)
```

#작업 경로 설정 - 실행 X

```
setwd("파일 경로")
```

#파일 불러오기 (csv파일) - 실행 X

```
mydata <- read.csv("파일명", fileEncoding="euc-kr")
```

#라이브러리 제공 기본 데이터(lalonde) 불러오기 및 해당 데이터를 mydata에 저장

```
data("lalonde")
```

```
mydata <- lalonde
```

# PSM Analysis Using R

## R을 활용한 PSM 분석 – 분석 데이터 설명

### ▪ lalonde 데이터 설명

- Dehejia & Wahba (1999)의 연구에서 사용한 데이터
  - ✓ Causal Effects in Nonexperimental Studies: **Reevaluating** the Evaluation of Training Programs
- 경제적 취약계층을 대상으로 이루어진 **미국 정부의 직업훈련 프로그램(NSW, national supported work)**이 해당 프로그램 참여자의 이후 소득에 미치는 영향을 분석하기 위한 데이터
  - ✓ 경제적 취약계층이 적절한 직업훈련을 받고 이를 바탕으로 안정적인 일자리를 얻어 자립할 수 있도록 지원하는 프로그램

- 분석 대상 인과 관계: **NSW 참여 → 높은 미래 소득**
  - 처치: **NSW 참여 여부**

# PSM Analysis Using R

## R을 활용한 PSM 분석 – 분석 변수

분석 대상:

**Treat → Re78**

### ■ 데이터셋 내 변수 목록

- **Treat**: NSW 참여자와 비참여자를 구분하는 더미 변수 (1: 참여)
- **Age**: 참여자의 나이
- **Educ**: 참여자의 교육 수준 (교육 년수)
- **Race**: 참여자의 인종 (흑인, 백인, 히스패닉)
- **Married**: 결혼 여부 (1: 결혼)
- **Nodegree**: 고등학교 졸업 여부 (1: 학위 없음)
- **Re74**: 프로그램 참여 전인 1974년 연 소득 (USD)
- **Re75**: 프로그램 참여 전인 1975년 연 소득 (USD)
- **Re78**: 프로그램 참여 후인 1978년 연 소득 (USD)

# PSM Analysis Using R

## R을 활용한 PSM 분석 – 데이터 둘러보기

#맨 위 몇 줄 보기

Head(mydata)

#자료 구조 보기

str(mydata)

#주요 통계량 보기

summary(mydata)

```
> head(mydata)
  treat age educ  race married nodegree re74 re75      re78
NSW1   1  37  11 black       1         1    0    0 9930.0460
NSW2   1  22   9 hispan      0         1    0    0 3595.8940
NSW3   1  30  12 black       0         0    0    0 24909.4500
NSW4   1  27  11 black       0         1    0    0  7506.1460
NSW5   1  33   8 black       0         1    0    0  289.7899
NSW6   1  22   9 black       0         1    0    0 4056.4940
```

```
> str(mydata)
'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 ...
```

```
> summary(mydata)
   treat      age      educ      race      married
Min.   :0.0000 Min.   :16.00 Min.   : 0.00 black :243 Min.   :0.0000
1st Qu.:0.0000 1st Qu.:20.00 1st Qu.: 9.00 hispan: 72 1st Qu.:0.0000
Median :0.0000 Median :25.00 Median :11.00 white :299 Median :0.0000
Mean    :0.3013 Mean    :27.36 Mean    :10.27          Mean    :0.4153
3rd Qu.:1.0000 3rd Qu.:32.00 3rd Qu.:12.00          3rd Qu.:1.0000
Max.    :1.0000 Max.    :55.00 Max.    :18.00          Max.    :1.0000

  nodegree      re74      re75      re78
Min.   :0.0000 Min.   :  0 Min.   : 0.0 Min.   : 0.0
1st Qu.:0.0000 1st Qu.:  0 1st Qu.: 0.0 1st Qu.: 238.3
Median :1.0000 Median :1042 Median : 601.5 Median : 4759.0
Mean    :0.6303 Mean    : 4558 Mean    : 2184.9 Mean    : 6792.8
3rd Qu.:1.0000 3rd Qu.: 7888 3rd Qu.: 3249.0 3rd Qu.:10893.6
Max.    :1.0000 Max.    :35040 Max.    :25142.2 Max.    :60307.9
```

# PSM Analysis Using R

## 1. 성향점수 추정

- 성향점수를 어떻게(어떠한 모형으로) 추정할 지를 결정하는 단계
  - 성향점수: 어떠한 처치를 받을 확률을 나타내는 지표
- 성향점수 추정 모형에 어떤 공변량(covariate)을 포함할 지를 결정해야 함
  - 공변량: 분석에서 결과(종속) 변수에 영향을 줄 수 있는, 관심 있는 주요 독립 변수 외 다른 변수
  - 즉, 혼동 변수 중 적절히 관측/측정되고 모형에 반영된 것을 공변량이라고 볼 수 있음

# PSM Analysis Using R

## 1. 성향점수 추정 – 로짓 모형 (1)

- 성향점수 추정 모형으로는 주로 로짓 모형(logit model)이 활용됨
  - 로짓 모형은 종속변수(Y)가 0 혹은 1인 경우 유용하게 활용될 수 있는 회귀모형
    - ✓ 로짓 모형은 로지스틱 회귀 모형이라고도 함 (혼용)
  - 처치 여부 역시 0 혹은 1의 더미 변수 형태로 표현될 수 있음

$$\begin{array}{l} \text{처치 여부} = \text{공변량1} + \text{공변량2} + \dots \\ \text{(종속변수)} \qquad \qquad \qquad \text{(설명변수들)} \end{array}$$

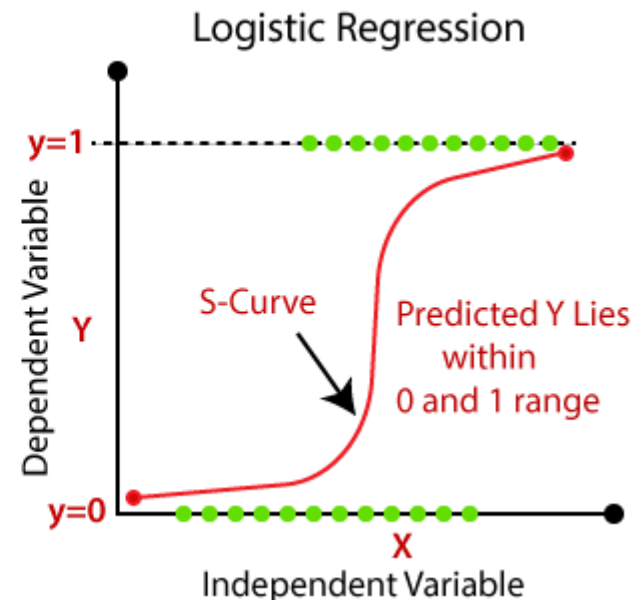
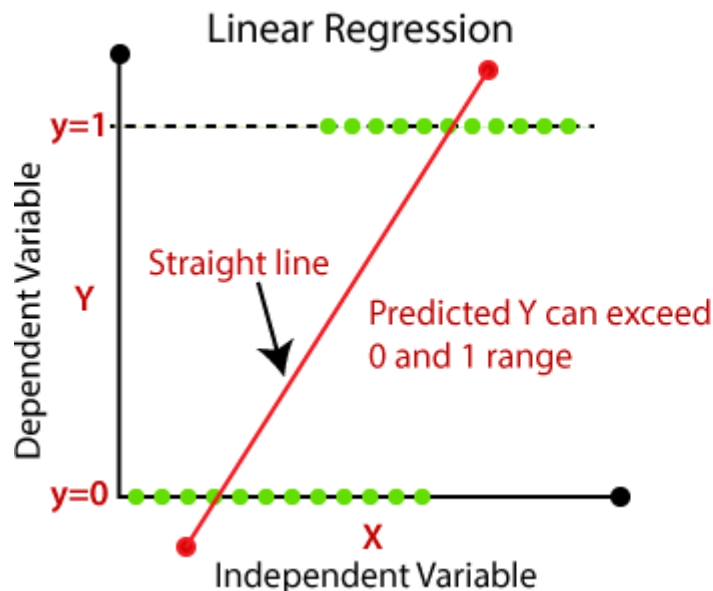
- 다른 모형(프로빗 등) 역시 활용될 수 있으나, 본 수업에서는 로짓 모형에 집중 (가장 널리 활용됨)

# PSM Analysis Using R

## 1. 성향점수 추정 - 로짓 모형 (2)

▪ Q. 그냥 선형회귀모형 쓰면 안되나요???

- A. 종속변수 Y가 0 혹은 1인 경우 적합도 측면에서 선형 함수보다 로지스틱 함수의 활용이 더 적절함



# PSM Analysis Using R

## 1. 성향점수 추정 – 어떤 모형이 좋을까?

- PSM의 주요 가정 중 하나는 처치 여부가 관측된 공변량에만 의존한다는 것
  - 즉, 관측된 공변량의 값이 모두 동일하다면, 처치 받을 확률 역시 동일하다고 가정
  - 해당 가정을 숙지하고 모형에 포함할 공변량을 정해야 함
- 앞서 살펴본 lalonde 데이터에서 공변량으로 모형에 포함할 변수는 무엇이 적절할까?
  - 나이, 교육 수준, 인종, 결혼 여부, 고등학교 졸업 여부, 74년 소득, 75년 소득
  - NSW 참여 여부 및 78년 소득은 인과성 분석 대상이므로 공변량으로 반영될 수 없음
    - ✓ 이 중 NSW 참여 여부(처치)는 성향점수 추정 모형의 종속 변수임



# PSM Analysis Using R

## PSM의 기본 가정

- **처치 여부가 관측된 공변량에만 의존함**
  - 두 관측치의 공변량이 동일하다면, 처치 확률 역시 동일함
  - 관측되지 않은 혼동 변수가 없는지 주의해야 함
- **처치군과 통제군의 성향 점수가 비슷한 범위 내에 존재함**
  - 모든 표본이 어느 정도는 처치를 받을 가능성이 있어야 함
    - ✓ 어떤 표본이 처치를 받을 가능성이 전혀 없거나 매우 확실한 경우, 이들은 매칭 대상에서 제외되는 것이 바람직함
    - ✓ 본 자료 초반에 잠시 살펴본 뇌졸중 연구에서 40세 미만 표본을 제외한 이유는?
  - 처치군과 통제군의 성향점수 범위를 검토해야 함

# PSM Analysis Using R

## 1. 성향점수 추정 - 실습

`matchit()` 함수는 성향점수 추정과 함께 매칭까지 한번에 하는 함수

#`matchit` 함수 활용    종속변수 ~ 설명변수1 + 설명변수2 + ...

설명서 보기: 콘솔에 “? `matchit`” 입력  
혹은 Google에 MatchIt 설명서 검색

```
mymodel <- matchit(treat ~ age + educ + married + re74,  
  data = mydata,
```

```
  distance = "glm",  
  method = "nearest")
```

`distance` 및 `method`는 매칭 방법 관련 (후술)

#logit 모형 추정 결과 불러오기 (psm 결과가 저장된 `mymodel`이라는 변수 내 `model` 부분)

```
summary(mymodel$model)
```

그냥 `summary(mymodel)` 하면 다른 것들이 나오고 아래는 안 나온다 (왜?)

```
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) -6.849e-01  4.905e-01  -1.396    0.163  
age          1.243e-02  1.053e-02   1.181    0.238  
educ         2.209e-02  3.736e-02   0.591    0.554  
married      -1.221e+00  2.365e-01  -5.161 2.45e-07 ***  
re74         -8.700e-05  2.214e-05  -3.929 8.51e-05 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

이 부분이 중요 하다가 보다는 이를 바탕으로 한 매칭 결과가 중요함

따라서 기본적으로는 일반 선형회귀모형과 마찬가지로 설명변수간 상관관계를 최소화하는 것이 바람직하겠지만, PSM에서는 매칭 결과에 주요 관심이 있기에 어느 정도 더 용인

# PSM Analysis Using R

## 로짓 모형 분석

**matchit()** 함수에서는 로짓 모형 분석과 매칭을 한번에 시행하지만,  
별도로 로짓 모형 분석을 하고 싶으면 **glm()** 함수를 활용할 수 있음

로짓모형임을 설정  
(종속변수가 이항)

#glm을 이용한 logit model 추정

```
logitmodel <- glm(treat ~ age + married + nodegree + re75, data = mydata, family = binomial)
```

종속변수 ~ 설명변수1 + 설명변수2 + ...

#추정 결과 요약

```
summary(logitmodel)
```

동일 결과 도출!

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.849e-01  4.905e-01  -1.396    0.163
age          1.243e-02  1.053e-02   1.181    0.238
educ         2.209e-02  3.736e-02   0.591    0.554
married      -1.221e+00  2.365e-01  -5.161 2.45e-07 ***
re74         -8.700e-05  2.214e-05  -3.929 8.51e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

종속변수가 0 혹은 1인 사례에서 예측 등을 하고 싶을 때 유용하게 활용 가능  
(이 경우 변수간 상관관계 등 검증 중요)

# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증

- 적절한 성향점수 추정 모형을 통해 성향점수를 추정했다면, 이제는 **추정된 성향점수를 바탕으로 어떻게 매칭을 할 것인지를 정해야 함**
  - 본 수업은 가장 널리 활용되는 1:1 매칭에 집중
- 크게 두 가지 기준을 설정해야 함
  - 성향점수를 구체적으로 어떻게 측정할 것인가? (distance)
  - 해당 거리를 바탕으로 어떻게 매칭할 것인가? (method)

#matchit 함수 활용

```
mymodel <- matchit(treat ~ age + married + nodegree + re75,  
  data = mydata,  
  distance = "glm",  
  method = "nearest")
```

# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증 - distance

### ▪ Distance 관련 주요 옵션

특정 방식이 무조건 더 좋고 그런 것은 없음  
세부 내용은 ? [matchit](#) 및 설명서 참조

- **glm: 로지스틱 회귀 활용 (기본)**
- 기타 다양한 회귀방법 활용 가능: lasso, ridge, elasticnet
- 머신러닝 방법론도 활용 가능: randomforest

다른 방식 활용 시 추가 패키지 설치를 요구하는 경우가 있는데, 지시에 따라 설치하면 됨

### ▪ glm이 아닌 다른 옵션을 distance에 사용할 경우:

- `summary(mymodel$model)` 입력 시 로지스틱 회귀 모형 결과가 아니라 사용한 방법에 맞는 모델 관련 내용이 출력됨

### ▪ 본 수업은 로지스틱 회귀 분석 결과를 활용하는 전통적인 방법에 초점을 맞추고 진행

# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증 - method

특정 방식이 무조건 더 좋고 그런 것은 없음  
세부 내용은 ? `matchit` 및 설명서 참조

### ▪ Method 관련 주요 옵션

- `nearest`: 익숙한 nearest neighbor (NN) 방식
- `optimal`: 전체 최적 매칭 방식
- 기타 다양한 매칭 방식도 존재함

다른 방식 활용 시 추가 패키지 설치를 요구하는 경우가 있는데, 지시에 따라 설치하면 됨

#`matchit` 함수 활용

```
mymodel <- matchit(treat ~ age + married + nodegree + re75,
```

```
  data = mydata,
```

```
  distance = "glm",  
  method = "nearest")
```

즉, 해당 옵션은

로지스틱 회귀 결과를 바탕으로 성향점수를 추정하고  
Nearest neighbor 방식으로 관측치를 매칭하겠다는 뜻

# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증 – 매칭 방법 비교

### ▪ Nearest neighbor vs. optimal 비교 예시

처치군	성향점수	대조군	성향점수
A	0.67	E	0.66
B	0.43	F	0.79
C	0.31	G	0.51
D	0.19	H	0.51
		K	0.42

**Nearest neighbor: A-E / B-K / C-G / D-H**

**Optimal: A-E / B-G / C-H / D-K**

Optimal 방식에서는 D가 그나마 가까운 K와 연결되는 방식을 선택할 수 있음

NN에서는 매칭 순서에 따라 후순위(C, D)는 성향점수 차이가 큰 대조군과 매칭될 수 있음

**Optimal 방식이 무조건 더 좋다고 볼 수는 없음**

→ 여러 방식 시도 및 비교를 통해 최적 모형을 정하자

# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증 – 무엇을 검증하나?

- 어떠한 방식으로 매칭할지를 결정했다면, 이제는 **매칭 결과를 검증**해볼 차례
- PSM의 목적은 **처치군에 맞는 적절한 대조군을 매칭**하는 것
  - 바람직한 처치군 대조군의 특징은 **처치를 제외하고는 동일(유사)한 특성을** 가지는 것
- 즉, **처치군과 매칭된 대조군 사이에 유의미한 차이가 존재하는지**를 분석해야 함
  - 모형에 반영된 공변량(혼동변수) 기준으로 검증 (SMD)
- 또한, PSM의 주요 가정 중 하나인 **“처치군과 대조군의 성향점수가 비슷한 범위 내에 존재한다”**는 것을 검증하기 위해, **처치군과 대조군의 성향점수 분포** 역시 비교해야 함
  - 시각화 도구 활용 가능



# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증 – summary 결과 해석

#이전에 추정한 mymodel의 summary 출력  
summary(mymodel)

distance 차이로부터 알 수 있는 것은?

무작위통제실험이었다면 distance 값이 어떻게 나타날까?

```
Call:
matchit(formula = treat ~ age + educ + married + re74, data = mydata,
        method = "nearest", distance = "glm")
```

distance는 성향점수

Summary of Balance for All Data: **처치군과 모든 대조군 혼동변수 평균값 비교**

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max
distance	0.3832	0.2660	0.9447	0.6451	0.2351	0.3927
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
married	0.1892	0.5128	-0.8263	.	0.3236	0.3236
re74	2095.5737	5619.2365	-0.7211	0.5181	0.2248	0.4470

Summary of Balance for Matched Data: **처치군과 "매칭된" 대조군 혼동변수 평균값 비교**

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
distance	0.3832	0.3739	0.0752	1.1053	0.0301	0.2162	0.0790
age	25.8162	24.8919	0.1292	0.5248	0.0718	0.2486	0.9511
educ	10.3459	10.4162	-0.0349	0.5106	0.0316	0.0865	1.2017
married	0.1892	0.2216	-0.0828	.	0.0324	0.0324	0.1104
re74	2095.5737	1916.6183	0.0366	1.5566	0.0304	0.2162	0.2421

Sample Sizes:

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

185개 처치군에 맞는 185개 대조군이 생성  
244개는 짝을 찾지 못하고...

# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증 - SMD (1)

#이전에 추정한 mymodel의 summary 출력  
summary(mymodel)

$$SMD = \frac{\bar{X}_t - \bar{X}_c}{SD_{pooled}}$$

처치군 평균      대조군 평균  
표준화 평균차이      처치군 + 대조군 표준편차

```
Call:
matchit(formula = treat ~ age + educ + married + re74, data = mydata,
        method = "nearest", distance = "glm")
```

Summary of Balance for All Data:

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max
distance	0.3832	0.2660	0.9447	0.6451	0.2351	0.3927
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
married	0.1892	0.5128	-0.8263	.	0.3236	0.3236
re74	2095.5737	5619.2365	-0.7211	0.5181	0.2248	0.4470

가장 중요 (Standardized Mean Difference, SMD)

Summary of Balance for Matched Data:

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
distance	0.3832	0.3739	0.0752	1.1053	0.0301	0.2162	0.0790
age	25.8162	24.8919	0.1292	0.5248	0.0718	0.2486	0.9511
educ	10.3459	10.4162	-0.0349	0.5106	0.0316	0.0865	1.2017
married	0.1892	0.2216	-0.0828	.	0.0324	0.0324	0.1104
re74	2095.5737	1916.6183	0.0366	1.5566	0.0304	0.2162	0.2421

Sample Sizes:

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

평균 차이가 유의미한 지 나타내는 지표 (0.1보다 작은 것이 바람직)  
절대값이 0.1보다 작다면 차이가 유의미하지 않다고 판단  
(절대값이 작을수록 차이가 유의미하지 않다고 봄)

# PSM Analysis Using R

매칭 검증에 t-검정/카이제곱검정을 쓰면 안된다는 이야기가 아님!

(R 내 타 기능으로 별도 검정 수행 가능 - p.48 참조)

단, t-검정과 카이제곱 검정의 경우 표본 크기가 매우 큰 경우

아주 작은 차이도 유의미하다는 결과를 도출할 수 있음

## 2. 매칭 및 매칭 결과 검증 – SMD (2)

$$SMD = \frac{\bar{X}_t - \bar{X}_c}{SD_{pooled}}$$

표준화 평균차이

대조군 평균

SMD 주요 장점

1) 표본 수 영향 X

2) 연속/범주형 변수 무관 간편 활용

두 그룹 공변량에 대해 풀링된(pooled) 표준 편차

cf)

$$t = \frac{\bar{X}_t - \bar{X}_c}{SE_{\bar{X}_t - \bar{X}_c}}$$

처치군과 대조군 평균 차이의 표준오차

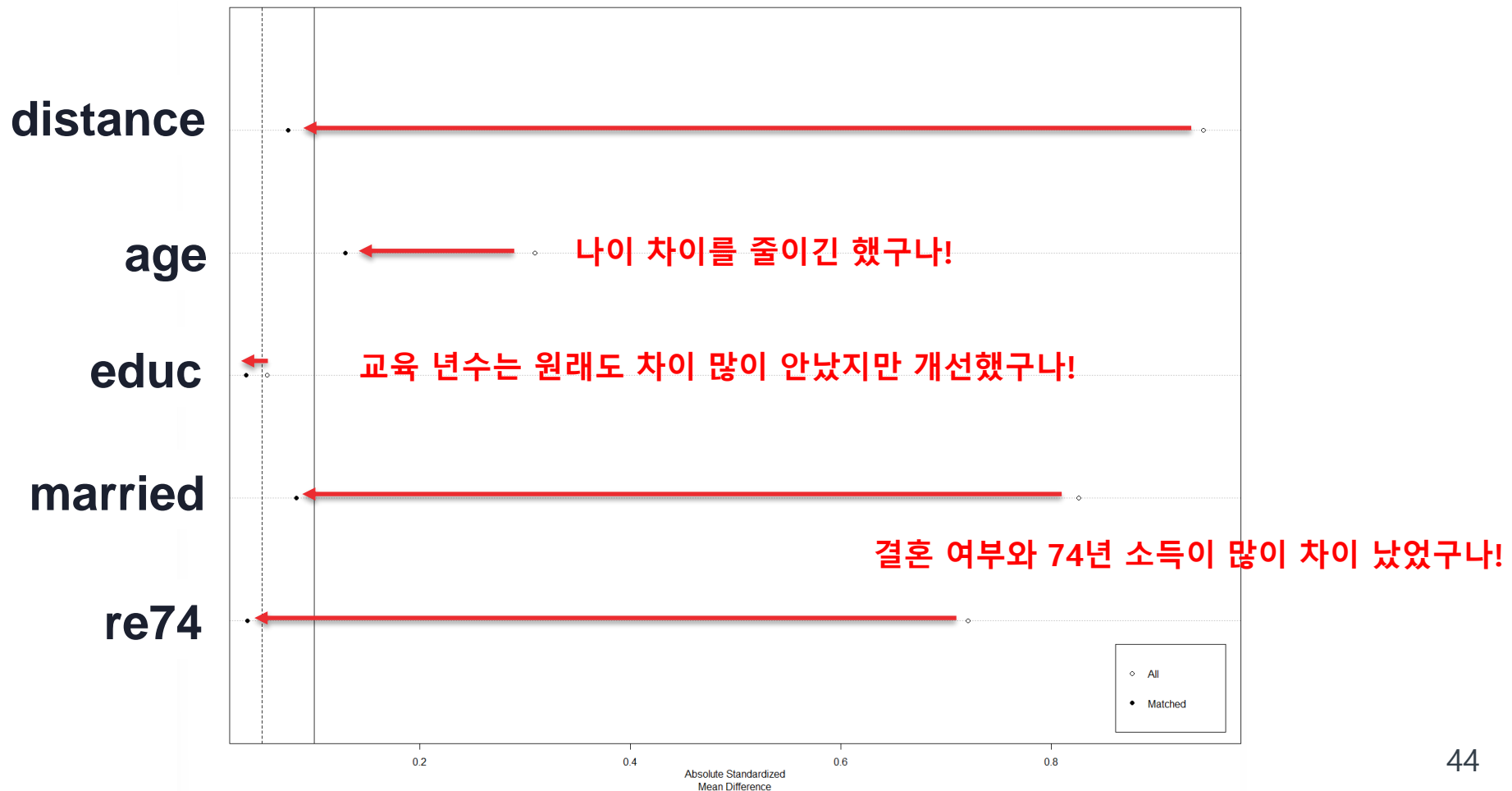
- ✓ SMD 값이 0에 가깝다 = 두 그룹 간의 공변량이 균형을 이루었다  
(통상적으로 0.1 이하) (무시할 수 있을 정도로 작다)
- ✓ SMD 값이 0에서 멀다 = 두 그룹 간의 공변량이 유의미하게 크다  
(다른 추정 모형 혹은 매칭 방식 등 통계적 조정 필요)
- ✓ SMD는 처치군과 대조군의 공변량 차이의 크기를 표준화하여 비교하는 데 초점을 두는 지표  
(t-검정과는 달리, 통계적 유의성에 관심이 있는 것이 아님)
- ✓ 추후 매칭된 처치군과 대조군을 바탕으로 두 그룹 사이에 인과성을 분석할 때에는 t-검정 등을 활용함 (통계적 유의성에 관심)

# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증 – 매칭에 따른 SMD 개선 시각화

#전체 대조군과 매칭된 대조군의 SMD 값을 시각화하여 비교해보자

`plot(summary(mymodel))`



# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증 – summary 결과 해석 (기타)

```
Call:
matchit(formula = treat ~ age + educ + married + re74, data = mydata,
  method = "nearest", distance = "glm")

Summary of Balance for All Data:
      Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
distance      0.3832      0.2660      0.9447      0.6451      0.2351      0.3927
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
married        0.1892      0.5128     -0.8263      0.0000      0.3236      0.3236
re74        2095.5737    5619.2365     -0.7211      0.5181      0.2248      0.4470

Summary of Balance for Matched Data:
      Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
distance      0.3832      0.3739      0.0752      1.1053      0.0301      0.2162      0.0790
age          25.8162     24.8919      0.1292      0.5248      0.0718      0.2486      0.9511
educ         10.3459     10.4162     -0.0349      0.5106      0.0316      0.0865      1.2017
married        0.1892      0.2216     -0.0828      0.0000      0.0324      0.0324      0.1104
re74        2095.5737    1916.6183      0.0366      1.5566      0.0304      0.2162      0.2421

Sample Sizes:
      Control Treated
All          429     185
Matched      185     185
Unmatched    244      0
Discarded     0      0
```

**VAR Ratio:** 처리군과 대조군 간의 공변량 분산 비율

(1에 가까울 수록 좋으며, 0.5~2 사이면 적절하다고 판단)

**eCDF Mean:** 매치된 표본에서 공변량의 누적 분포함수(eCDF) 평균 차이

(작을수록 좋음 – 공변량 분포가 비슷하다는 의미)

**eCDF Max:** 매치된 표본에서 eCDF 차이 최대값

(작을수록 좋음 – 어떤 구간에서도 차이가 작다는 의미)

**Std. Pair Dist.:** 표준화된 매칭 쌍 간 거리 평균

(낮을수록 좋음 – 더 유사한 표본끼리 매칭)

# PSM Analysis Using R

## 2. 매칭 및 매칭 결과 검증 - 매칭 분포 시각화

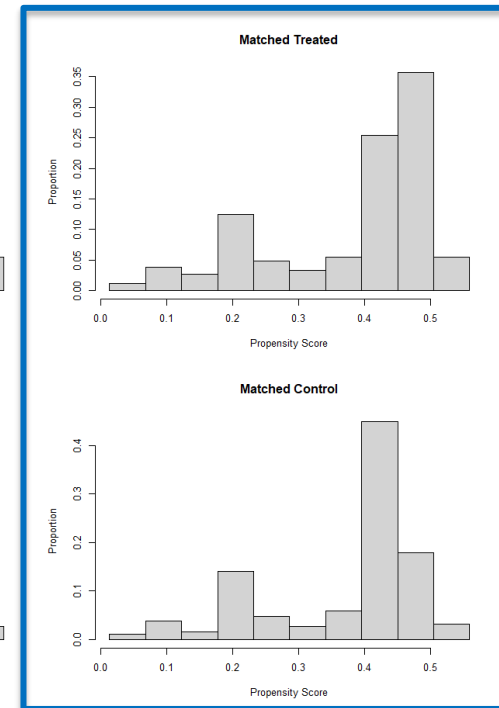
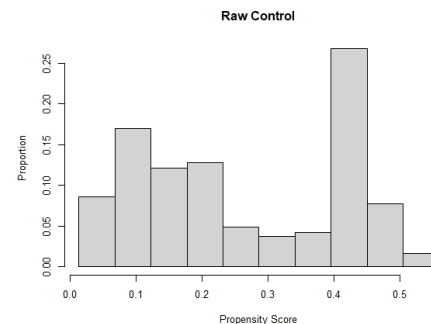
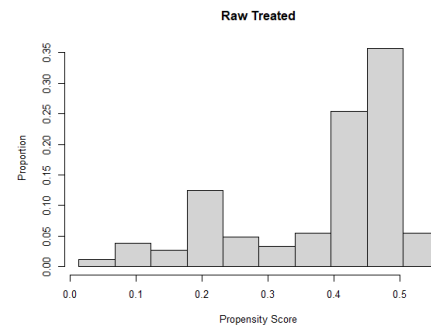
#성향점수 분포를 jitter된 시각화 자료로 살펴보자

```
plot(mymodel, type = "jitter")
```

jitter의 경우 명령어 시행하면 잠시 대기상태가 되는데, 이름을 나타내고 싶은 관측치(들) 클릭 후 ESC를 누르면 해당 관측치명이 표기됨

#성향점수 분포를 히스토그램으로 나타내어보자

```
plot(mymodel, type = "hist")
```



# PSM Analysis Using R

## 3. 인과성 추론

- 매칭이 적절히 (최대한) 이루어졌다고 판단된다면, 이제 구성된 처치군과 대조군을 이용해 인과성 추론을 해볼 차례
- 처치군과 대조군 사이에 인과 추론의 대상 현상이 얼마나 차이나는지 비교
  - 예시 데이터 기준: 처치군(NSW 이수)과 대조군(NSW 미이수) 사이에 NSW 프로그램 이후인 1978년 소득이 얼마나 차이가 나는지?

# PSM Analysis Using R

## 3. 인과성 추론 – 매칭 데이터 활용

#매칭된 데이터 분리

```
matched_data <- match.data(mymodel)
```

#매칭된 데이터 대상 t-검정 수행

```
with(matched_data, t.test(re78 ~ treat))
```

참고1) 매칭된 데이터 t-검정 방법

- `with(matched_data, t.test(age ~ treat))`
- `with(matched_data, t.test(educ ~ treat))`
- `with(matched_data, t.test(re74 ~ treat))`

참고2) 매칭 데이터 카이제곱검정 방법

- `with(matched_data, chisq.test(table(matched_data$treat, matched_data$married)))`

처치군과 대조군의 78년 소득(re78) 비교

```
Welch Two Sample t-test

data: re78 by treat
t = -1.7292, df = 333.67, p-value = 0.08471
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -2630.9570  169.3691
sample estimates:
mean in group 0 mean in group 1
    5118.350      6349.144
```

t=1.73, p-value=0.08

10% 신뢰수준에서 유의한 차이

대조군

처치군

- ✓ 즉, NSW 프로그램 이수는 1231 달러 (평균치 대비 21%) 만큼의 미래 소득 개선 효과가 있었다고 주장할 수 있음  
(통계적으로 유의미한 차이이며, 우연한 차이가 아님)



# PSM Analysis Using R

## 3. 인과성 추론 – 추가적 고려 사항

- 좀 더 엄밀하게 주장을 하기 위해서는...
- **매칭된 표본의 특성을 살펴보아야 함**
  - 제외된 표본에 대한 인과성 추론은 어려우므로 분석에 포함된 표본 범위 내에서 해석과 주장이 이루어져야 함
  - **분석 범위의 명확한 정의 필요**
- **매칭에도 불구하고 처치군과 대조군 사이에 여전히 공변량 차이가 존재한다면, 해당 공변량 차이가 인과성 추론에 미치는 영향을 평가해야 함**
  - 앞선 분석의 경우, 처치군과 대조군이 **나이** 부분에서 차이가 약간 있는 것으로 나타났으므로 이를 **인지/유의**해야 함
    - ✓ 처치군이 약간 더 어림
    - ✓ 해당 차이가 인과성 추론에 문제가 될까?  
(나이가 좀 더 어린 것이 결과에 큰 영향을 미칠까?)

# PSM Analysis Using R

## 3. 인과성 추론 – 데이터 둘러보기 및 csv 파일 작성

#매칭된 데이터 살펴보기

head(matched\_data)

```
> head(matched_data)
  treat age educ race married nodegree re74 re75 re78 distance weights subclass
NSW1   1  37  11 black      1         1    0    0 9930.0460 0.2309803         1         1
NSW2   1  22   9 hispan     0         1    0    0 3595.8940 0.4470411         1        98
NSW3   1  30  12 black     0         0    0    0 24909.4500 0.4882687         1       109
NSW4   1  27  11 black     0         1    0    0  7506.1460 0.4734475         1       120
NSW5   1  33   8 black     0         1    0    0  289.7899 0.4755216         1       131
NSW6   1  22   9 black     0         1    0    0 4056.4940 0.4470411         1       142
```

#매칭된 데이터 기초 통계량 검토

summary(matched\_data)

distance: 성향점수

weights: 가중치 (1:1 매칭은 항상 1)

subclass: 같은 번호인 경우 매칭된 짝

#매칭된 데이터를 csv 파일로 빼내기

write.csv(matched\_data, file = "matched\_data\_lalonde.csv")

빼내서 다양한 추가분석 가능

```
> summary(matched_data)
   treat      age      educ      race      married
Min.   :0.0   Min.  :16.00   Min.   : 1.00   black :207   Min.   :0.0000
1st Qu.:0.0   1st Qu.:19.00   1st Qu.: 9.00   hispan: 42   1st Qu.:0.0000
Median :0.5   Median :23.00   Median :11.00   white :121   Median :0.0000
Mean    :0.5   Mean    :25.35   Mean    :10.38                Mean    :0.2054
3rd Qu.:1.0   3rd Qu.:29.00   3rd Qu.:12.00                3rd Qu.:0.0000
Max.    :1.0   Max.    :55.00   Max.    :18.00                Max.    :1.0000

 nodegree      re74      re75      re78      distance
Min.   :0.0000   Min.   : 0   Min.   : 0   Min.   : 0.00   Min.   :0.01294
1st Qu.:0.0000   1st Qu.: 0   1st Qu.: 0   1st Qu.: 39.16   1st Qu.:0.28471
Median :1.0000   Median : 0   Median : 0   Median : 3897.50 Median :0.43258
Mean    :0.6649   Mean    :2006 Mean    :1386 Mean    : 5733.75 Mean    :0.37857
3rd Qu.:1.0000   3rd Qu.:1496 3rd Qu.:1531 3rd Qu.: 8797.93 3rd Qu.:0.45980
Max.    :1.0000   Max.   :35040 Max.   :25142 Max.   :60307.93 Max.   :0.54680
```

참고) dplyr 패키지 활용

install.packages("dplyr")

library(dplyr)

summary(filter(matched\_data, treat == 1))

# **Quick Review with** **Another Dataset**

# Quick Review

앞서 배운 내용을 다른 데이터를 이용하여 빠르게 복습하고 연습해보자

## ■ 분석 대상 데이터

- 국내 소비자 약 1105명 대상 설문조사 데이터 (가상의 데이터)

## ■ 분석 대상 인과 관계

(있다고 가정)

- 전기요금 실시간 조회 어플리케이션 활용

→ 시간별 차등 요금제 참여 의향 개선 (1: 의향 있음 ~ 0: 의향 없음)  
(이용 시간대별로 전기 요금 단가가 달라지는 요금제도)

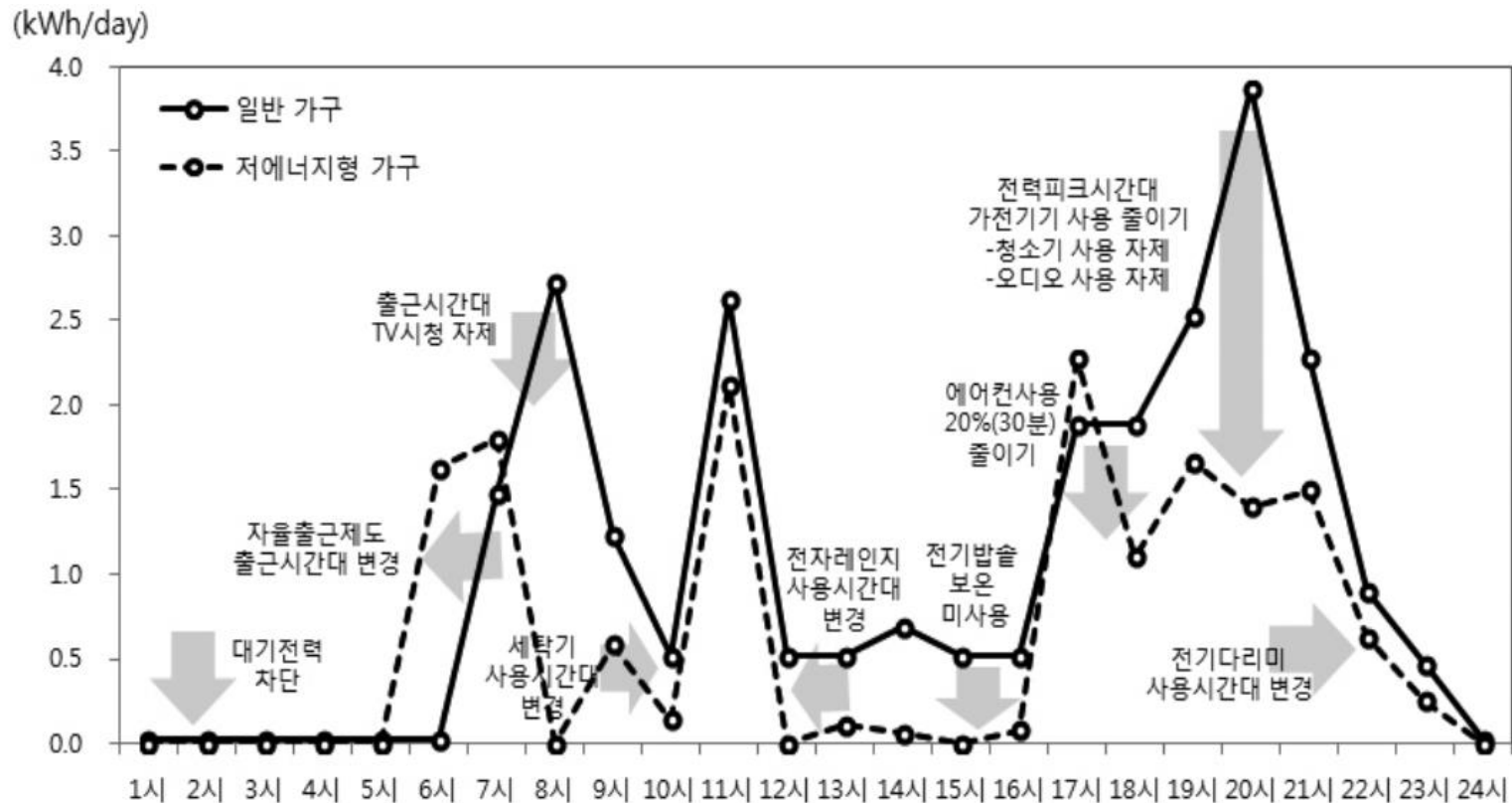
## ■ 문제 상황

인과성이 있다면 해당 어플 사용을 유도하여  
시간별 차등 요금제 참여 의향을 개선시키고자 함

- 표본의 대표성이 없음
- 전기요금 실시간 조회 어플리케이션을 전체 표본 중 약 26%만 활용

# Quick Review

## 번외) 시간대별 전력 사용량과 가능한 에너지 절약 행위



〈그림 7〉 에너지 절약 실천 가구의 전력 피크 저감효과

# Quick Review

## 변수 목록

- treat: 실시간 조회 어플리케이션 활용 여부 (1: 활용, 0: 미활용)
- tou\_intent: 시간별 차등 요금제 참여 의향 (1: 있음, 0: 없음)
- gender: 성별 (남성: 1)
- age: 나이 (세)
- region: 지역 (후술)
- edu: 교육 연수 (후술)
- h\_income: 가구소득 (백만원/월)
- i\_income: 개인소득 (백만원/월)
- h\_mem: 가구원 수
- elec\_bill: 월 평균 전기요금 (만원/월)
- bill\_satis: 전기요금 만족도 (1: 매우 불만족 ~ 5: 매우 만족)
- politic: 정치 성향 (1: 매우 진보적 ~ 5: 매우 보수적)
- support: 정부 정책 지지도 (1: 매우 지지하지 않음 ~ 5: 매우 지지함)

# Quick Review

## 변수 목록 – 추가 설명 (1)

### ■ 거주지역

번호	지역	번호	지역
1	서울	10	강원도
2	부산	11	충청북도
3	대구	12	충청남도
4	인천	13	전라북도
5	광주	14	전라남도
6	대전	15	경상북도
7	울산	16	경상남도
8	세종	17	제주도
9	경기도		

# Quick Review

## 변수 목록 – 추가 설명 (2)

- 교육 연수

무학	초등학교	중학교	고등학교	대학교	대학원
0	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15 16	17 18 19 20 21 22



# Quick Review

## 0. 자료 불러오기

#데이터 불러오기    **설정된 경로에 파일이 있어야 함**

```
mydata2 <- read.csv("BEDA_data1_rev1.csv")
```

#데이터 둘러보기

```
summary(mydata2)
```

```
> summary(mydata2)
```

pid	treat	tou_intent	gender	age	region
Min. : 1	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :20.00	Min. : 1.000
1st Qu.: 277	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:34.00	1st Qu.: 1.000
Median : 553	Median :0.0000	Median :0.0000	Median :1.0000	Median :46.00	Median : 6.000
Mean : 553	Mean :0.2624	Mean :0.1846	Mean :0.5068	Mean :45.56	Mean : 5.758
3rd Qu.: 829	3rd Qu.:1.0000	3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:56.00	3rd Qu.: 8.000
Max. :1105	Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :69.00	Max. :17.000

edu	h_income	i_income	h_mem	elec_bill	bill_satis
Min. : 0.00	Min. : 0.000	Min. : 0.000	Min. :1.000	Min. : 1.000	Min. :1.000
1st Qu.:14.00	1st Qu.: 6.000	1st Qu.: 4.000	1st Qu.:2.000	1st Qu.: 3.000	1st Qu.:2.000
Median :16.00	Median : 7.000	Median : 5.000	Median :3.000	Median : 5.000	Median :3.000
Mean :15.24	Mean : 6.871	Mean : 4.984	Mean :2.691	Mean : 5.478	Mean :2.728
3rd Qu.:16.00	3rd Qu.: 9.000	3rd Qu.: 7.000	3rd Qu.:4.000	3rd Qu.: 7.000	3rd Qu.:3.000
Max. :22.00	Max. :10.000	Max. :10.000	Max. :9.000	Max. :16.000	Max. :5.000

politic	support
Min. :1.000	Min. :1.000
1st Qu.:3.000	1st Qu.:1.000
Median :3.000	Median :2.000
Mean :2.949	Mean :2.321
3rd Qu.:3.000	3rd Qu.:3.000
Max. :5.000	Max. :5.000

# Quick Review

## 성향점수 추정 & 매칭 및 매칭 결과 검증

성별, 나이, 가구소득, 가구원 수, 전기요금 수준을 공변량으로 포함

```
mymodel2 <- matchit (treat ~ gender + age + h_income + h_mem + elec_bill,  
  data = mydata2,  
  distance = "glm",  
  method = "nearest")
```

 로짓 모형 + nearest neighbor

```
summary(mymodel2)
```

```
Summary of Balance for Matched Data:
```

	Means Treated	Means Control	Std. Mean Diff.
distance	0.2905	0.2903	0.0026
gender	0.6379	0.6207	0.0359
age	44.7379	43.7345	0.0772
h_income	7.2483	7.2586	-0.0048
h_mem	2.7690	2.8138	-0.0363
elec_bill	5.5966	5.4655	0.0402

# Quick Review

## 매칭 결과 검증 (시각화)

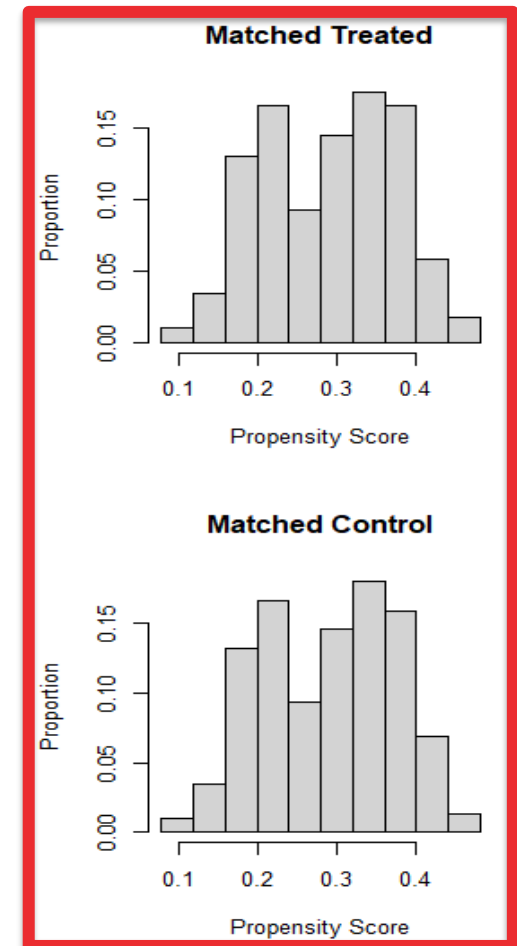
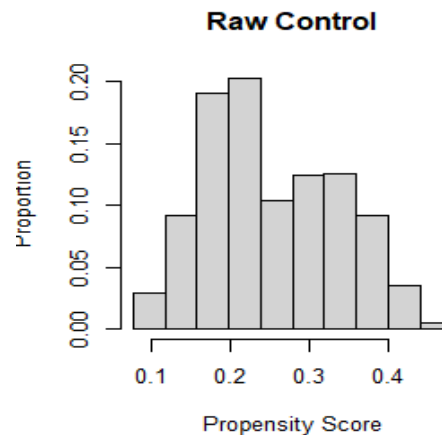
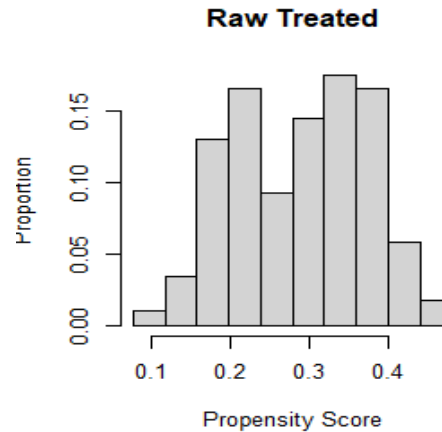
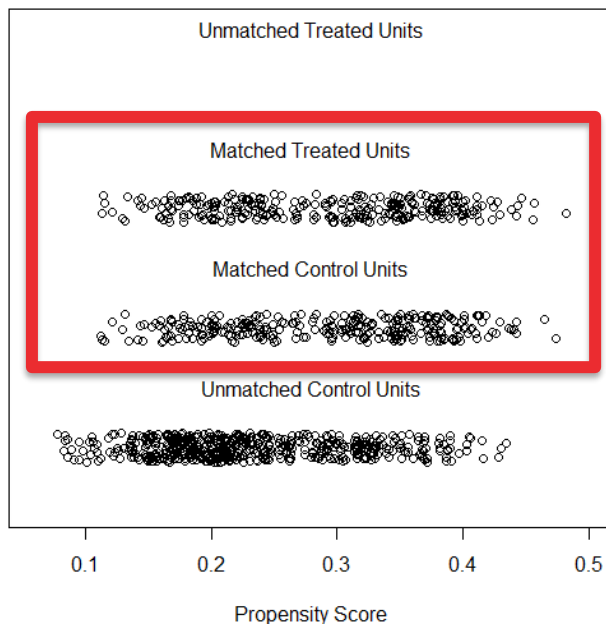
#주요 시각화 분석 시행

```
plot(summary(mymodel2))
```

```
plot(mymodel2, type = "jitter")
```

```
plot(mymodel2, type = "hist")
```

Distribution of Propensity Scores



# Quick Review

## 인과성 추론

#매칭 데이터 구성

```
matched_data2 <- match.data(mymodel2)
```

	0	1
0	257	33
1	213	77

#카이제곱 검정을 위한 테이블 구성

검증 대상이 범주형 더미 변수이므로 카이제곱검정 수행

```
table_for_test <- table(matched_data2$treat, matched_data2$tou_intent)
```

#카이제곱 검정

```
chisq.test(table_for_test)
```

```
Pearson's Chi-squared test with Yates' continuity correction  
data: table_for_test  
X-squared = 20.743, df = 1, p-value = 5.252e-06
```

p-value가 0.1보다 작으므로 (심지어 매우 작음)  
treat에 따라 tou\_intent의 유의미한 차이가 존재함  
(수치를 별도로 계산해보면 11.4% vs 26.5%)

# **Comparing PSM with Other Methods**

# Multiple Linear Regression Model

- Recall) 다중선형회귀모형
- 구축된 모형을 통해 예측 등 추가분석 가능
  - 선형성, 다중공선성, 오차관련 주요 가정 등을 만족시켜야 함

## ▪ 다중(multiple) 선형회귀모형

- 하나의 종속변수(Y)가 여러 개(하나 이상)의 설명변수(X)와 어떠한 관계를 가지는가?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

- $\beta_0$ :  $X_1$ 과  $X_2$ 가 모두 0일 때의 Y의 값 (해석적 의미 크게 없는 경우 많음)
  - $\beta_1$ :  $X_2$ 의 영향이 조정된 상태에서,  $X_1$ 의 단위 변화에 따른 Y의 변화량
  - $\beta_2$ :  $X_1$ 의 영향이 조정된 상태에서,  $X_2$ 의 단위 변화에 따른 Y의 변화량
- 즉, 설명변수(X)가 여러 개인 경우 각 X 앞에 곱해진 모수( $\beta$ )의 의미는 **다른 설명변수들의 영향이 모두 조정된 상태에서** 각 X의 단위 변화에 따른 Y의 변화량

# PSM vs Linear Regression

PSM은 다중선회귀분석 대비 어떠한 이점이 있는가?

- 선형회귀분석의 경우 혼동변수를 통제변수(control variable) 형태로 모형에 포함(모델링)하여 그 영향을 통제함
  - 선형 가정을 기반으로 존재하지 않는 관측치에 대한 예측, 변수의 중요도 평가 등 다양한 분석이 가능함 (인과성 추론만이 목적 아님)
  - 단, 혼동변수가 종속변수에 '선형'으로 영향을 미친다고 가정하며, 경우에 따라 (특히 분석 대상 현상이 복잡한 경우) 이러한 선형 가정은 생각보다 큰 제약일 수 있음
- PSM의 경우 관찰 데이터에 매칭 기법을 적용하여 준실험적 조건을 구성, 강력한 인과성 추론 기반을 제공하며 선형 가정을 완화할 수 있음
  - 인과성 추론이 중요한 목적
  - 단, 매칭되지 못한 일부 관측치를 버리게 될 수 있으며, 매칭되지 못한 관측치와 연관된 그룹에 대해서는 설명하기 어려움

# PSM vs Linear Regression

## PSM vs 선형회귀분석 예시

- **A대학 취업지원 프로그램이 취업률 개선에 미치는 영향을 분석하기 위해 데이터를 수집했다고 해보자**
  - 분석 대상 인과 관계: 취업지원 프로그램 이수 → 취업률 개선
  - 그런데 **놀랍게도, 프로그램 이수자 모두가 남성이었다고 해보자**
- 이 경우 PSM을 활용한다면 보유한 **처치군(프로그램을 이수한 남성)**에 대응하는 적절한 **대조군**을 매칭, 취업지원 프로그램이 취업에 미친 영향(**인과성**)을 효과적으로 분석할 수 있음
  - 단, 혼동변수로 성별이 고려되어 **대조군이 모두 남성으로 매칭된다면 분석에 포함되지 못한 그룹(여성)**에 대해 담론을 확장하기는 어려움
- 선형회귀분석을 활용할 경우 **모든 데이터를 활용할 수 있고**, 추정된 모형을 바탕으로 존재하지 않는 관측치(**여성 이수자**)에 대한 예측도 가능
  - 단, 주요 가정 만족에 대한 엄밀한 검토가 필요함



# PSM vs Linear Regression

## PSM vs 선형회귀분석 결론

- 연구의 목적, 가용한 데이터, 연구 질문의 특징 등을 종합적으로 고려하여 적절한 분석 방법을 선택할 수 있음
  - 예) 앞선 취업지원 프로그램의 사례에서 여성 이수자에 대한 효과성을 논의하는 것은 중요한 문제이겠지만, 건강한 20대가 뇌졸중에 걸릴 확률이나 R&D 인력이 없는 기업에 대한 R&D 지원 효과성에 대한 것을 논의하는 것은 비교적 중요한 문제일까?
- 여러 방법론의 주요 가정과 한계를 이해하고, 이를 종합하여 분석하고 해석하여 결론을 내리는 것도 좋은 방법임

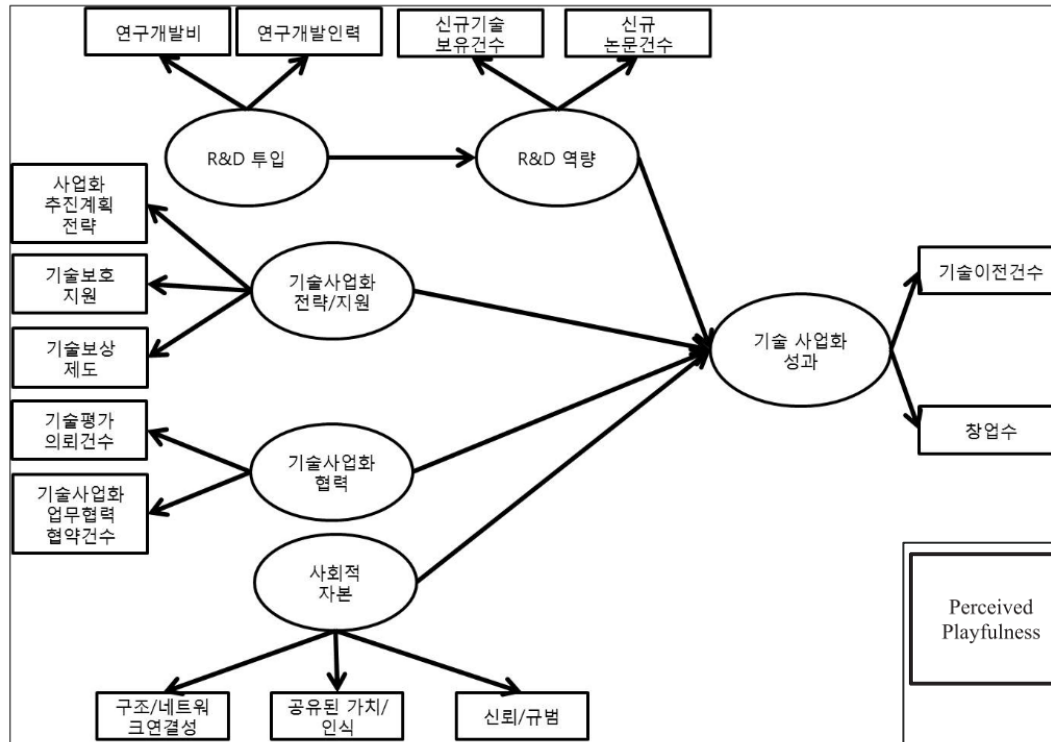


맞는 도구를 쓰자!



# Structural Equation Modelling

번외, Recall) SEM (Structural Equation Modelling, 구조방정식)



SEM 예시

SEM의 일종인 TAM(의 일종)  
(Technology Acceptance Model)

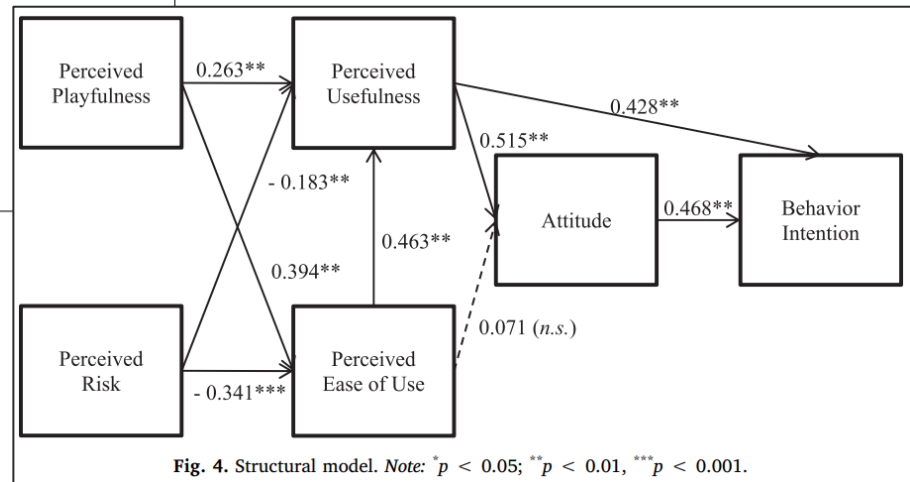


Fig. 4. Structural model. Note: \* $p < 0.05$ ; \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

# PSM vs SEM

## 번외) PSM vs SEM

- SEM(구조방정식, Structural Equation Model) 역시 인과성 추론에 자주 활용되는 방법론 중 하나
  - 복잡한 개념을 **잠재변수(latent variable)**라는 개념을 통해 추정하고 이들 사이의 **복잡한 인과 관계**를 모델링할 수 있음
    - ✓ 이에 따라 특수한 연구설계 및 비교적 많은 표본이 필요함
  - SEM의 가장 큰 단점은 연구자가 **인과의 방향을 (비록 근거를 제시하긴 하지만) 자의적으로 설정**한다는 것
- 즉, PSM이 특정 처치에 따른 인과성 추론에 주요 목적이 있다면, SEM은 수치화하기 어려운 잠재변수 간의 복잡한 인과관계 식별에 주요 목적이 있음

**Recap**

# Recap

## PSM이란 무엇이고 그 의미는 무엇인가?

- **PSM**은 인과성 분석을 수행함에 있어 **무작위통제실험**이 어려운 경우 자주 활용되는 **매칭 방법** 중 하나
  - 처치군과 대조군 사이의 차이를 통제하기 위한 통계적 기법
  - 혼동 변수들을 바탕으로 추정된 **성향 점수(propensity score)**를 바탕으로 **적절한 비교 대상을 짝지어줌(matching)**
- 실무에서는 직접 무작위통제실험을 하기보다는 이미 존재하는 데이터를 분석하게 되는 경우가 (훨씬) 더 많음
- PSM은 이미 확보된 데이터를 바탕으로 추가적인 처리를 거쳐 **실험에 준하는 분석 환경을 구축**할 수 있도록 해줌
  - 준실험(quasi-experimental) 방법
  - 관찰 연구 vs 실험 연구?

# Recap

## PSM의 한계점

- **데이터 손실 (data loss)**
  - PSM은 표본들을 서로 매칭해주는 것이기 때문에, **매칭되지 못한 표본은 분석에서 제외**되어 표본 수가 감소할 수 있으며, 이에 따라 **결과 일반화에 제약**이 발생할 수 있음
- **관찰되지 못한 혼동 변수(unobserved confounding variable)**
  - PSM에서는 고려된 혼동 변수들만 처치 여부에 영향을 미친다고 가정
  - **미처 고려되지 못한 혼동 변수로 인한 편향**을 완전히 배제할 수 없음
- **불완전 매칭 (imperfect matching)**
  - PSM을 통해 이루어진 매칭이 적절하지 않을 수 있음
- 적절히 설계된 무작위통제실험만큼의 인과관계를 확립하기는 어려움

# Recap

## PSM 분석 단계

- PSM 분석의 과정은 아래 3단계로 요약할 수 있음
  1. 성향점수 추정
  2. 매칭 및 매칭 결과 검증
  3. 인과성 추론
- 각 단계와 연관된 주요 개념 및 유의 사항 숙지
- PSM의 주요 가정 숙지

# Recap

## PSM과 다른 모형의 비교

- **선형회귀분석**의 경우 인과성 추론 외에 다양한 목적으로도 활용되며, 선형 가정을 기반으로 다양한 응용이 가능함
- **PSM**은 인과성 추론에 주요한 목적이 있으며, 이를 위해 관찰 데이터에 매칭 기법을 적용하여 준실험적 조건을 구성함
- **연구의 목적, 가용한 데이터, 연구 질문의 특징** 등을 종합적으로 고려하여 적절한 분석 방법을 선택할 수 있음
- 여러 방법론의 주요 가정과 한계를 이해하고, 이를 종합하여 분석하고 해석하여 결론을 내리는 것도 좋은 방법임



# **Appendix:**

# **R Statistical Software**

# R Statistical Software

## R 개요

- 1995년 뉴질랜드 오클랜드 대학의 **R**oss Ihaka와 **R**obert Gentleman에 의해 개발
- 현재는 R development core team이라는 **비영리단체**에 의해 개발 및 유지보수가 이루어지고 있음
- 무료로 제공되며, 학계, 산업계 등에서 데이터 분석을 위해 널리 사용됨
  - 특히, **통계학** 등 사회과학 연구자들에게 인기가 많음



# R Statistical Software

## R 장단점

### ■ 장점

- 무료로 제공됨
- 다양한 통계적 분석 및 시각화 도구를 제공함
- 유저 수(**installed base**)가 상당히 많음  
(전문 및 일반 사용자들의 신규 라이브러리 개발 및 업데이트 활발, 다양한 보완 소프트웨어, 학습도구 및 질의응답 커뮤니티 활성화 등)

### ■ 단점

- 프로그래밍에 익숙하지 않다면 타 통계패키지(SPSS, SAS, Stata 등)에 비해 학습 난이도가 있음
- 프로그래밍에 매우 익숙한 사용자 입장에서조차 아쉬운 측면이 있음  
(비효율 등)
- 일반 사용자들이 개발한 패키지는 상용 통계 패키지와 달리 꾸준히 버전업 및 유지보수가 되지 않으며, **개발자가 결과에 대해 책임지지도 않음**
- **한글** 데이터를 활용할 때 까다로운 부분이 있음

# R Statistical Software

## R 설치 (1)

- <http://www.r-project.org>에서 무료 다운로드



[\[Home\]](#)

**Download**

[CRAN](#)

**R Project**

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting Bugs](#)

[Conferences](#)

[Search](#)

[Get Involved: Mailing Lists](#)

[Get Involved: Contributing](#)

## The R Project for Statistical Computing

### Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

### News

- **R version 4.3.3 (Angel Food Cake)** has been released on 2024-02-29.
- **Registration for useR! 2024** has opened with early bird deadline March 31 2024.
- **R version 4.3.2 (Eye Holes)** has been released on 2023-10-31.
- **R version 4.2.3 (Shortstop Beagle)** has been released on 2023-03-15.
- You can support the R Foundation with a renewable subscription as a [supporting member](#).

# R Statistical Software

## R 설치 (2)

Indonesia

<https://cran.usk.ac.id/>

Universitas Syiah Kuala

Iran

<https://cran.um.ac.ir/>

Ferdowsi University of Mashhad

Italy

<https://cran.mirror.garr.it/CRAN/>

Garr Mirror, Milano

<https://cran.stat.unipd.it/>

University of Padua

Japan

<https://cran.ism.ac.jp/>

The Institute of Statistical Mathematics, Tokyo

<https://ftp.yz.yamagata-u.ac.jp/pub/cran/>

Yamagata University

Korea

<https://cran.yu.ac.kr/>

Yeungnam University

Mexico

<https://cran.itam.mx/>

<https://www.est.colpos.mx/>

Morocco

<https://mirror.marwan.ma/cran/>

Netherlands

<https://mirrors.evoluso.com/CRAN/>

<https://mirror.lyrahosting.com/CRAN/>

### The Comprehensive R Archive Network

#### Download and Install R

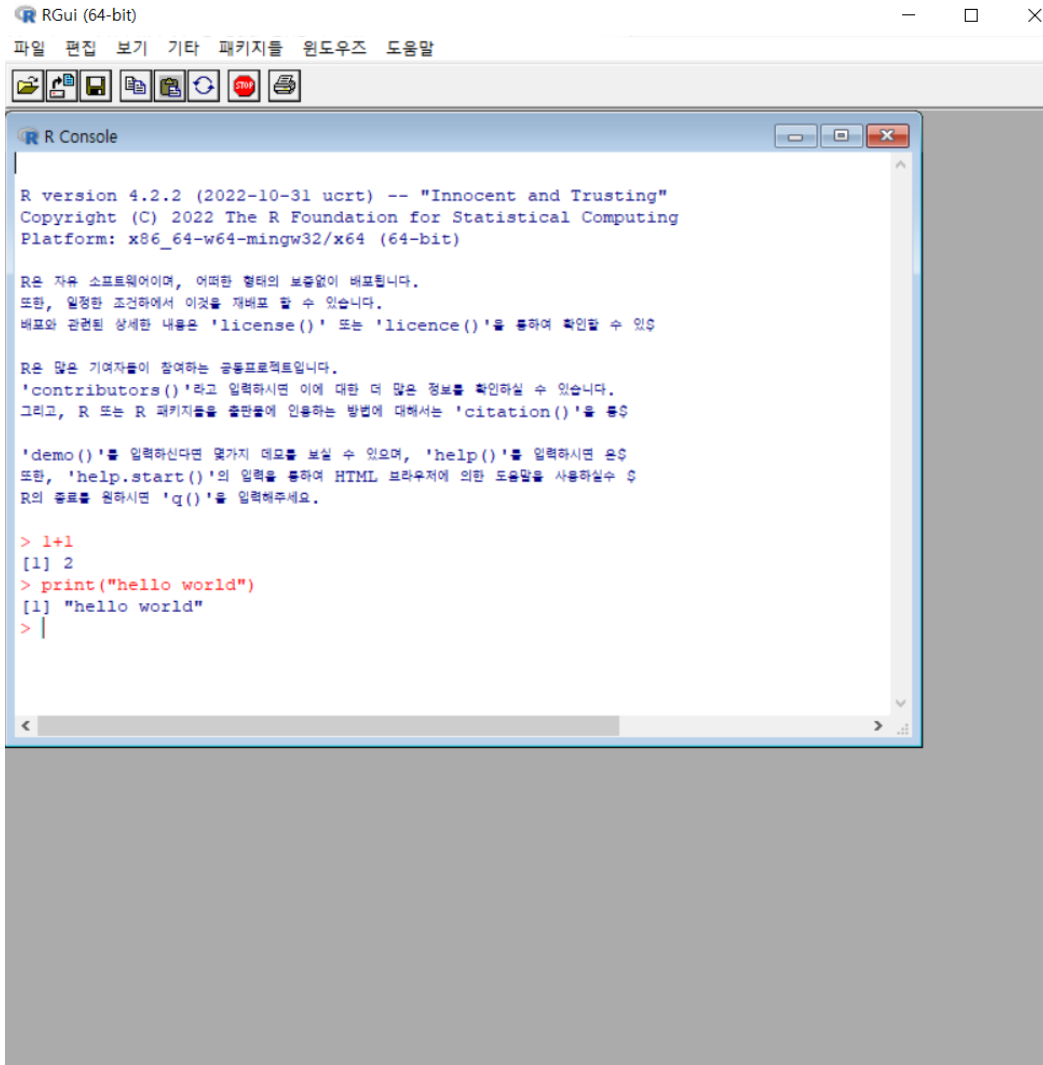
Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#) ([Debian](#), [Fedora/Redhat](#), [Ubuntu](#))
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

# R Statistical Software

## R 기본 UI (RGui)



The screenshot shows the RGui (64-bit) window. The title bar reads "RGui (64-bit)". The menu bar includes "파일", "편집", "보기", "기타", "패키지를", "윈도우즈", and "도움말". The toolbar contains icons for file operations and execution. The R Console window is open, displaying the following text:

```
R version 4.2.2 (2022-10-31 ucrt) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R은 자유 소프트웨어이며, 어떠한 형태의 보증없이 배포됩니다.
또한, 일정한 조건하에서 이것을 재배포 할 수 있습니다.
배포와 관련된 상세한 내용은 'license()' 또는 'licence()'를 통하여 확인할 수 있습니다.

R은 많은 기여자들이 참여하는 공동프로젝트입니다.
'contributors()'라고 입력하시면 이에 대한 더 많은 정보를 확인할 수 있습니다.
그리고, R 또는 R 패키지들을 출판물에 인용하는 방법에 대해서는 'citation()'을 참조하십시오.

'demo()'를 입력하신다면 몇가지 예제를 보실 수 있으며, 'help()'를 입력하시면 도움말
또한, 'help.start()'의 입력을 통하여 HTML 브라우저에 의한 도움말을 사용할 수 있습니다.
R의 종료할 원하시면 'q()'를 입력하십시오.
```

The user has entered the following commands and received the output:

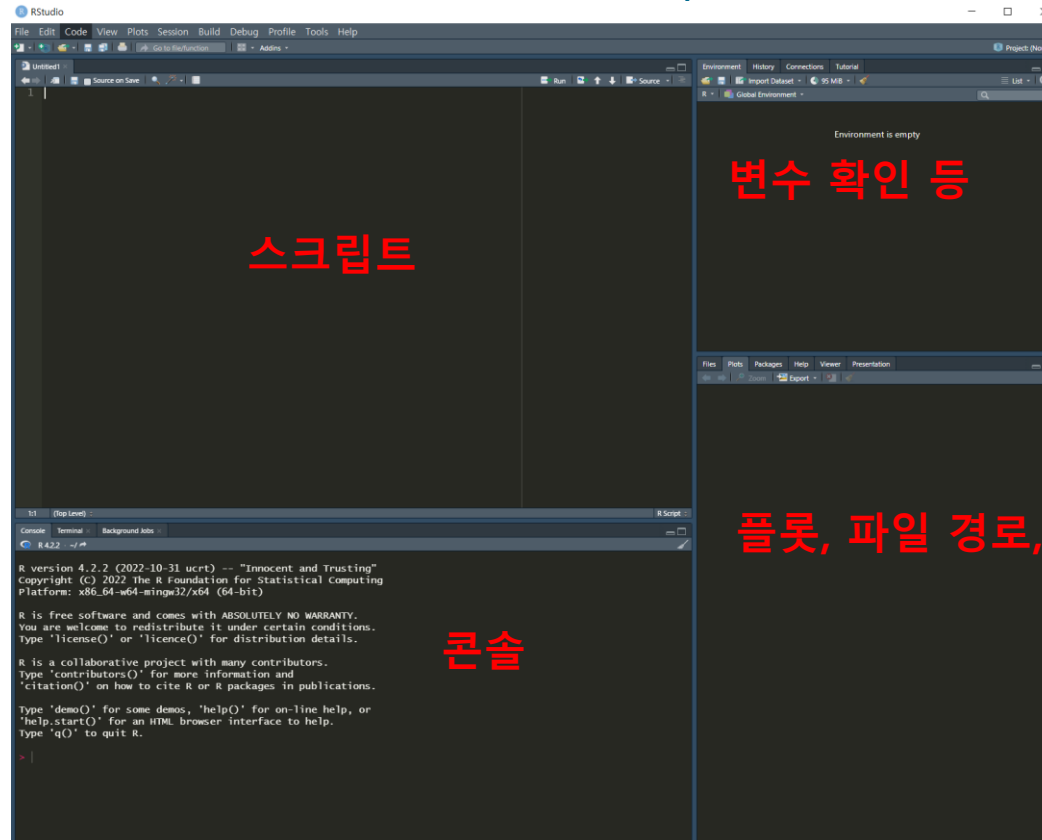
```
> 1+1
[1] 2
> print("hello world")
[1] "hello world"
> |
```



# R Statistical Software

## R Studio

- 더 보기 좋고 편리한 UI를 위해 **R Studio**를 설치하여 활용하는 것을 권장
- <https://posit.co/download/rstudio-desktop/>에서 Rstudio Desktop 설치



# R Statistical Software

## R 패키지 활용법

- 기본적인 기능(선형회귀분석, t-검정 등)은 별도 패키지 설치 없이도 가능
  - 주요 기능은 이미 내장
  - R에서 기본 제공하는 몇 가지 샘플 데이터도 있음  
<https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html>
- 그러나 **기개발된 다양한 패키지들을 활용**하기 위해서는 패키지 설치 및 임포트(불러오기)가 필요함
- 패키지 설치 커맨드: `install.packages("패키지명")`
- 패키지 임포트: `library("패키지명")`
  - Python에서 `import`와 같은 기능



# R Statistical Software

## R에서 한글이 깨져 보일 때 해결법

- Rstudio 좌상단 File > Reopen with Encoding > Show all encodings 체크박스 체크 > euc-kr 선택 > ok 클릭

