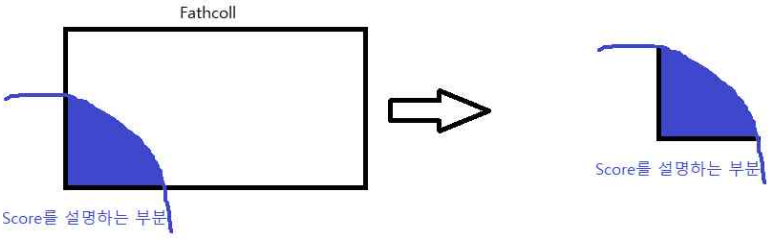
1. Problem set5: Question2-(5)

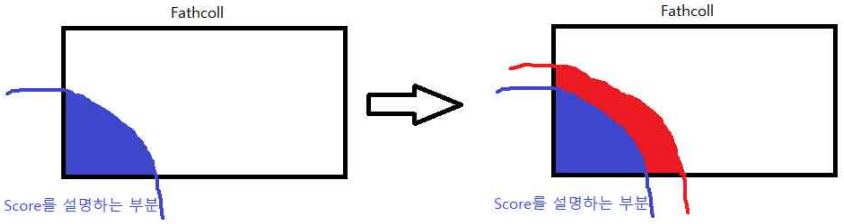
부모님의 대학원 졸업 여부나 전공을 더미변수로 형태로 추가하는 것이 mothcoll과 fathcoll이 score를 설명하는 부분을 따로 빼오는거라고 TA 세션(그림1)에서 설명해주셨는데 이 부분이 잘 이해가 가지 않습니다.

저는 부모님의 대학원 졸업 여부나 전공이라는 "새로운 정보"를 사용하여, 부모님의 학부 졸업 여부만으로는 설명하지 못한 부분(error term:u)을 줄여주니 score를 설명하는 부분이 확장된다고 이해했는데 어디를 잘못 생각한건지 모르겠습니다.

부모님이 학부를 졸업했냐 안했냐를 아는 것과 거기에 더해 부모님이 데이터 사이언스를 전공했다 등을 아는 것은 분명 다르다고 생각해서 그림2와 같은 관점으로 접근했습니다.



<그림1> TA 세션 설명



<그림2>

2. Lecture note 7&8: Question 1

HGPA나 SAT가 학생의 "IQ"라는 공통 변수의 영향을 받을 것 같은데 문제에 나와있는 식에다가 IQ까지 넣어서 regression을 돌린다면 어떻게 될지 궁금합니다.

1^{st} moment 관점에서는 IQ가 계수를 가져가서 HGPA와 SAT의 계수가 더 이상 유의해지지 않고 2^{nd} moment 관점에서는 correlation이 높아서 세 변수의 standard error가 크게 나올 것 같다고 하면 논리적일까요?

3. Lecture note 7&8: Question 5

5번에서 omitted variables가 많을 것 같아서 fraction online의 진정한 효과를 보려고 IV를 쓴다고 이해를 했는데요. (2)와 (3) column을 보니 IV가 OVB를 진정으로 제거해줄 수 있는지 의문이 생깁니다. OVB 때문에 사용한 IV인데 library visit까지 고려해야 한다는 걸 깨닫고 (3)-IV를 돌릴거면, (2)-IV는 OVB를 해결해주지 못하는걸로 보이는데 좀 더 근본적으로 IV가 OVB 있는 상태에서 사용했을 때는 언제든지 반박을 받을 수 있는 방법이고 말짱도루묵이지않냐는 생각이 들었습니다.

4. **Lecture note 7&8: Question 6-(b)**

(a)에서 omitted variable로 owner의 능력이나 노력을 생각할 수 있는데 House price growth를 변수로 추가한다고 해서 능력이나 노력을 고려해주는 건 아닌 것 같은데, 그러면 여전히 OVB의 위험에 노출되어있어 endogeneity를 해결할 수 없지 않나요?

5. Lecture note 7&8: Question 8- 전반적인 논리

A- 강의에서 사용한 논리는 'H=P+R이며, (2)에서 H의 계수가 0.21로 (1)에서 R의 계수와 같다 -> R이 M을 대부분 설명하지 P가 M을 설명하지 않는다'였습니다.

그런데 식 (1)에서 R의 t값이 2보다 작아 insignificant하고, 다른 방면으로 식 (2)에서 H의 t값은 2보다 작아서 insignificant하다고 판단할 수 있습니다. 그러면 R이 H 분산의 대부분을 차지한다해도 M을 설명한다고 주장하기 힘들 것 같다는 생각인데요. 수업 끝나고 대표님께 여쭤봤을때 식 (1)과 (2) 자체도 공부의 효율성을 반영하지 못하여 OVB에 노출되어있어서 그렇다고 말씀해주셨던걸로 기억하는데, 그렇게 치면 R의 분산이 H의 분산과 비슷하다는 결론 빼고는 어떠한 결론도 낼 수 없지 않나하는 생각입니다.

B- 예전에는 대표님께서 R이 평균 대비 분산이 커서(swing이 커서) M을 잘 설명할 것 같다고 예상을 하고 식 (1)을 봤더니 R의 t값이 2보다 작아서 M의 swing이 작은가보다는 관점으로 접근하셨던걸로 기억합니다. 그걸 증명해주는게 식 (1)에서 0.15, 0.21에 비해서 매우 큰 45.6이라는 유의한 상수항이였구요. 그렇다면 swing이 작은 H와 P가 M을 잘 설명해줘야 하는데 식 (2)를 보면 t값이 둘 다 2보다 작다. 그래서 omitted variable이 존재한다고 생각해야한다고 이해를 했었는데요.

B 논리도 이해가 안가는 부분이 "평균 대비 분산"으로 설명력을 예상한 점인데요. 분산을 계산할 때 이미 평균을 고려하는데 왜 평균 대비 분산으로 비교를 하는지 잘 모르겠습니다. R의 SD가 2.1로 P의 SD인 10.1보다 작으니 식 (1)에서 $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$로부터 R의 s.e가 P의 s.e보다 크다고 생각하는데요(즉, P와 H가 오히려 R보다 swing이 크다고 생각합니다.)

A와 B는 서로 다른 주장이고 둘 중 어느 관점으로 답안을 작성해야할지 모르겠고, A,B 논리에 각각 의문을 제기해볼 수 있을 것 같아서 혼란스러운 상태입니다.

6. Lecture note 7&8: Question 8-(7)

대표님께서 Measurement error가 있어도 t값은 커질지 작아질지 알 수 없다고 강의시간에 말씀해주셨는데요. attenuation bias는 생겨도 variance가 커질지 작아질지 알 수 없기 때문이라고 알고 있었는데, DBDM Lecture note6 5p(그림3)를 보면 t값은 biased downwards로 또 나와 있어서 어떻게 정답을 작성해야 하는지 헷갈립니다.

2.3 Measurement error for t-statistics

(Mathematical derivation in this subsection is optional for MBA)
However, the t-statistic will be biased downwards. The t-ratio converges to

$$\frac{\operatorname{plim} t}{\sqrt{n}} = \frac{\operatorname{plim} \widehat{\beta}}{\sqrt{s}} = \frac{\lambda\beta}{\sqrt{\lambda x + \lambda(1-\lambda)\beta^2}}$$
$$= \sqrt{\lambda} \cdot \frac{\beta}{\sqrt{s + (1-\lambda)\beta^2}}$$

which is smaller than $\beta/\sqrt{s}$.

<그림 3>