Chapter06

선형회귀

l 회귀모형의 진단과 선별

FASTCAMPUS
ONLINE
금융공학/퀀트 l

강사. 장순용

I 키포인트

- 선형회귀 모형의 진단.
- t 검정을 적용한 회귀 계수의 유의성 확인.
- F 검정을 적용한 회귀 모형의 설명력 확인.
- 결정계수 R^2, MSE, VIF 등과 같은 진단 수치 확인.
- 정보량과 선형회귀 모형의 선별.

I 선형회귀 진단: t 검정

Question : 모형의 설명변수들은 통계적으로 의미 있나?

$\Rightarrow$ 개개의 회귀 계수에 대한 t 검정.

I 선형회귀 진단: t 검정

- 개개의 회귀 계수에 대한 양측검정 (t 검정)을 실행한다.

귀무가설 $\boldsymbol{H_0} : \beta_i = 0$

대립가설 $\boldsymbol{H_1} : \beta_i \neq 0$

$\Rightarrow$ t 검정 통계량과 p-값 사용

l 선형회귀 진단: t 검정

- 개개의 회귀 계수에 대한 양측검정 (t 검정)을 실행한다.

  귀무가설 $\boldsymbol{H_0} : \beta_i = 0$

  대립가설 $\boldsymbol{H_1} : \beta_i \neq 0$

  $\Rightarrow$ t 검정 통계량 $= \dfrac{\widehat{\beta_i}}{\beta_i \text{의 표준오차}}$

  $\Rightarrow$ p-값이 임계치 이하인 경우 (< 0.05), $\boldsymbol{H_0}$ 기각 후 $\boldsymbol{H_1}$채택.

  X_i을 모형에 포함시키는 것은 **정당**함.

I 선형회귀 진단: F 검정

Question : 회귀 모형은 설명력을 제공하나?

I 선형회귀 진단: F 검정

Question :

회귀모형의 독립변수 중 최소 한 개라도 종속변수를 설명하는 역할을 하고 있나?

I 선형회귀 진단: F 검정

- 모든 회귀 계수에 대한 F 검정을 실행한다.

귀무가설 $\boldsymbol{H_0}$: $\beta_1 = \beta_2 = \cdots = \beta_K = 0.$

대립가설 $\boldsymbol{H_1}$: 적어도 하나의 β_i 가 0과 다르다.

$\Rightarrow$ F 검정 통계량과 p-값 사용.

I 선형회귀 진단: F 검정

- 모든 회귀 계수에 대한 F 검정을 실행한다.

귀무가설 $\boldsymbol{H_0}$: $\beta_1 = \beta_2 = \cdots = \beta_K = 0.$

대립가설 $\boldsymbol{H_1}$: 적어도 하나의 β_i 가 0과 다르다.

$\Rightarrow$ F 검정 통계량 $= \dfrac{\text{설명할 수 있는 오류(분산)}}{\text{설명할 수 없는 오류(분산)}}$

$\Rightarrow$ p-값이 임계치 이하인 경우 ($<$ 0.05), $\boldsymbol{H_0}$ 기각 후 $\boldsymbol{H_1}$ 채택.

회귀 모형은 조금이라도 설명력 있음.

I 선형회귀 진단: 결정계수

- 결정계수 R^2 는 대표적인 진단 척도중의 하나이다.
- $0 < R^2 < 1$이며 R^2이 1에 가까울 수록 좋다.

$$R^2 = 1 - \frac{SSE}{SST}$$

$$\text{with } SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \text{ and } SST = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

I 선형회귀 진단: 결정계수

- 모형이 복잡해 질수록 R^2은 증가한다. ⇒ R^2만을 기준으로 모형을 만들면 과적합 현상이 쉽게 발생하니 주의한다.
- 독립변수가 **하나 뿐인** 경우에는 R^2 는 X와 Y 사이의 상관계수의 제곱과 값이 같다.

$$R^2 = Cor(X,Y)^2$$

I 선형회귀 진단: MSE, RMSE, MAE

- MSE와 RMSE는 예측값과 실제값 사이의 차이를 나타낸다.

  ⇒ 작을수록 좋다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{MSE}$$

- MAE도 MSE와 유사한 의미의 수치이다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

I 선형회귀 진단: VIF

- 다중공선성의 정도는 개개 설명변수의 VIF (Variance Inflation Factor)를 사용하여 가늠해 볼 수 있다.

  → VIF > 5 : 강한 다중공선성 존재.

  → VIF > 10 : 심각한 수준의 다중공선성 존재.

- VIF 수치가 큰 경우 모형 간추리기가 필요할 수 있다.

I 선형회귀 진단: VIF

- 개개 설명변수 X_i에 대한 VIF는 다음과 같은 방식으로 구한다.

1). 변수 X_i를 종속변수의 역할에 놓고 나머지 설명변수로 선형 회귀식을 만든다:

$$X_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{i-1} X_{i-1} + \beta_{i+1} X_{i+1} + \cdots + \varepsilon$$

2). 해당 결정계수 R_i^2를 사용하여 VIF_i를 계산한다:

$$VIF_i = \frac{1}{1 - R_i^2}$$

I 정보량과 모형 선별

- 정보량 (Information Criteria):

$$AIC = -2\frac{Log\ likelihood}{N} + 2\frac{p}{N}$$

$$BIC = -2\frac{Log\ likelihood}{N} + p\frac{Ln(N)}{N}$$

$$Log\ Likelihood = -\frac{N}{2}\left(1 + Ln(2\pi) + Ln\left(\frac{SSE}{N}\right)\right)$$

I 정보량과 모형 선별

- 정보량 (Information Criteria):

$$AIC = -2\frac{Log\ likelihood}{N} + 2\frac{p}{N}$$

$$BIC = -2\frac{Log\ likelihood}{N} + p\frac{Ln(N)}{N}$$

$\Rightarrow$ AIC (또는 BIC)를 최소화 하려고 한다.

$\Rightarrow$ AIC (또는 BIC)는 두개의 상반된 트렌드의 합이다.

I 정보량과 모형 선별

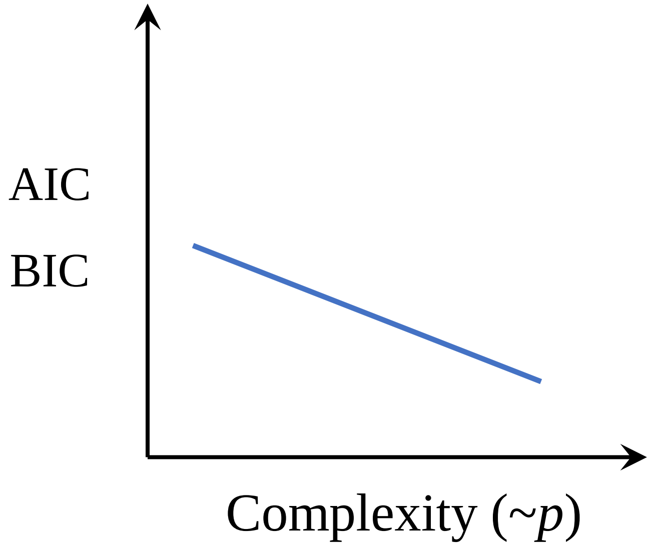
- 정보량 (Information Criteria):

$$AIC = -2\frac{Log\ likelihood}{N} + 2\frac{p}{N}$$

$$BIC = -2\frac{Log\ likelihood}{N} + p\frac{Ln(N)}{N}$$

$$\sim \ -Log\ likelihood$$

모형이 복잡할 수록 감소.



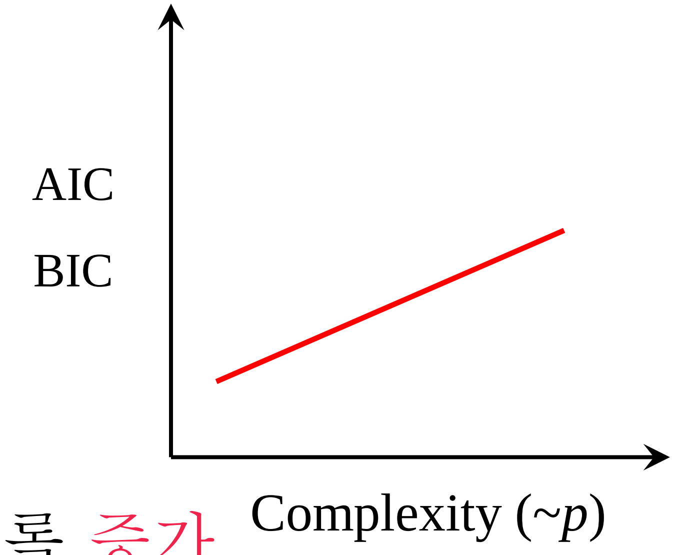
I 정보량과 모형 선별

- 정보량 (Information Criteria):

$$AIC = -2\frac{Log\ likelihood}{N} + 2\frac{p}{N}$$

$$BIC = -2\frac{Log\ likelihood}{N} + p\frac{Ln(N)}{N}$$

$\sim p$



모형이 복잡할 수록 증가.

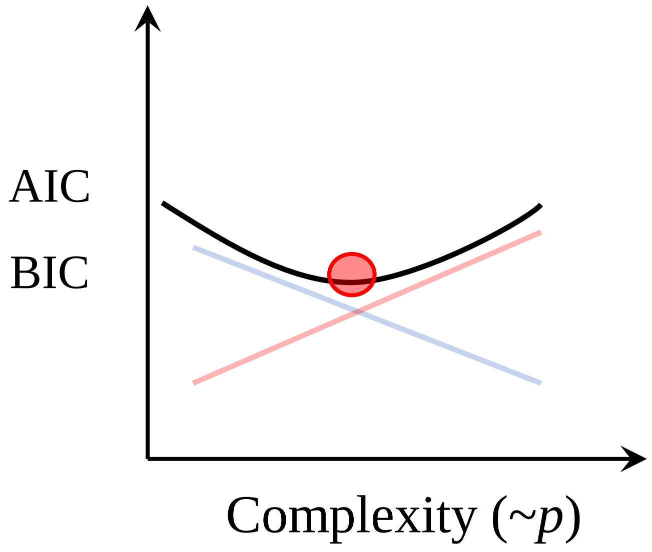
p는 모형의 파라미터의 수.

I 정보량과 모형 선별

- 정보량 (Information Criteria):

$$AIC = -2\frac{Log\ likelihood}{N} + 2\frac{p}{N}$$

$$BIC = -2\frac{Log\ likelihood}{N} + p\frac{Ln(N)}{N}$$



⇒ 합이 최소인 **최적점**이 있다.

I 회귀모형의 선별

- 회귀모형의 선별 방법:

  ⇒ R^2은 1에 가까워져야 한다.

  ⇒ AIC가 감소하는 방향으로 최적화 진행.

  ⇒ 모형이 잘못된 방향으로 변경되면, AIC는 감소하는 대신 증가한다.

Stop!

I 끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.