

05.04

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

조정 결정 계수와 함께 많이 쓰이는 모형 비교 기준은 최대 우도에 독립 변수의 갯수에 대한 손실(penalty)분을 반영하는 방법이다. 이를 정보량 기준(information criterion)이라고 하며 손실 가중치의 계산 법에 따라 AIC (Akaike Information Criterion)와 BIC (Bayesian Information Criterion) 두 가지를 사용한다.

AIC는 모형과 데이터의 확률 분포 사이의 Kullback-Leibler 수준을 가장 크게하기 위한 시도에서 나왔다. BIC는 데이터가 exponential family라는 가정하에 주어진 데이터에서 모형의 likelihood를 측정하기 위한 값에서 유도되었다. 둘 다 값이 작을 수록 올바른 모형에 가깝다.

$$AIC = -2 \log L + 2K$$

$$BIC = -2 \log L + K \log n$$

데이터 ↑ 인수를 (독립변수 수 ↑)
페널티 (BIC 값을 크게 만든다)

K ↑ 가 페널티!

→ log-likelihood

모형 성능 비교!

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	108.1
Date:	Wed, 30 Oct 2019	Prob (F-statistic):	6.72e-135
Time:	19:54:19	Log-Likelihood:	-1498.8
No. Observations:	506	AIC:	3026.
Df Residuals:	492	BIC:	3085.
Df Model:	13		
Covariance Type:	nonrobust		

K ↑ 로다
차감.

R-squared: ⇒ 모형 성능

OLS Adj. R-squared: ⇒ K ↑ 에 따른 가계변화 확인

F-statistic:] $H_0: R^2 = 0$
검정!

Log-Likelihood: ⇒ 클수록 좋다.

AIC:]
BIC:]
→ 0에
→ 작을수록 좋다.

* log L : log-likelihood
MLR ⇔ 가능성도 최대!
이 모형에서 최대의 가능성도 의미!

값을 수를 줄다.

(-1400 vs -400)

↓ 이게 더 좋은 것

AIC/BIC

: 값이 작을수록 좋다.

1) 분산 분석(ANOVA)

선형회귀 성능 평가

RSS : 잔차제곱합 크면 성능 안 좋음

*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 선형 회귀 모델 성능 비교분석 용이

[ANOVA]

1. 3가지 값 정의 (1page)

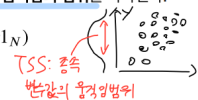
TSS : 종속변수값의 분산

ESS : 예측값의 분산

RSS : 잔차의 분산

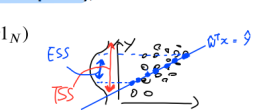
$TSS = ESS + RSS$

종속 변수 y 의 분산(샘플의 갯수로 나누지 않았으므로 정확하게는 분산이 아니지만 여기에서는 분산이라는 용어를 사용하자)을 나타내는 **TSS(total sum of square)**라는 값을 정의한다. **TSS는 종속변수값의 움직임의 범위를** 나타낸다.

$$TSS = \sum_{i=1}^N (y_i - \bar{y})^2 = (y - \bar{y}1_N)^T (y - \bar{y}1_N)$$


위 식에서 $\bar{y}1_N$ 는 \bar{y} 이라는 스칼라가 N 번 반복된 브로드캐스팅 벡터다.

마찬가지로 회귀 분석에 의해 예측한 값 \hat{y} 의 분산을 나타내는 **ESS(explained sum of squares)**,

$$ESS = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 = (\hat{y} - \bar{y}1_N)^T (\hat{y} - \bar{y}1_N)$$


잔차 e 의 분산을 나타내는 **RSS(residual sum of squares)**도 정의할 수 있다.

$$RSS = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = e^T e$$

RSS : 오(잔차)²
***잔차 : $y - \hat{y}$**

위 식에서 \hat{y} 는 모형 예측값 \hat{y} 의 평균이다.

또한 **ESS**는 모형에서 나온 예측값의 움직임의 범위, **RSS**는 잔차의 움직임의 범위, 즉 오차의 크기를 뜻한다고 볼 수 있다.

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

분산 분석표 : ANOVA + F test \Rightarrow 둘 다 모형 성능 표현

분산 분석의 결과는 보통 다음과 같은 분산 분석표를 사용하여 표시한다. 아래의 표에서 N 은 데이터의 갯수, K 는 모수의 갯수를 뜻한다.

source	degree of freedom	sum of square	mean square	F test-statistics	p-value
Regression	$K - 1$	ESS	$s^2 = \frac{ESS}{K - 1}$	$F = \frac{s^2}{s^2}$	p-value
Residual	$N - K$	RSS	$s^2 = \frac{RSS}{N - K}$		
Total	$N - 1$	TSS	$s^2 = \frac{TSS}{N - 1}$		
R^2		ESS/TSS			

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

이 때 \hat{w} 값은 기대값이 0인 정규 분포에서 나온 표본이므로 예측값 $\hat{y} = \hat{w}^T x$ 는 정규 분포의 선형 조합이라서 마찬가지로 정규 분포를 따른다. 그리고 잔차(residual)는 오차(disturbance)의 선형 변환으로 정규 분포를 따르므로 ESS와 RSS의 비율은 F 분포를 따른다.

$$\frac{ESS}{K - 1} \div \frac{RSS}{N - K} \sim F(K - 1, N - K)$$

따라서 이 값을 회귀 분석 F-검정의 검정통계량으로 사용할 수 있다.

$$H_0: R^2 = 0$$

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

```
In [4]:
sm.stats.anova_lm(result)
```

```
Out[4]:
```

	df	sum_sq	mean_sq	F	PR(>F)
X	1.0	188589.613492	188589.613492	179.863766	6.601482e-24
Residual	98.0	102754.337551	1048.513648	NaN	NaN

$H_0: R^2 = 0$ 에 대한 F검
(H_0 기각!)

5) 결정계수, 상관계수

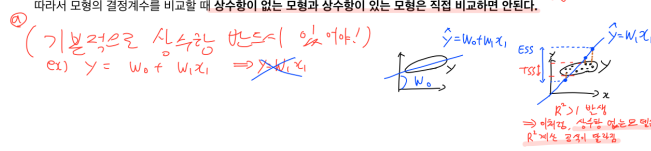
$$R^2 = \frac{ESS}{TSS} = \text{상관계수 (실제, 예측값의)} \quad (7\text{page})$$

$$\rho_{\hat{y}\hat{x}} = R^2 = \frac{ESS}{TSS}$$

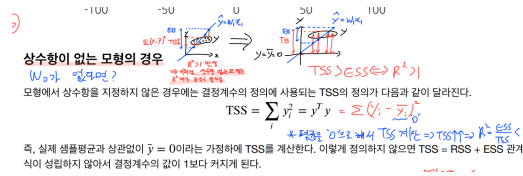
(실제, 예측의 상관계수 = 모형 성능)
상관계수 \uparrow 인수록 당연히 정확도 \uparrow RSS $\downarrow \Leftrightarrow R^2 \uparrow$

6) 상수항이 없는 모형의 경우

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



- 모형에 w_0 가 없다면?
 - TSS를 평균 = 0 으로 놓고 계산! ==> TSS > ESS 가능해짐 (8page, b)



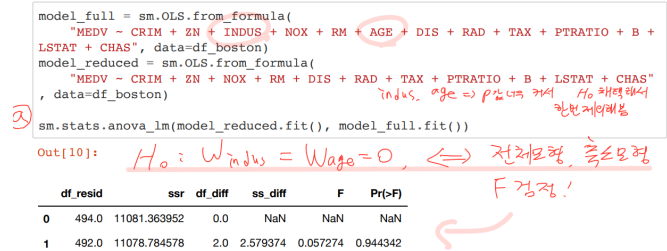
7) F 검정을 이용한 모형 비교

- 모형 비교 : 전체모형 vs 축소모형

*축소모형 : 일부 feature 제거

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

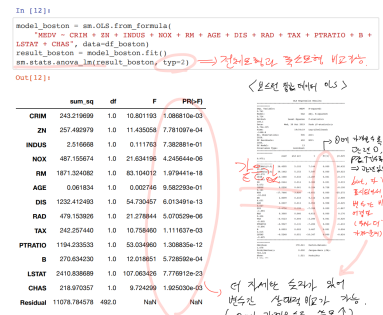
$H_0 : w_2 = w_3 = 0 \iff$ 모형 비교 (10page, a)



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)`



9) 조정 결정계수

조정 결정계수
 feature를 1개라도 더 넣으면, 성능이 좋아질까? 항상 그렇다.

R^2 vs $TSS > ESS$
 $\therefore \frac{ESS + 0}{TSS + 0} < \frac{ESS}{TSS}$

이러한 독립 변수 추가 효과를 상쇄시키기 위한 다양한 기준들이 제시되었다. 그 중 하나가 다음과 같이 독립 변수의 갯수 K에 따라 결정계수의 값을 조정하는 조정 결정계수이다

$$R^2_{adj} = 1 - \frac{n-1}{n-K} (1 - R^2) = \frac{(n-1)R^2 + 1 - K}{n-K}$$

여기서 n 은 데이터 개수, K 은 feature 개수 (feature, 변수 수)
 데이터가 많을수록 R^2 은 높아진다
 feature가 많을수록 R^2 은 낮아진다

9) 정보량 기준

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

조정 결정계수와 함께 많이 쓰이는 모형 비교 기준은 최대 우도에 독립 변수의 갯수에 대한 손실(penalty)분을 반영하는 방법이다. 이를 정보량 기준(Information criterion)이라고 하며 손실 가중치의 계산 법에 따라 AIC (Akaike Information Criterion)와 BIC (Bayesian Information Criterion) 두 가지를 사용한다.

AIC는 모형과 데이터의 확률 분포 사이의 Kullback-Leibler 수준을 가장 크게하기 위한 시도에서 나왔다. BIC는 데이터가 exponential family라는 가정하에 주어진 데이터에서 모형의 likelihood를 측정하기 위한 값에서 유도되었다. 둘 다 값이 작을 수록 올바른 모형에 가깝다.

$$AIC = -2 \log L + 2K$$

$$BIC = -2 \log L + K \log n$$

데이터가 많을수록 (독립변수 수)가 많아지면 BIC가 커진다
 K가 커지면 AIC가 커진다

