**< 다중선형회귀분석 회귀계수 구하기 >**

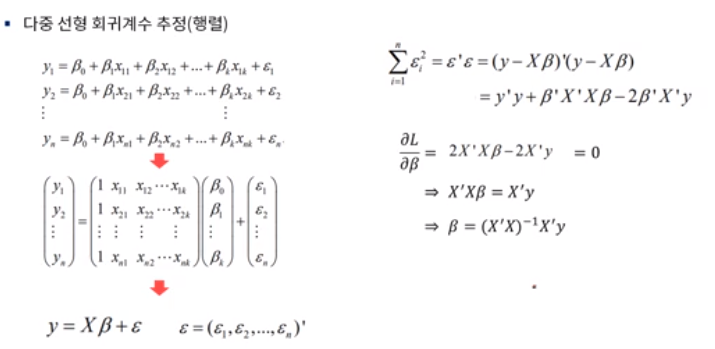
단순선형회귀분석 : 변수가 1개인 경우

Y\_hat = beta0 + beta1\*x

다중선형회귀분석 : 변수가 여러 개인 경우

Y\_hat = beta0+beta1\*x1+beta2\*x2 + …. + beta(p)\*x(p)

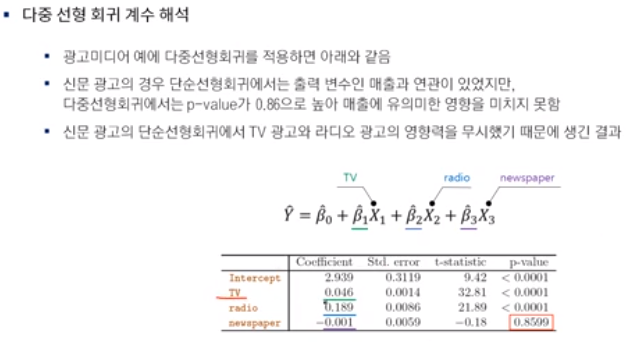
다중 선형 회귀계수 추정 : 회귀계수를 추정하는 것은 SSE를 최소화하는 방향으로 계산



Beta식을 결과로 보자 : 역행렬이 안구해지면 beta값이 의미가 없다

역행렬이 구해지지 않는 조건을 알아야한다.

**< 회귀계수와 모델의 검증 >**



다른 변수가 고정되어있을 때 TV가 1단위 증가할 때 매출액이 0.046단위 증가한다 . 그때 p-value가 매우 낮기 때문에 유의미한 변수로 볼 수 있다.

단순선형회귀분석에서는 하나의 변수에 대해 가설을 검증하면 모델도 검증이 되지만 다중선형회귀분석에서는 다른 변수들에대해서 검증이 되어야 모델을 검증할 수 있다.

다중 선형 회귀 모델 검정

* 귀무가설 : B1=B2=B3 … Bp=0
* 대립가설 : 하나의 회귀계수라도 0이 아니다. ( 매우 귀무가설을 기각하기 쉽다 ) 즉 설명력이있는 변수가 존재한다.
* F검정을 통해서 검정 F-statistics(MSR/MSE) 커질수록 p-value가 작아진다

??????????????????

변수가 추가되면 추가 될수록 귀무가설을 기각하기 쉽다 변수가 추가되면 MSR부분은 커지고 MSE부분은 작아진다 -> 따라서 F통계량은 커지게되고 p-value는 작아진다.

R-squared = SSR/SST 도 마찬가지로 커진다.

결국 변수가 매우많을때는 F통계량과 R^2를 마냥 믿을 수는 없다.

변수가 많아지면 자연스럽게 다중공선성 현상이 발생한다.

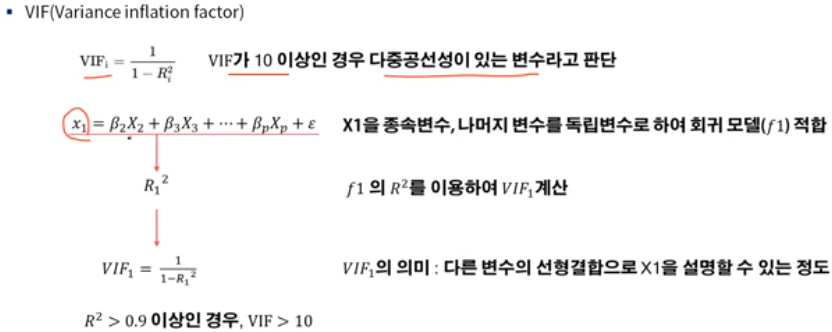
**< 다중공선성 >**

Y를 설명하는 x의 변동성이 중복되다 보니 겹치는 변동성에 대해서 중복으로 가져가지 못함.

* 다중공선성을 진단하는 방법

VIF(variance inflation factor) , correlation matrix(상관행렬)

하나의 변수를 예상할 수 있다면 사용하지 않아도 되지않는가



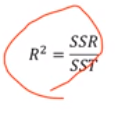
굳이 R^2가 크면 다중공선성이 있다고 판단하지 왜 VIF를 사용하는가?

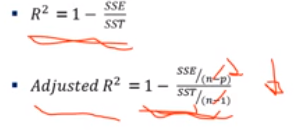
* 증가하는 R^2에 대해서 증가하는 VIF의 가중치를 구하기 위해 이런 식을 잡음 ex) R^2 = 0.9이면 VIF=10 / R^2 = 0.95이면 VIF=20

다중공선성을 근본적으로 해결하는 방법은 아직 없다. 최근 머신러닝 기법들은 중요 특징을 뽑는 알고리즘이 있다. 그러나 사람이 판단해서 변수를 제외하고 넣어주는 경우와 모든 변수를 넣는경우는 매우 다를 수 있다. ( 머신러닝 알고리즘이 최대한 많은 변수들을 판단하고 가지고가려하기 떄문에 정확도에 문제가 생길수도있다. )

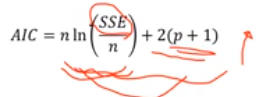
* 따라서, 중요변수 100개를 넣는 것이 모든변수 10000개를 넣는것보다 무조건 좋다

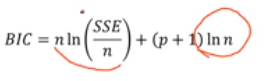
**< 회귀모델의 성능지표 >**





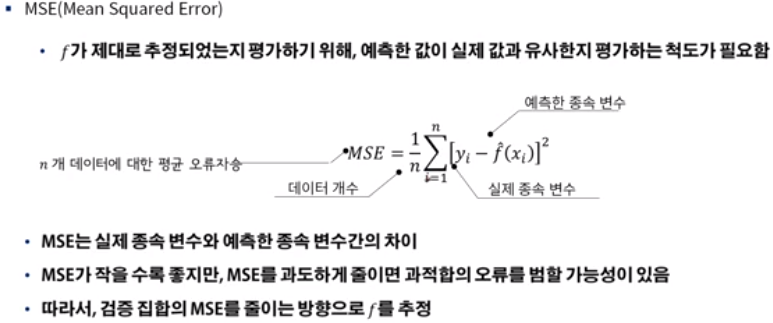
R^2는 변수의 수가 많아질 수 록 의미없어질 수 있으므로 변수의 개수로 패널티를 줌





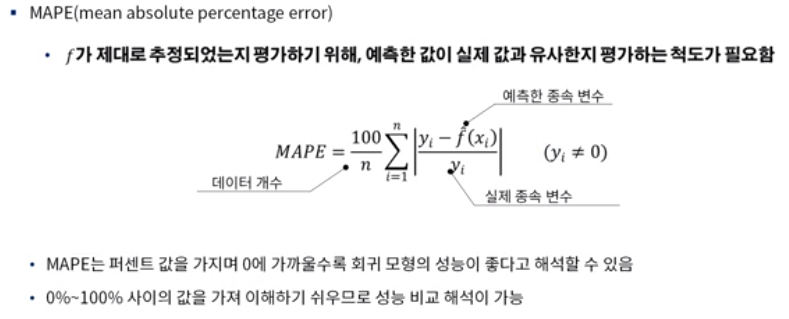
AIC의 단점은 표본 n이 커질 때 부정확해짐 따라서, 이를 보정

**< 모형의 성능 지표 > 주로 사용하게 될 성능지표( ML에서 자주 사용 )**



MSE의 단점은 단위가 없음. 따라서 MSE가 0.5라고 하면 이게 잘나온건지 모름.

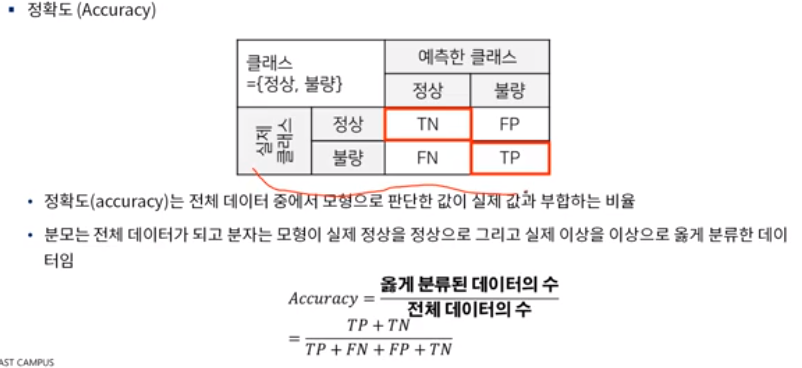
여러 모델과의MSE를 비교해보면서 무엇이 좋은 모델인지 상대적으로 파악은 가능함



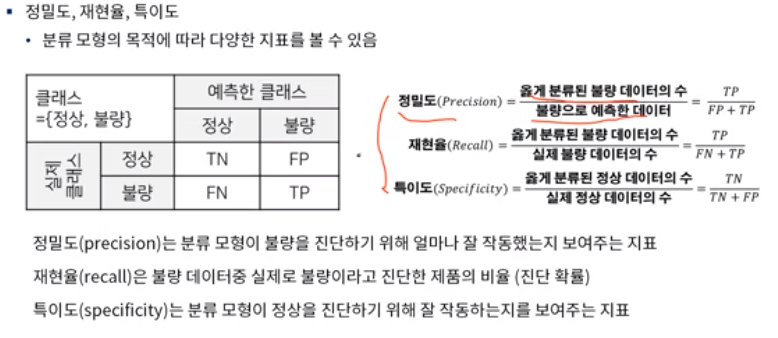
MSE와는 다르게 결과가 퍼센트값으로 나오기 때문에 척도로서 사용가능하다

MSE나 MAPE이나 상황에 따라 다르게 쓰이지만 보통 MSE가 쓰인다 ( Y가 연속형이나 실수값일 때 즉, 회귀모델일때 )

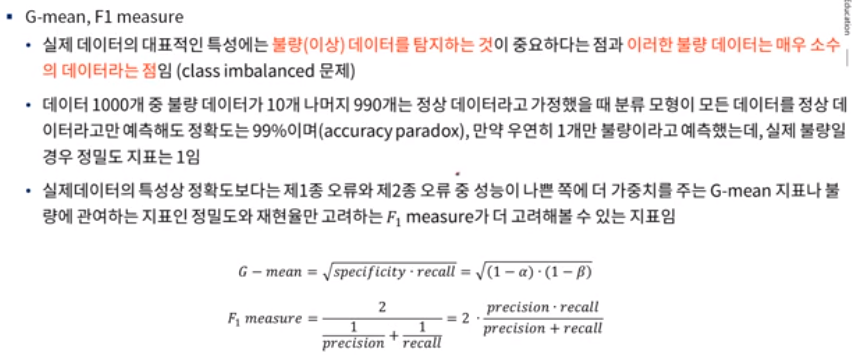
Classification에서는 Accuracy를 사용한다

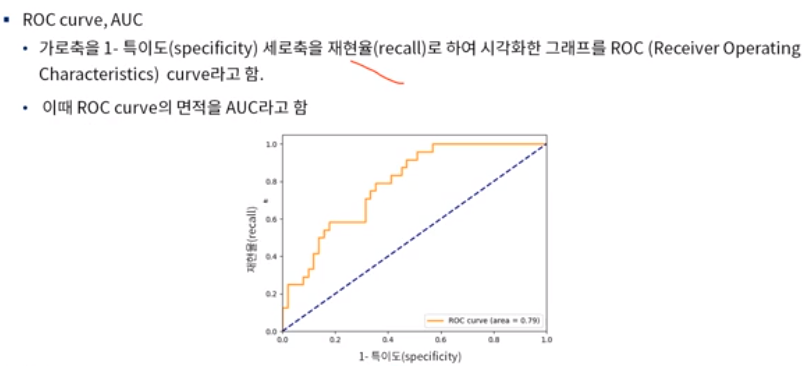


그러나 자료의 비율이 밸런스하지 않을경우, accuracy가 무의미해질 수 있기 때문에 다른 성능지표를 봐야한다

****

**자주 사용하는 지표 클래스예측모델에서는 accuracy는 별로안쓰고 F1많이 쓴다**





모델이 쓸모없으면 직선에 붙어있다. 위쪽으로 볼록할수록 좋은 모델이다. AUC는 (0~1 사이 값을갖고 1에 가까울수록 좋은 모델이다