R 교육 세미나

ToBig's 10기 황이은

Liner Regression

선형회귀분석 보충 자료

선형회귀모형에 관하여

Unit 01 | 회귀분석 보충자료

회귀분석…… 왜 쓰는 건데?

1. 독립변수(X)와 종속변수(Y)가 관련(상관)이 있는가?

2. 독립변수(X)를 이용해 Y를 예측할 수 있는가?
*독립변수, 종속변수가 모두 연속형일 때 회귀분석을 사용할 수 있다.

Unit 01 | 회귀분석 보충자료

회귀분석 예시 – 실제 상황

집값

지하철과의 거리

주차장의 유무

도심과 떨어진 정도

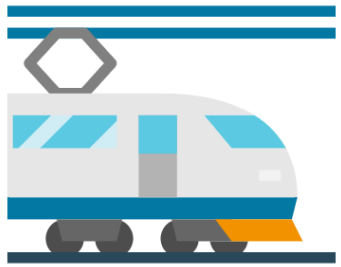
방의 개수

편의시설 개수

.....

Unit 01 | 회귀분석 보충자료

회귀분석 예시 – 데이터



우리가 알고 있는 것1 (독립변수):
지하철과의 거리

우리가 알고 있는 것 2(독립변수):
방의 개수

우리가 알고 싶은 것(종속변수):
집값

Unit 01 | 회귀분석 보충자료

회귀분석 예시 – 식으로 표현

$$y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$$

X_1 Input으로 들어오는 '지하철과의 거리'

X_2 Input으로 들어오는 '방의 개수'

y 집값

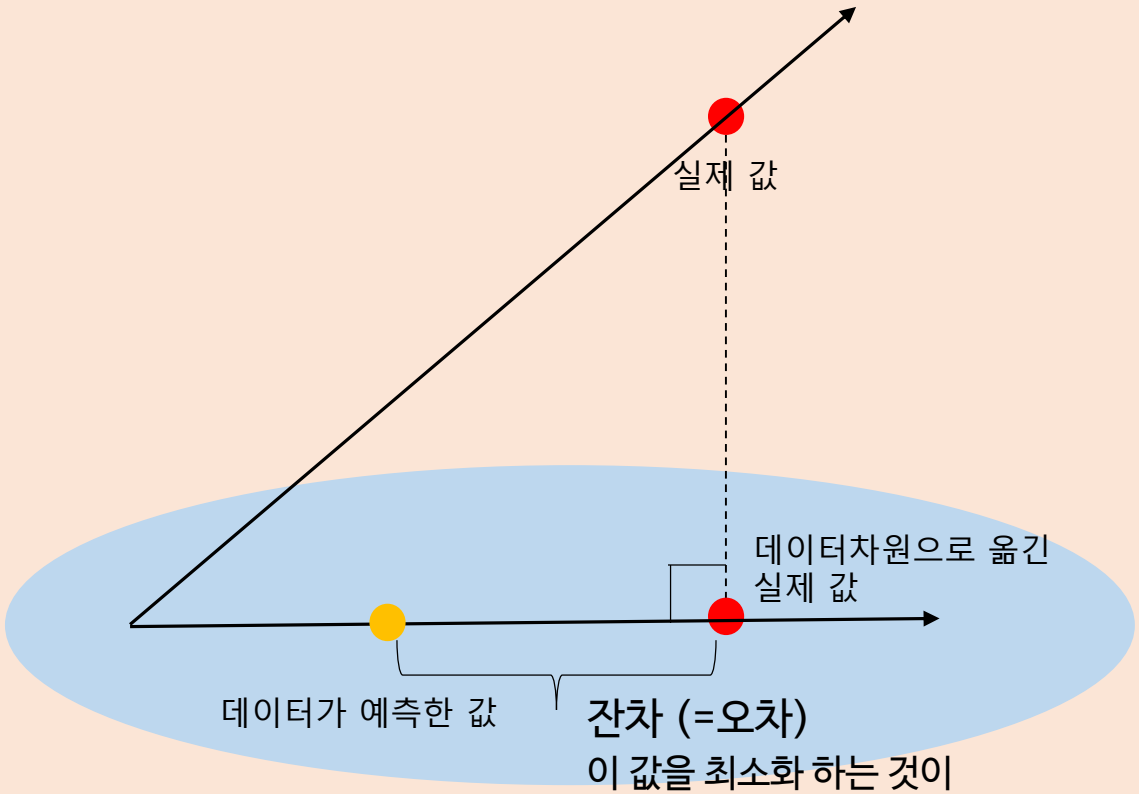
b_0 b_1 b_2

이 값들을 적절히 조정함으로써
(회귀계수) y를 예측 -〉 OLS

다른 관점으로 보면,
각 변수에 대한 가중치라고도 볼 수 있음

Unit 01 | 회귀분석 보충자료

직관적으로 이해하자!



실제 차원
데이터의 차원보다 고차원.
우리가 알 수 없는 세상

데이터의 차원
표현할 수 있는 차원이
유한하기 때문에 실제값과
차이가 발생!

P-value에 관하여

Unit 01 | 회귀분석 보충자료

P-value는 뭘까??

P-value를 만나보기 전에, 귀무가설과 대립가설에 대해 알아보자

귀무 가설(null hypothesis, 기호 H0) 또는 영 가설은 통계학에서 처음부터 버릴 것을 예상하는 가설이다.

대립 가설(alternative hypothesis, 기호 H1) 또는 연구 가설 또는 유지 가설은
귀무가설에 대립하는 명제이다. 보통, 모집단에서 독립변수와 결과변수 사이에 어떤 특정한 관련이 있다는 꼴이다.

Unit 01 | 회귀분석 보충자료

P-value는 뭘까??

선형회귀에서의 귀무가설과 대립가설

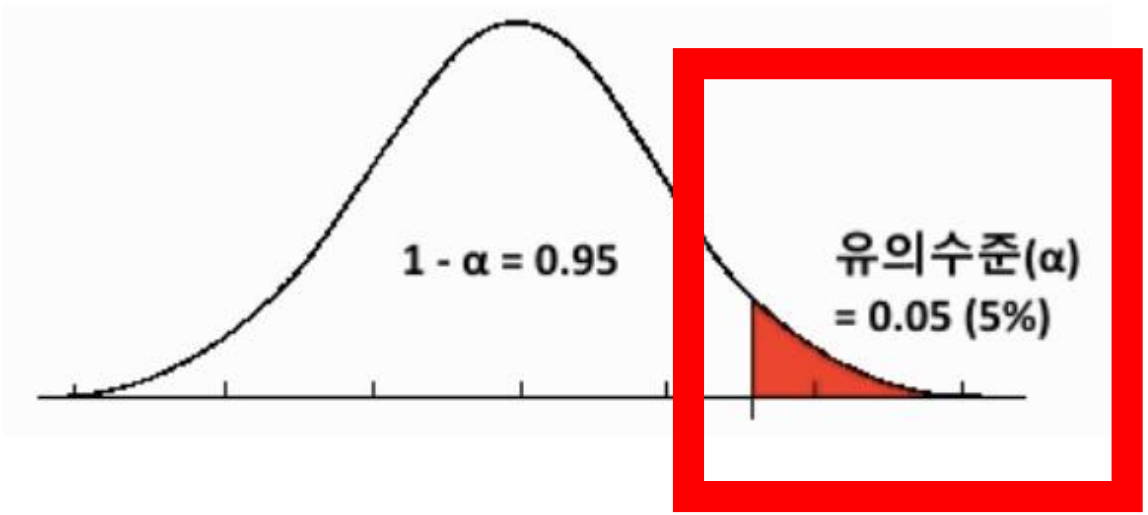
귀무가설 $H_0: b_1 = 0$

즉, '지하철과의 거리는 집값에 영향을 미치지 않는다'고 가정!!!!

대립가설 $H_1: b_1 \neq 0$

Unit 01 | 회귀분석 보충자료

P-value는 뭘까??



만약 p-value(=유의수준)가 0.05이하이면,
귀무가설을 기각하고 대립가설을 채택!

즉, 귀무가설이 빨간색 네모쳐진 곳에 위치한다는 뜻임!
이는, 신뢰구간 95%를 벗어난 곳
= 귀무가설이 95%신뢰구간안에 안 들어간다.

따라서, 귀무가설
(독립변수가 종속변수에 영향을 미치지 않는다)
신뢰할 수 없는 수준에 있음
-> 귀무가설을 기각하고 대립가설
(독립변수가 종속변수에 영향을 미친다)를 채택!

Unit 01 | 회귀분석 보충자료

선형회귀모형에서 변수 선정

전진선택법, 후진제거법, 단계적 선택법

Ex) 변수가 100개인 데이터가 있다고 가정
-〉 모두다 쓸거야???
독립변수가 종속변수에 영향을 미치지 않는 얘들(=p-value가 0.05 이상)은 필요 없잖아~
오히려 노이즈를 만듬(=설명력을 떨어뜨림)

★ 우린, 독립변수가 종속변수에 영향을 미치는 변수들만 보고싶어 ★
=p-value가 0.05이하인 변수들

Unit 01 | 회귀분석 보충자료

선형회귀모형에서 변수 선정

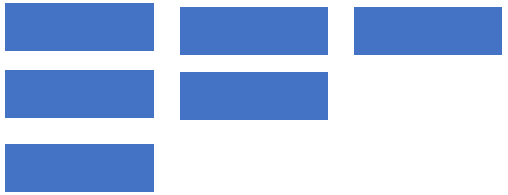
전진선택법, 후진제거법, 단계적 선택법

변수 1

전진선택법

P-value값이 낮은(=유의한)
변수를 하나씩 추가

전체 변수들



후진제거법

모든 변수를 넣고
P-value값을 기준으로 하나씩 뺌

Unit 01 | 회귀분석 보충자료

선형회귀모형에서 변수 선정

전진선택법, 후진제거법, 단계적 선택법

단계적 선택법

전진선택법과 후진제거법의 혼합형태!

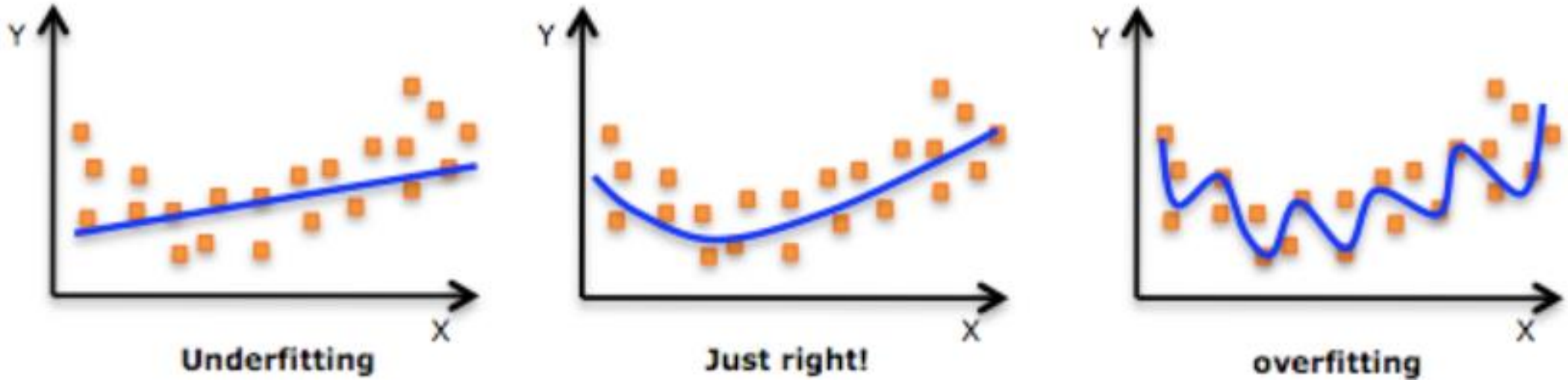
처음엔 전진선택법으로 한 변수를 추가한 뒤, 제거했다가 추가했다가를 반복하면서 AIC값이 낮아지는 변수를 선택하는 방법

-〉 시간이 오래걸림

기타 보충 자료

Unit 01 | 회귀분석 보충자료

오버피팅(과적합)과 언더피팅



오버피팅(과적합)과 언더피팅

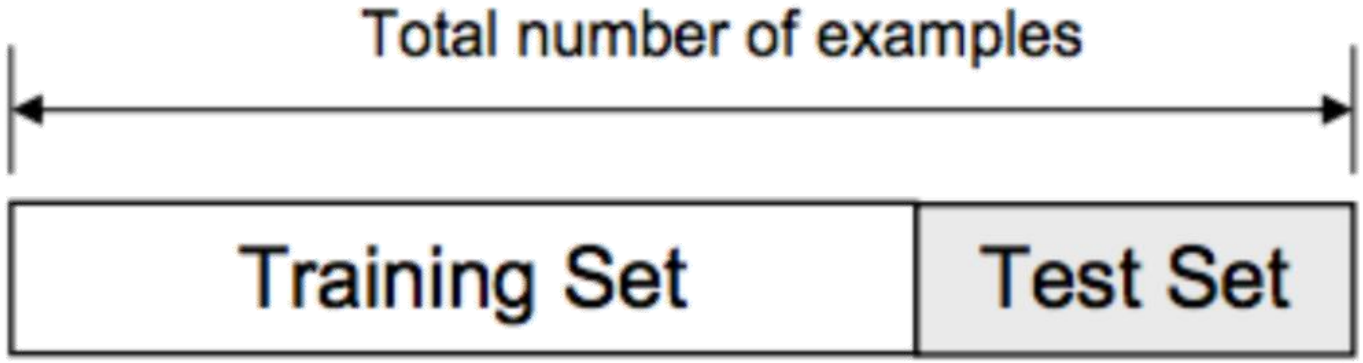
왜 오버피팅(과적합)이 문제가 될까???

오버피팅이 되면, 새로운 데이터가 들어왔을 때 이를 잘 예측하지 못함ㅠㅠ

우린, 과거에서 패턴을 찾을 찾는게 목적이 아님!
새로운 데이터(미래)가 들어와도 잘 작동하는 모델을 만들고 싶어!!!

Unit 01 | 회귀분석 보충자료

Train, Test, Validation 구분



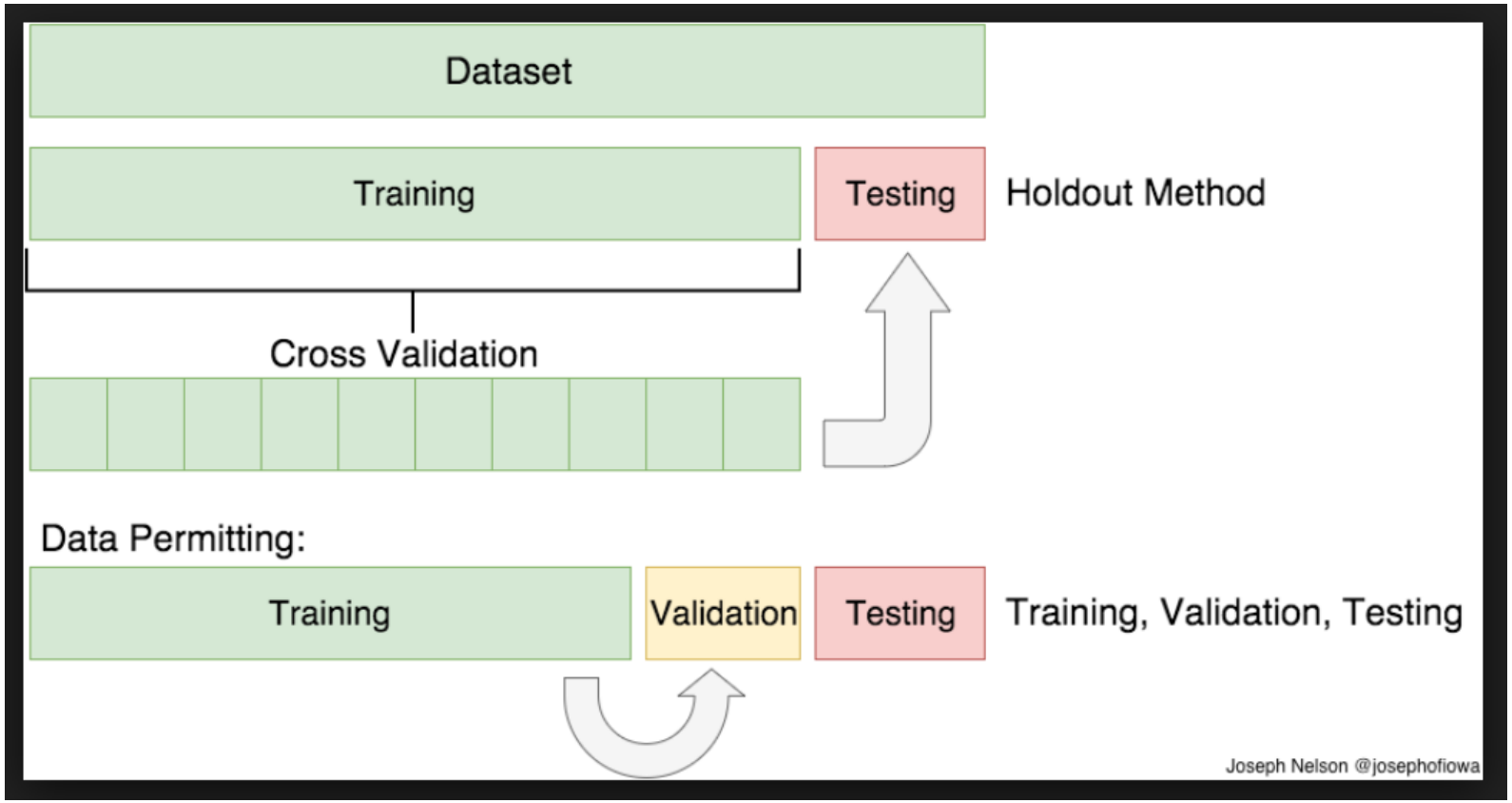
전통적인 방법: 7:3으로 train과 test를 분리함

하지만, 요즘같이 BIG데이터 시대엔
9:1도 충분하다! >> a.k.a 분석자 주관쓰~
예를들어 총 데이터가 십만개라고 할때, 9:1로 나누면 test데이터가 만개 확보 됨!

단!!!!!!!!!!train과 test의 분포는 같아야함!!!!

Unit 01 | 회귀분석 보충자료

Train, Test, Validation 구분



Unit 01 | 회귀분석 보충자료

실제로 한번 해보자!

R을 통해 실제 회귀분석을 해보겠습니다.

*데이터: 배틀그라운드 유저 데이터

*종속변수: 승률(winPlacePer)

*PUBG.r 파일을 실행시켜주세요

들어주셔서 감사합니다.