경영경제데이터분석

보충자료2

최 현 홍
(hongchoi@khu.ac.kr)

Contents

보충자료2

- 선형회귀모형 vs PSM with 예시 데이터

Linear Regression Model & PSM with Example Data

Supplementary2

데이터 설명

- 활용 자료: A대학 취업지원 프로그램 관련 성과 자료 (가상) (BEDA_support2.csv)
  - **result**: 취업 여부 (취업: 1, 미취업: 0)
  - **program**: 취업지원 프로그램 참여 여부 (참여: 1, 미참여: 0)
  - **male**: 남성 여부 (남성: 1, 여성: 0)
  - **gpa**: 학점 (4.3 만점)
  - 관측치 수: 35

- 분석 대상 인과 관계: **program → result**

- 고려 가능한 혼동 변수: **male, gpa**

이외에도 혼동변수는 매우 많겠지만...
예시를 위한 분석

- 데이터 특이 사항: 프로그램 참여 인원이 전부 남성임

Supplementary2

선형회귀분석 시

```
#데이터 불러오기
mydata3 <- read.csv("BEDA_support2.csv")

#데이터 둘러보기
summary(mydata3)

#선형회귀모형 분석
myresult <- lm(result ~ program + male + gpa, data = mydata3)

##선형회귀모형 분석 결과 보기
summary(myresult)
```

$$Y_{result} = \beta_0 + \beta_1 X_{program} + \beta_2 X_{male} + \beta_3 X_{gpa} + \varepsilon$$

종속변수: 취업 여부

설명변수: 프로그램 참여 여부, 성별, 학점

성별, 학점은 통제변수 개념으로 포함

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4347     0.6093  -0.713 0.480920
program       0.6703     0.1774   3.778 0.000674 ***
male         -0.5069     0.1619  -3.131 0.003787 **
gpa           0.3300     0.1697   1.944 0.060967 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Supplementary2

선형회귀분석 결과의 해석 각 설명변수가 결과에 미치는 영향 수준 식별

- 모수 추정 결과 $Y_{result} = \beta_0 + \beta_1 X_{program} + \beta_2 X_{male} + \beta_3 X_{gpa} + \varepsilon$

Variables	Coefficient	Std.Error	P-value	
Intercept (β_0)	-0.4347	0.609	0.481	
program (β_1)	**0.670**	**0.177**	**0.001**	**1% 유의수준에서 유의**
male (β_2)	-0.507	0.162	0.003	1% 유의수준에서 유의
gpa (β_3)	0.330	0.170	0.061	10% 유의수준에서 유의

- **(성별 및 학점이 통제되었을 때) 프로그램 참여 시 취업률이 67% 높음**
- (프로그램 참여 여부 및 학점이 통제되었을 때) 남성의 취업률이 50.1% 낮음
- (프로그램 참여 여부 및 성별이 통제되었을 때) 학점이 1점 더 높을 때마다 취업률이 33% 높아짐

참고) 엄밀한 분석 시 현 데이터는 설명변수간 높은 상관관계로 인해 편향이 발생할 가능성이 높으므로, 자료 추가 확보 등의 노력이 필요할 수 있음
(본 분석은 예시 및 비교를 위한 분석임에 유의하자)

Supplementary2

선형회귀분석을 이용한 예측

실제로는 존재하지 않는 관측치에 대한 예측도 가능

설명변수명에 맞춰 예측하고 싶은 X값 투입
(아래 예시는 여성, 프로그램 참여, 학점 4.0)

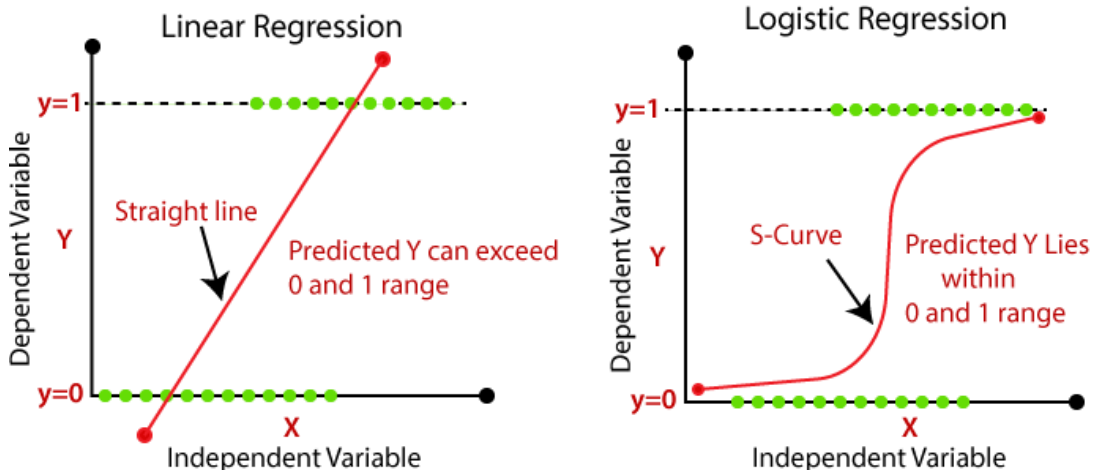
#선형회귀모형 예측

```
predict(myresult, newdata = data.frame(male = c(0), program = c(1), gpa = c(4)))
```

회귀분석 결과

```
> predict(myresult, newdata = data.frame(male = c(0), program = c(1), gpa = c(4)))
       1
1.555416
```

취업률이 155%...?



Supplementary2

로짓선형회귀분석모형 및 예측

```
#로짓선형회귀모형 분석
myresult2 <- glm(result ~ male + program + gpa, data = mydata3, family = binomial)

#로짓선형회귀모형 분석 결과 보기 (직접 해석 X)
summary(myresult2)

#로짓선형회귀모형 예측
predict(myresult2, newdata = data.frame(male = c(0), program = c(1), gpa = c(4)), type = "response")
```

```
> predict(myresult2, newdata = data.frame(male = c(
        1
0.9978569
```

99.8%

```
#로짓선형회귀모형 예측 - 복수
```

- `predict(myresult2, newdata = data.frame(male = c(0,0), program = c(1,0), gpa = c(4,4)), type = "response")`

```
> predict(myresult2, newdata = data.frame(male =
onse")
        1         2
0.9978569 0.9169397
```

학점 4.0인 여학생이 프로그램 참여 시 99.8% 미참여 시 91.7%

Supplementary2

PSM 분석 시 - 매칭

혼동변수로 성별 및 학점 고려

```
#PSM 매칭 시행
mymodel3 <- matchit (program ~ male + gpa,
                     data = mydata3,
                     distance = "glm",
                     method = "nearest")

#매칭 결과 요약
summary(mymodel3)
```

```
Call:
matchit(formula = program ~ male + gpa, data = mydata3, method = "nearest",
    distance = "glm")

Summary of Balance for All Data:
         Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
distance        0.5107        0.1957          3.1223     0.1685    0.3043     0.60
male            1.0000        0.4000          1.4491          .    0.6000     0.60
gpa             3.5500        3.5080          0.0988     1.1289    0.0500     0.14

Summary of Balance for Matched Data:
         Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
distance        0.5107        0.4893          0.2124     6.8950    0.0786      0.3          0.4434
male            1.0000        1.0000          0.0000          .    0.0000      0.0          0.0000
gpa             3.5500        3.4600          0.2118     7.2545    0.1100      0.3          0.4471

Sample Sizes:
          Control Treated
All            25      10
Matched        10      10
Unmatched      15       0
Discarded       0       0
```

학점 차이가 좀 아쉽지만, 적절한 매칭이라고 가정하고 넘어가보자

Supplementary2

PSM 분석 시 – 매칭 데이터 요약 및 인과 분석

#매칭 데이터 추출

matched_data3 <- match.data(mymodel3)

#매칭 데이터 요약

summary(matched_data3)

전부 남성

```
> summary(matched_data3)
      pid             result          male      program          gpa           distance
 Min.   : 1.00   Min.   :0.00   Min.   :1   Min.   :0.0   Min.   :2.800   Min.   :0.3341
 1st Qu.:10.75   1st Qu.:0.00   1st Qu.:1   1st Qu.:0.0   1st Qu.:3.400   1st Qu.:0.4746
 Median :18.50   Median :1.00   Median :1   Median :0.5   Median :3.400   Median :0.4746
 Mean   :19.00   Mean   :0.55   Mean   :1   Mean   :0.5   Mean   :3.505   Mean   :0.5000
 3rd Qu.:28.75   3rd Qu.:1.00   3rd Qu.:1   3rd Qu.:1.0   3rd Qu.:3.525   3rd Qu.:0.5052
 Max.   :35.00   Max.   :1.00   Max.   :1   Max.   :1.0   Max.   :4.200   Max.   :0.6642
```

#인과성 분석

with(matched_data3, t.test(result ~ program))

```
> with(matched_data3, t.test(result ~ program))

        Welch Two Sample t-test

data:  result by program
t = -4.2, df = 16.691, p-value = 0.0006237
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -1.0521318 -0.3478682
sample estimates:
mean in group 0 mean in group 1
            0.2             0.9
```

t=-4.2, p-value = 0.0006

인과성 존재한다고 주장할 수 있으나,
표본 및 매칭 관련 한계 언급 필요