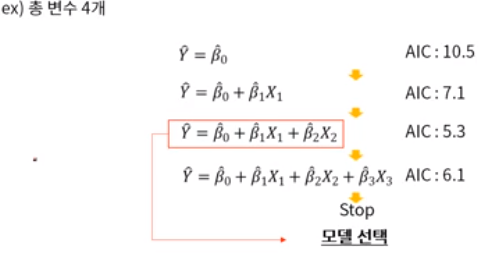
**< 다중선형회귀분석 변수선택법 >**

변수선택

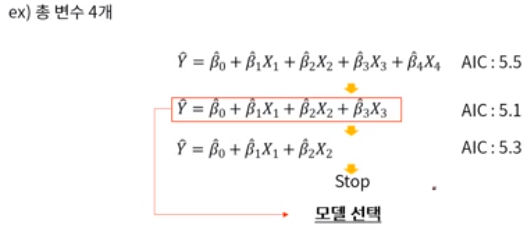
* 변수가 여러 개일때 최적의 변수 조합을 찾는 기법
* 변수가 p개일 때 변수의 총 조합은 2^p개로 너무 많다
* 총 변수들의 조합 중 최적의 조합을 찾기 위한 차선의 방법(optimal은 아니다. Optimal은 다해봐야 최선이다)

1. Feedforward Selection : 변수를 추가해가며 성능지표를 비교해가는 방법

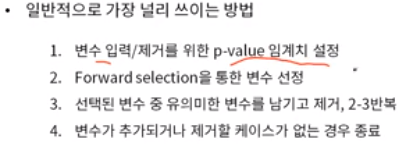


AIC가 낮았다가 높아지면 즉, 최소값이 나오면 STOP 모든변수의 조합을 고려하진 않지만 어느정도 나옴, 빠르다 / AIC가 낮으면 낮을수록 좋음

1. Backward Elimination : 변수를 제거해가며 성능지표를 비교해가는 방법



1. Stepwise :
   1. 가장 유익한 변수를 추가하거나 유의하지 않은 변수를 제거하는 방법
   2. Feedforward와 Backward를 섞은 기법

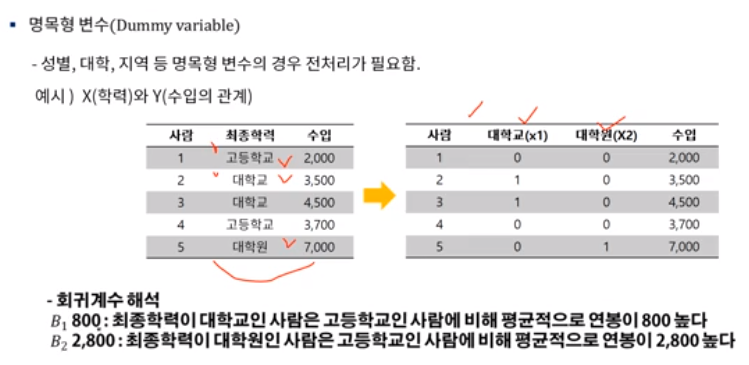
****

**< 교호작용 > 참고만 할정도로 이해**

교호작용 (interaction term)

* 변수간의 시너지 효과 x1,x2는 y에 영향을 끼치지않지만 결합하면 중요한 영향을 끼칠수있음
* 교호작용은 일반적으로 도메인지식에 근거하여 추가하여야한다 ( 정확한 근거가없으면 잘 쓰지 않는다 )

명목형 변수( dummy variable )

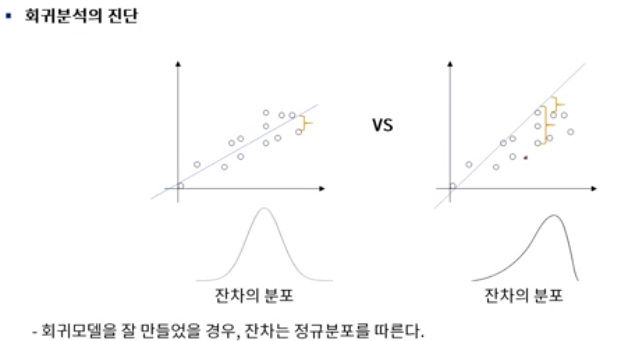


‘원핫인코딩’을 통해서 많은 변수들을 줄일 수 있다. 그 때 어떻게 해석을 하는가? -> 회귀계수 해석 예시부분 참고하기

**< 회귀분석의 진단 > 최적의 모델 만들고 정확도측정하고, 이 과정에 대해서 맞는 모델인가를 판단 이 과정이 있어야 그 모델이 조금 더 좋은모델일지 아닐지를 판단할 수 있다. 통계학에선 매우 중요한 단계 \***

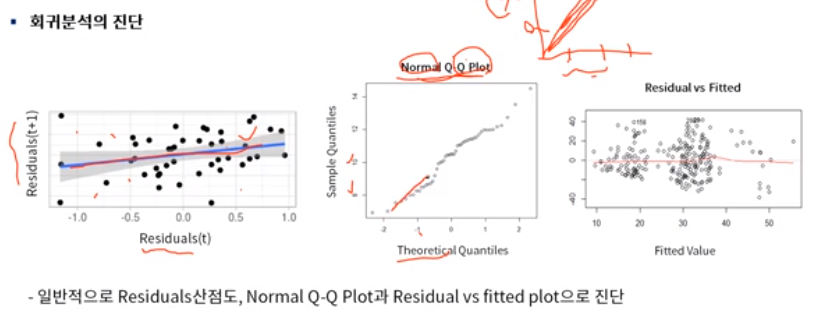
* 회귀분석에서는 잔차에 대한 세가지 가정이 존재: 정규성, 독립성, 등분산성
* 세가지 가정을 만족할 시 잘 만들어진 회귀모델이라 판단.

왜 이 세가지 가정을 만족해야하는가?

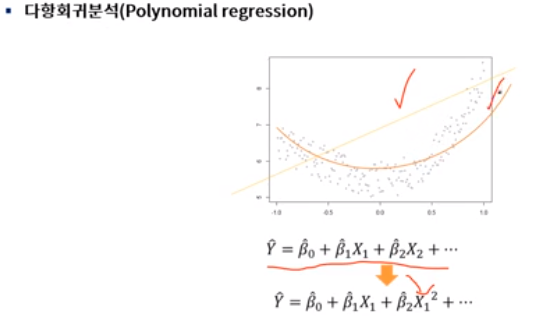
* 

회귀직선을 잘 찾으면 정규분포를 따른다.

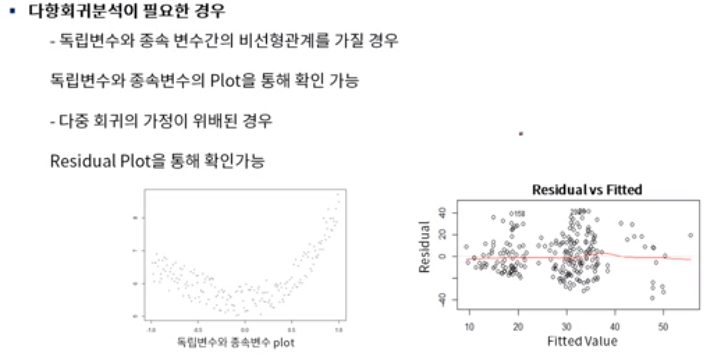
다중회귀선형분석에서는 시각화해서 잘 맞는지 확인을 하기 어렵기 때문에 아래의 방법을 통해 판단한다.



**< 다항회귀분석 >**2차항을 넣은 비 선형 회귀분석이다.



X,y가 비선형 관계일 때 사용한다. 일반적으로 데이터만 보고 알긴 힘들기 때문에 위의 모델의검증(잔차검증)을 하고나서 개선된 모델을 찾기 위해 시도해볼 수 있다.



X1과 Y와의 관계가 2차식과 같은 비선형이라면 2차항만 고려하는 것이 일반적. 3차항 이상은 거의 없음 다중공선성이 심해지기도 하고 overfitting이 심해진다. 비선형에서 항을 추가하는 것은 매우 잘 고민해야한다.