04.03 분산 분석과 모형 성능

모형 성능 비교 :

- 1) R^2
- 2) R^2_{adj}
- 3) F-test
- 4) ANOVA
- 5) Log-likelihood
- 6) AIC, BIC

1) 분산 분석(ANOVA)

선형회귀 성능 평가

RSS : 잔차제곱합 크면 성능 안좋음

*문제점 : 데이터 갯수 많아지면 무조건 커짐 / 모형 간 단위가 서로 다르다면 RSS로 모형 간 성능 비교 불가

ANOVA(대안) : 잔차를 쓰긴 쓰지만, 기준 동일화 위해 정규화된 잔차를 사용 => 모형 간 선형회귀 모델 성능 비교분석 용이

[ANOVA]

1. 3가지 값 정의 (1page)

TSS : 종속변수값의 분산

ESS : 예측값의 분산

RSS : 잔차의 분산

TSS = ESS + RSS

2. 시사점

1) TSS > ESS > RSS
2) 성능이 좋은 모형일 수록 RSS는 작아진다 (TSS와 ESS가 같아짐)

3. 코드

regressionresult 객체

TSS => result.uncentered_tss

ESS => result.mse_model

RSS => result.ssr

2) 결정계수

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$
$$0 \leq R^2 \leq 1$$

상수항이 없는 모델의 경우, 기존의 R^2 을 사용하면 범위의 문제가 생김

상수항이 없는 모델인 경우, TSS를 평균 = 0으로 놓고 계산하는 것으로 변경

3) 분산분석표

ANOVA, F-test 결과를 함께 보여줌 (5page)

*ANOVA, F-test : 둘 다 모형의 성능을 표현

4) ANOVA, F-test 관계

ANOVA로 얻은 ESS, RSS 를 통해 F 분포를 따르는 통계량을 얻을 수 있다. (5page, a)

이 값을 F-test의 검정통계량으로 사용

코드 : sm.stats.anova_lm(result) (6page)

5) 결정계수, 상관계수

$R^2 = \frac{ESS}{TSS}$ = 상관계수(실제, 예측값의) (7page)

6) 상수항이 없는 모형의 경우

- 기본적으로 상수항 반드시 있어야 함 ==>> 그래야 절편을 갖고, 본래 데이터에 근사한 선형을 자유롭게 만들 수 있음 (8page, a)

- 모델에 w_0가 없다면?
  - TSS를 평균 = 0 으로 놓고 계산! ==>> TSS > ESS 가능해짐 (8page, b)

7) F 검정을 이용한 모형 비교

- 모형 비교 : 전체모형 vs 축소모형

  *축소모형 : 일부 feature 제거

  *코드 : sm.stats.anova_lm(model_reduced.fit(), model_full.fit())

$H_0 : w_2 = w_3 = 0$ <==> 모형 비교 (10page, a)

8) F 검정을 이용한 변수 중요도 비교

- 데이터를 제외해야 하는 경우 : 어느 feature가 y값에 영향을 덜 미쳐서, 제외해도 되는가?

*코드 : sm.stats.anova_lm(result_boston, type=2) (11page, 상단)

9) 조정 결정계수

- 11page 하단, 12page a

9) 정보량 규준

- AIC, BIC : 작을 수록 올바른 모형에 가까운 것(좋은 것) (12page, 하단)