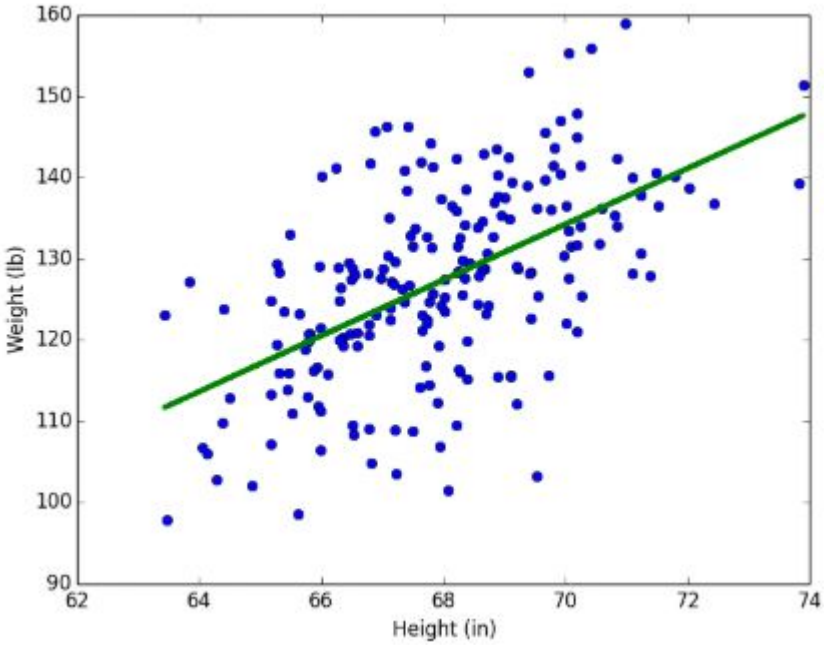
# Linear Regression(선형 회귀)

## 1.설명

데이터를 놓고 그걸 가장 잘 설명할 수 있는 선을 찾는 분석방법.




출처 : (https://hleecaster.com/ml-linear-regression-concept/)

예를 들어 위 그림처럼 키와 몸무게를 놓고 가장 잘 설명할 선을 그어놓으면 특정 사람의 키를 바탕으로 몸무게를 예측할 수 있다.다만, 이 값은 어디까지나 근사치이다.

> y = ax+b (단일 선형 회귀) y = a0 + a1x1 + a2x2 + ... anxn (다중 선형 회귀 -다양한 매개변수로 예측값을 선형적으로 예측)

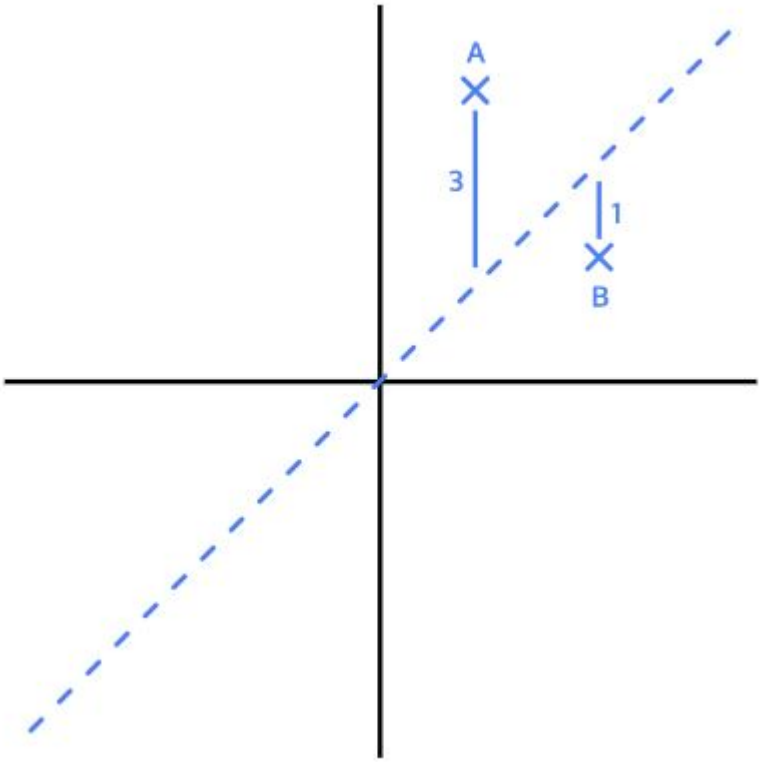
에서 우리가 가진 데이터를 가장 잘 설명할 수 있는 최고의 기울기 a와 y절편 b를 구하는것이 선형 회귀 분석의 목적이고, 이를 토대로 임의의 데이터의 예측을 하는 것이다.

---

## 2.특징

### 2-1.손실

데이터가 모두 일차원 선위에 존재하는 선형 데이터가 아닌 경우 모두 오차가 발생한다.




출처 : (https://hleecaster.com/ml-linear-regression-concept/)
위의 그림처럼 A와 B는 각각 3과 1만큼의 오차가 발생했다. 그럼 우리가 그은 선을 이용한 예측 값과 실측 값 사이의 오차를 측정하기위해서는 모든 오차(3이나 1)에 제곱을 해주어야 한다. A오차는 9이고 B는 1이 된다. 이러한 방식을 평균 제곱 오차(Mean Square Error, MES)라고 한다.

**선형 회귀 모델의 목표는 모든 데이터로부터 나타나는 오차의 평균을 최소화할 수 있는 최적의 기울기와 절편을 찾는 것이다.**

## 2-2.경사하강법(Gradient Descent)

선형회귀에서 최적화 알고리즘(최적의 모델을 찾는 것)으로 경사하강법을 사용한다. 단순히 생각하면 비용함수를 최소로 만드는 기울기와 y절편을 찾는 것이다.

---

## 2-3. 장단점

- 장점 : 해석이 간단, 통계적 유의성 쉽게 검정 가능 .
- 단점 :선형 모델이라는 가정이 근간이기때문에 실제 모델이 선형적이지 않는경우 모델의 오차가 크다. 예측 정확도의 경우 좋지 않을 가능성이 크다.특이점에 영향을 받을 가능성이 크므로 사용을 자제

## 적합한 사례

- 나이,성별,bmi,아이,흡연 여부 등 다양한 조건에 따른 다중선형회귀를 통한 의료비 예측
- 광고 지출 증가에 따른 판매 영향 예측