클린업 2주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 클린업 1-3주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 16시 00분에 발표됩니다.
- 제출기한은 **목요일 자정까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요. 또한 불성실한 제출로 인한 경고를 2회 받으실시도 퇴출이니 유의해주세요.

Chapter 0 : Theory

통계에서 가장 기본적으로 배우게 되는 선형회귀분석은 예측을 위해 사용되는 모델이기도 하지만 인과관계를 추정하기 위한 모델이기도 합니다. 인과관계를 상정하기 위해서는 독립변수와 오차 항 간의 공분산이 0($Cov(X,e)=0$)이어야 한다는 조건이 붙습니다. 이때 공분산이 0인 경우를 외생성, 공분산이 0이 아닌 경우를 내생성이라고 합니다.

다음 문제를 풀어보면서 통계에서 잘 다루지 않는 인과관계에 대해서 공부를 해보겠습니다. 문제는 가상의 인물과 가상의 데이터로 구성되어 있습니다.

S대 통계학과의 J교수는 학생들의 출석률이 학습 성취도에 강한 영향을 미친다고 생각합니다. 이에 S대 본부에 더욱 강력한 출결 제도를 요구하기 위해 출석율이 기말고사 성적에 양의 관계를 가지는 것을 보이고자 합니다. 아래는 J교수의 "통계적 OOO" 과목 학생들의 출석률(atndrte)과 이전 학기까지 성적 (priGPA), 그리고 표준화된 중간고사 성적(stndmid)을 독립변수로 하여 표준화된 기말고사 성적(stndfin)를 예측한 모델입니다.

$$\widehat{\text{stndfn}} = -0.021\text{atndrte} + 0.082\text{stndmid} - 0.555\text{priGPA} + 0.011\text{atndrte} * \text{priGPA}$$

	Atndrte	Stndmid	priGPA	Atndrte*priGPA
Coefficient	-0.021	0.082	-0.555	0.011
Se	(0.002)	(0.011)	(0.078)	(0.004)
	n=680	$R^2 = 0.201$		

이때 기말고사 성적, 중간고사 성적은 제대로 채점되었고, 이전 학기까지 성적은 대학 본부로부터 제공받아 측정 오차가 없다고 가정합니다. 출석률은 전자 출결로 측정되었으나 출석률을 제출하고 싶지 않은 학생은 교수에게 제출하지 않을 수 있습니다.

문제1. 각 회귀계수들의 t value를 구하고 회귀계수들이 통계적으로 유의한지 설명해주세요.

문제2. J 교수는 학생들이 수업을 들었음에도 전자 출결을 하지 않는 경우와 수업을 듣지 않았음에도 전자 출결을 하는 이른바 '출튀' 문제 때문에 얻은 출석률에 측정 오류가 있다는 것을 알았습니다. 이때 출석률의 회귀계수에 어떤 영향을 주는지 설명해주세요.

(HINT) 독립변수에 measurement error가 있는 경우 $X_i = X_i^* + \mu_i$로 회귀 모델을 두고 풀게 됩니다. 이때 X_i는 잘못 측정된 값, X_i^*은 원래 값, μ_i는 오차를 의미하는 확률변수입니다.

문제 3. J 교수는 기말고사를 치른 800 명의 학생 중 120 명이 출석률 제출에 동의하지 않았다는 것을 알았습니다. 이 경우에 출석률의 회귀계수에 어떤 영향을 주는지 설명해주세요.

(HINT) 이 경우는 1936 년 미국 대선 선거 여론 조사와 같은 경우입니다. 이때 240 만명의 우편 설문지를 조사한 결과 공화당의 앨프 랜던이 57% 지지를 얻어 선거에서 승리할 것으로 예상했지만, 실제로는 민주당의 플랭클린 루즈벨트가 61%의 득표율로 재선을 하였습니다.

문제4. 이 회귀모델에서 출석률의 회귀계수 해석과 관련하여 stndmid, priGPA가 무슨 역할을 하는 지 설명해 주세요. 또한 위 모델에서 출석률(atndre)에 영향을 줄 수 있는 잠재변수를 말해주세요.

(HINT) 출석률에 영향을 줄 수 있는 관측하지 못하는 잠재변수는 정답이 없으므로 자유롭게 적어주세요.

문제5. J교수는 추가적으로 강의실과 학생들의 거주지의 거리(dist)를 대학 본부로부터 받았습니다. 이 변수를 사용하여 회귀모델을 자유롭게 재구성해주세요.

(HINT1) dist와 atndre 사이에 어떤 관계가 있을지 생각해보세요.

(HINT2) dist는 atndre를 제외한 다른 변수들과는 낮은 상관관계를 보일 것입니다.

Chapter 1 : Data Processing & Background

이번 패키지에서는 사회과학에서 통계가 어떻게 사용되는지를 살펴보겠습니다. 패키지의 흐름은 "입시 제도에서 나타나는 적응의 법칙과 엘리트 대학 진학의 공정성"이라는 논문을 따라 가고 있으며 Chapter 0에서 다뤘듯이 특정 변수(가족 배경)가 종속 변수(엘리트 대학 진학 여부)에 어떤 영향을 주는 지에 대해 집중하고 있습니다.

가설에 맞게 데이터 전처리를 하는 것은 중간고사 이후 주제분석 큰 경험이 될 것이니 모두 포기하지 말고 끝까지 해봅시다. 이 논문의 가설은 "입시전형과 관계없이 가족의 사회경제적 배경이 높을수록 상위권 대학 진학의 확률이 커지지만, 그 정도는 입시전형마다 다르다"입니다.

우리가 사용할 데이터는 대졸자 직업 이동 경로 조사(GOMS)로 2017년과 2018년에 조사된 데이터를 다룹니다. 변수가 많은 관계로 각 변수에 대한 정보는 같이 첨부된 code book에 나타나 있으니 참고하시면 분석을 진행하면 되겠습니다.

[조건 : 가급적 tidyverse 이외의 패키지 사용 금지, %>% 연산자를 최대한 활용하여 간결한 코드 작성]

문제1. 주어진 데이터는 stata 13에서 작성된 데이터입니다. 이를 불러오기 위해 readstata13 라이브러리를 설치하고 read.dta13 함수를 이용해 데이터를 불러와주세요.

문제2. 저희가 사용할 변수는 goms2017에서 g161pid~g161gradum과g161f___(ex: g161f001, ...) g161p___(ex:g161p001, ...), goms2018에서 g171pid~g171gradum, g171f___(ex: g171f001, ...) g171p___(ex:g171p001, ...)에 해당하는 변수입니다. 이를 제외한 나머지 변수를 삭제해주세요.

(HINT 1) Select 함수의 starts_with를 사용하면 쉽게 변수를 선택할 수 있습니다.

문제3. goms_2017에서 g161f___ 중 g161f001~g161f017(g161f170은 포함되어야합니다.)을 제외한 나머지 g161f___ 변수를 지워주세요. goms_2018에서 g171f___ 중 g161f001~g161f017을 제외한 나머지 g171f___ 변수를 지워주세요.

문제4. goms_2017에서 g161p___ 중 g161p018~g161p036을 제외한 나머진 g161p___ 변수를 지워주세요. 마찬가지로 goms_2018에서 g171p___ 중 g171p018~g171p036을 제외한 나머진 g171p___ 변수를 지워주세요.

문제5. goms_2018에서 추가적으로 g171dpmt_n과 g171major_n을 제거해주세요.

문제6. goms_col_17의 colnam을 goms_2017에 id 기준으로 붙여주세요. 마찬가지로 goms_col_18의 colnam을 goms_2018에 id 기준으로 붙여주세요.

문제7. 두 데이터프레임을 병합하기 위해 변수 이름을 동일하게 맞춰줘야 할 필요성이 있습니다. 각 데이터프레임 앞에 붙은 조사년도 접두사(g161, g171)를 제거해주세요. (g161school -> school)

(HINT) renamewith와 gsub를 이용하면 쉽게 패턴을 찾아 이름을 변경할 수 있습니다.

문제8. 두 데이터프레임을 병합해주세요. 단 조사년도를 확인하기 위해서 새롭게 파생변수(year)를 만들면서 병합해주세요.

(HINT) bind_rows를 이용하면 r 기본함수 rbind를 이용한 것보다 빠르게 데이터프레임을 병합할 수 있습니다.

문제9. 이 분석은 4년제를 대상으로 진행합니다. School 변수에서 4년제 대학생이 아닌 경우를 제거해주세요.

문제10. 2015년에 입학 사정관제가 학생부종합전형으로 개편되었습니다. 그러나 이 두 전형은 이질적인 특성을 가지고 있으므로 구분해야 합니다. 본 연구는 2009년부터 2013년의 대학입학자를 대상으로 하므로 1988~1994년 출생자(birthy)가 아닌 경우 제거해주세요.

[보너스 문제 1] 입학 사정관제와 학생부 종합전형이 이질적인 특성을 가지고 있음에도 불구하고 구분하지 않아 연구 대상에 포함하면 가설과 관련하여 어떤 문제가 발생하는지 설명해주세요.

문제11. 또한 고교 졸업 후 바로 대학시장에 졸업하지 않고 노동시장에서 경험을 쌓은 후 대학에 진학한 비전형적인 수험생을 제외하기 위해 고등학교 졸업시기(f001)와 대학입학 시기(f011)의 간격이 3년 이상인 경우도 제거해주세요.

문제12. 본 연구는 입학 대학이 종속 변수이기 때문에 졸업대학과 고교 졸업 후 입학 대학이 다른 편입(f010)생 또한 제거해주세요.

문제 13. 입학 대학이 경희대, 고려대, 서강대, 서울대, 서울시립대, 성균관대, 연세대, 이화여대, 중앙대, 한국외국어대, 한양대, 모든 의약대학, 한국과학기술원, 포항공과 대학교이면 1로 그렇지 않다면 0으로 바꿔주세요.

- 의과 대학은 majorcat을 통해서 확인할 수 있습니다.

본 연구에서는 가족배경이 엘리트 대학 진학률에 어떤 영향을 미치는지 확인하려고 합니다. 이때 가족배경은 사회경제적 배경과 출신 고등학교 위치로 나누었고, 이 중 사회경제적 배경은 대학 입학 당시의 부모의 소득(p034), 현재 부모의 자산 정도(p036), 부모 직업의 위계(p032, p033), 부모의 교육수준(p028z, p031z)으로 측정했습니다.

문제 14. 대학 입학 당시의 부모의 소득은 구간별로 측정되어 있습니다. 측정된 구간의 크기가 다르기 때문에 p034 변수를 중위값으로 대체하고 관측치의 입학년도 별(f011) 백분위 값으로 대체해주세요.

example) 월 소득 700~1000만원 미만의 경우 850으로 대체. 같은 850만원이어도 2009년 입학자는 86분위 2011년 입학자는 83분위.

- 모름 및 무응답의 경우는 제거해주세요. 안계심(사망)의 경우는 소득 없음과 마찬가지로 중위값을 0으로 처리해주세요.
- 백분위 값은 소수점 첫째 자리에서 반올림해주세요.

(HINT) group_by와 quantile 이용하여 입학년도 별 백분위를 쉽게 구할 수 있습니다.

문제 15. 현재 부모의 자산 정도도 마찬가지로 중위값으로 대체하고 조사연도(2017, 2018)별로 백분위 값으로 대체해주세요.

문제 16. 교육 수준은 부모 중 교육수준이 높은 쪽의 응답을 아래의 교육년수로 전환한 후 입학년도 별 백분위 값으로 전환해주세요.

초졸이하 6, 중졸 9, 고졸 12, 전문대졸 14.5, 대졸 16, 대학원졸 18

문제 17. 부모 직업의 위계는 각 직업별 대졸이상 학력자의 비율로 변경하고 부모의 평균값을 사용해주세요. 만약 부모의 직업변수가 결측치인 경우 부모의 학력 변수를 이용하여 동일 학력자의 평균 직업 위계 점수를 부여해주세요. 이후 위 값들을 다른 변수와 마찬가지로 백분위 값으로 전환해주세요.

문제 18. 마지막으로 가족배경 종합지표를 만들어 주세요. 가족배경 종합지표는 소득, 자산, 교육, 직업 위계의 백분위 값을 단순 합산한 후 이 값의 백분위를 다시 계산한 뒤 10으로 나누어 계산합니다.

example) 소득 85, 자산 30, 교육 50, 직업 70이라면 235로 계산하고 235의 백분위를 구한 뒤 10으로 나눕니다.

[보너스 문제 2] 소득, 자산, 교육, 직업 위계의 백분위 값을 PCA를 진행하고 분산을 제일 잘 설명하는 축으로 선택해주세요. 이후 min-max scaling을 통해 0-100 사이의 값을 가지게 만들어주세요.

두 변수 중 어떤 변수가 가족배경 종합지표로 적절한지 설명해주세요.

문제 19. 출신 고등학교 위치 정보(f006, f007)를 서울, 경기도, 기타 광역시, 기타 광역도로 나누고 서울 강남 3구(강남구, 서초구, 송파구)의 경우 새로운 가변수를 만들어 주세요.

문제 20. 입시전형(f013, f170)은 수능 위주 정시 전형, 내신위주 전형, 학생부 종합 전형, 논술위주 수시전형, 기타 전형(실기 위주, 면접 위주, 서류 위주, 기타)로 나누어 주세요. f013은 정시-수시 여부, f170은 대학 입학 당시 전형을 의미합니다. 첫번째 응답은 수시이나 두번째 응답은 수능 위주 정시와 같은 응답이 일치하지 않는 경우는 기타로 분류해주세요.

문제 21. 통제 변수로는 인구학적 변수, 기술변수 그리고 출신 고교 유형을 통제하였습니다. 인구학적 변수는 성별(sex), 광역시도별 출생 지역(p20), 출생연도(birthy), 부친 부재 여부(p032), 모친 부재 여부(p033)를 포함합니다. 기술 변수는 대학입학 연도(f011), GOMS의 조사연도, 재수 여부, 삼수 여부를 포함합니다. 출신 고교 유형(f009)은 일반-인문계, 일반-자연계, 특목고-자사고, 기타 고교로 나누었습니다.

- 부친 부재 여부 및 모친 부재 여부는 각각 p032, p033을 이용해서 파생 변수를 만들어주세요.
- 재수 여부, 삼수 여부는 대학입학 연도(f011)와 고등학교 졸업 연도(f001)의 차이가 1인 경우는 재수, 2인 경우는 삼수로 파생 변수를 만들어주세요.

Chapter2 Modeling

이번 연구에서 사용한 모델은 Y과 0, 1일 때 X와의 관계를 선형으로 가정하는 선형확률모델(Linear Probability Model, LPM)입니다. 연구에서 사용된 모델은 아래와 같습니다.

$$y = \alpha + \beta SES + \sum_i \gamma_i App\,Type_i + \sum_i \delta_j (SES \times App\,Type_i) + \sum_j \zeta_j X_j + \epsilon$$

여기서 Y는 엘리트 입학(1, 기타는 0)을 나타내는 변수이고, SES는 가족 배경변수, AppType은 대학입시 전형 유형, X는 J개의 통제 변수입니다. 아래의 문제를 통해 모델을 해석해보겠습니다.

문제1. 모델링은 종속 변수와 독립 변수를 선형으로 가정하는 선형확률모델을 사용했습니다. 위의 데이터로 모델을 학습시키고 선형확률모델을 사용했을 때 문제점을 설명하고 그럼에도 선형확률모델을 쓴 이유에 대해서 설명해주세요.

문제2. 챕터 2에서 만들었던 가족 배경 변수와 대학인시 전형 유형, 통제 변수들로 구성된 모델을 학습시켜 주세요.

문제3. 연구의 가설인 "입시전형과 관계없이 가족의 사회경제적 배경이 높을수록 상위권 대학 진학의 확률이 커지지만, 그 정도는 입시전형마다 다르다"를 증명하기 위해선 위에서 어떤 변수가 유의해야 하는지 말하고 유의한지 확인해주세요.

문제4. 다른 변수들의 계수를 확인하고 통제 변수가 가설 검증을 위해 어떠한 영향을 주는지 설명해주세요.

문제5. 위의 연구에서 영향력을 보기 위해서 사용한 SES 변수에 내성성이 없는지 확인하고 있다면 어떤 데이터가 추가로 필요한지 설명해주세요.

(보너스 문제 2) 만약 연구의 가설이 "가족의 사회경제적 배경이 높을수록 상위권 대학 진학의 확률이 커진다"라면 어떤 식으로 모델이 구성되어야 하는지 설명해주세요.

(보너스 문제 3 로지스틱 회귀로 학습시키고 계수를 해석했을 때 연구에서 발생할 수 있는 문제점에 대해서 설명해주세요.