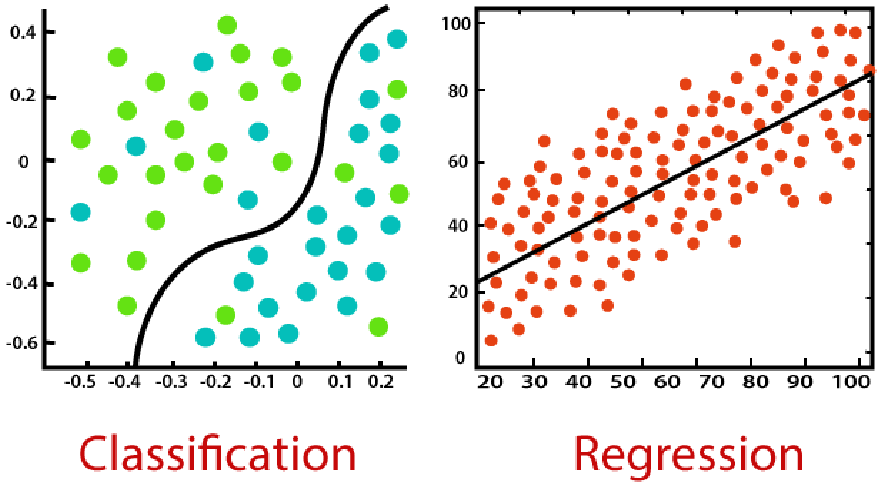
2. Regression

회귀분석

도입

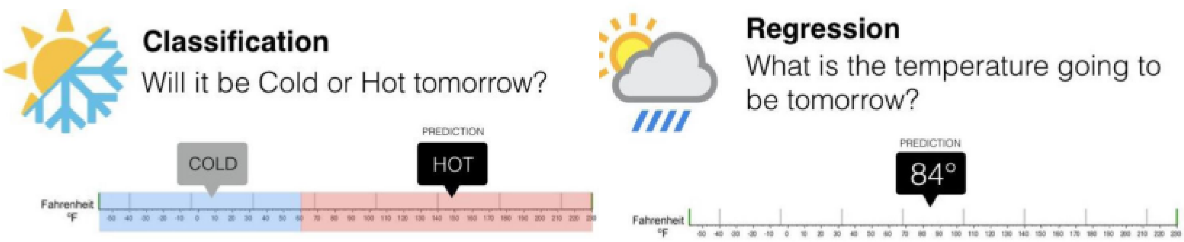
- (다변량) 데이터



  - 연속형 → 수치 예측(특정값)
  - 범주형 → 범주 예측(분류)

Classification vs Regression

- 분류(Classification):
  - 범주형 클래스 레이블(이산형 또는 명목형)을 예측
  - 클래스 레이블을 가진 훈련 세트를 학습하여 모델을 구축
  - 구축된 모델을 사용하여 새로운 데이터를 분류
- 회귀(Regression):
  - 연속적인 값을 가지는 함수를 모델링
  - 모델을 사용하여 알려지지 않았거나 누락된 값을 예측



유사점

- 모델 구축: 두 방법 모두 특정 입력에 대해 예측을 하기 위해 모델을 구성

차이점:

- 분류: 범주형 클래스 레이블을 예측. 예를 들어, 이메일이 스팸인지 아닌지를 결정
- 회귀: 연속 공간에서 값을 예측. 예를 들어, 주택 가격이나 온도 같은 연속적인 값을 예측

회귀(Regression)에 대하여:

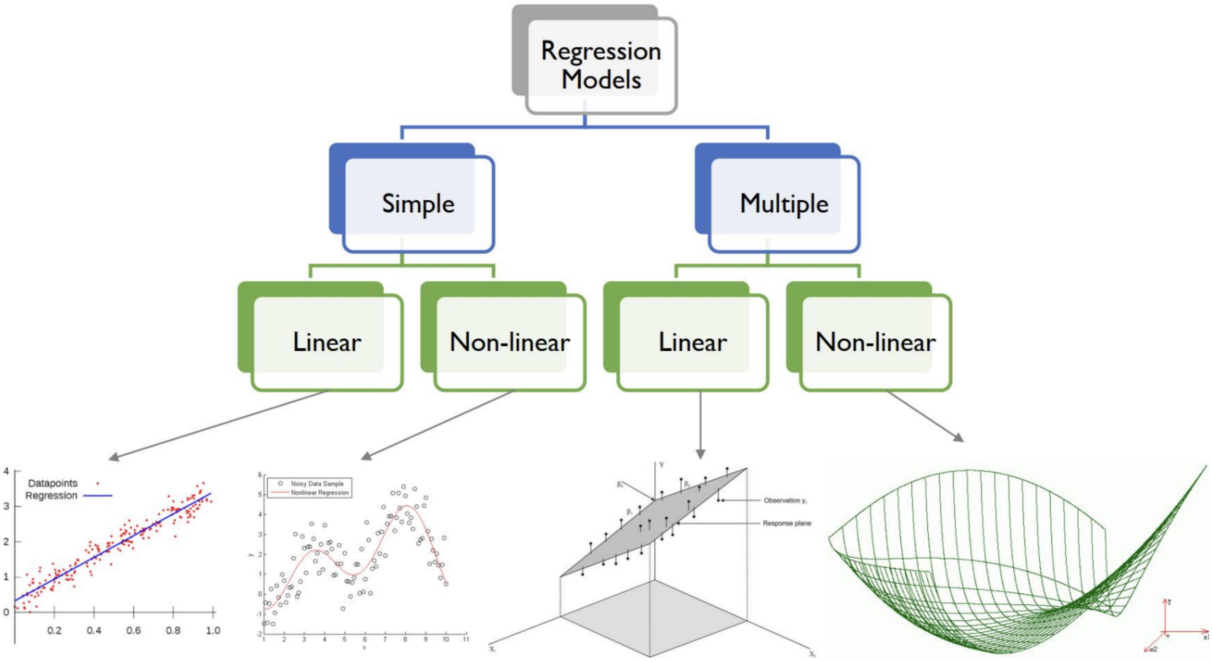
- 예측 변수(독립 변수)와 응답 변수(종속 변수) 간의 관계를 모델링
- 회귀의 유형:
    - 선형 회귀 및 다중 선형 회귀: 변수 간의 선형 관계를 모델링
    - 비선형 회귀: 변수 간의 비선형 관계를 모델링
    - 기타 회귀 방법:
        - 일반화 선형 모델
        - 포아송 회귀
        - 로그-선형 모델
        - 회귀 트리 등

X와 Y사이의 관계

- 확정적(Deterministic) 관계 : 수치예측 $Y = w_0 + w_1X$
- 확률적(Stochastic) 관계: $Y = w_0 + w_1X + \epsilon$

선형회귀

- 선형? 벡터공간의 성질을 보존하는 공간
- 일차식으로 표현되는 공간
- $w_0 + w_1x, w_0 + w_1x + w_2y, \dots$
- $w_0 + w_1x, +w_2x^2 \dots$



1. **단순 선형 회귀(Simple Linear Regression):**
    - 종속 변수 y 와 하나의 독립 변수 x 에 의존하는 선형 함수로 구성됨.
    - 수식: $y = w_0 + w_1 x$
        - 여기서 w_0는 y절편, w_1 은 기울기로, 회귀 계수임.
    - **훈련(Training):**
        - 최적의 직선을 추정하기 위해 필요.
        - w_0 과 w_1 은 훈련 데이터를 사용하여 추정됨.
    - 예시: 집 가격 예측.
2. **다중 선형 회귀(Multiple Linear Regression):**
    - 하나 이상의 독립/입력 변수를 포함.
    - $X =< x_1, x_2, x_3, \ldots, x_n >$: 입력 변수들.
    - y : 종속/출력/목표 변수.
    - 훈련 데이터는 $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ 형태로 구성됨.
    - 예를 들어, 2차원 데이터의 경우 수식은 $y = w_0 + w_1 x_1 + w_2 x_2$ 가 될 수 있음

Idea

1. **아이디어(Idea):**
    - 입력 변수들과 목표 변수 사이의 관계가 항상 선형이라고 가정.
    - 목표 변수 Y 와 일련의 input feature 들의 $x_1, x_2, \ldots, x_p$ 사이에 선형 관계를 적합(fit)시킴.
2. **훈련(Training):**

- 적절한 계수 집합 β 를 찾는 것이 목적.

3. **모델 비교(Which Model is the Best?):**
   - 최적의 β 를 얻는 방법에 대한 고려.

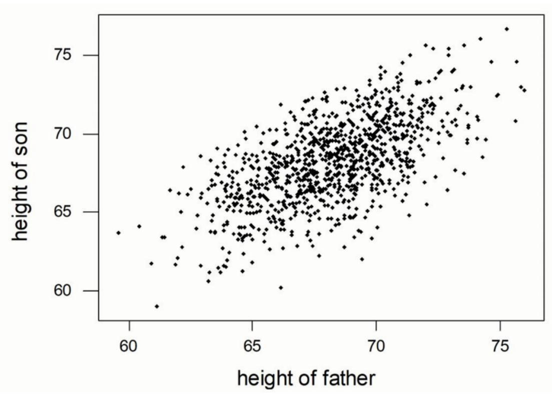
선형 회귀에서 필요한 주요 수식

- **단일 입력 변수(Single Input)의 경우:**
  - $Y = \beta_0 + \beta_1 x_1$
  - 여기서, β_0 는 y절편, β_1 은 기울기
- **다중 입력 변수(Multiple Inputs)의 경우:**
  - $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$
  - 이 경우에는 각 입력 변수 x_i 에 대응하는 계수β_i가 존재

Which model is the Best?

- 최적의 β_0, β_1 는 어떻게 구할수 있을까?

$$E(Y) = f(X) = \beta_0 + \beta_1 X_1$$



Model Training

1. **Cost(또는 loss, error) function E 정의:**
   - 모델 예측과 실제 값 사이의 차이를 나타냄.
   - 일반적으로, $E = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
     - 여기서 n 은 데이터 포인트의 수, y_i 는 실제 값, $\hat{y}_i$ 는 모델에 의해 예측된 값.

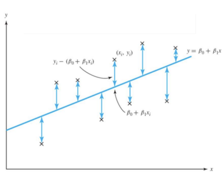
1. **최적화 문제 해결:**
   - 오차 함수 E 를 최소화하는 최적의 β 찾기.

$$E(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

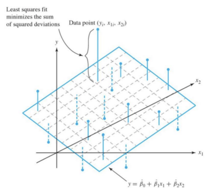
- 제곱 오차 함수(Squared Error Function)를 사용할 때 일반적인 형태

$$E(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}))^2$$

- **goal** : $\arg\min_{\beta} E(\beta)$
  - 함수 $E(\beta)$를 최소화하는 β 값을 찾는다.



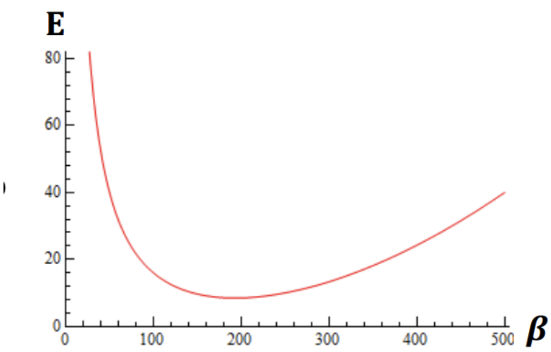
< Single input >

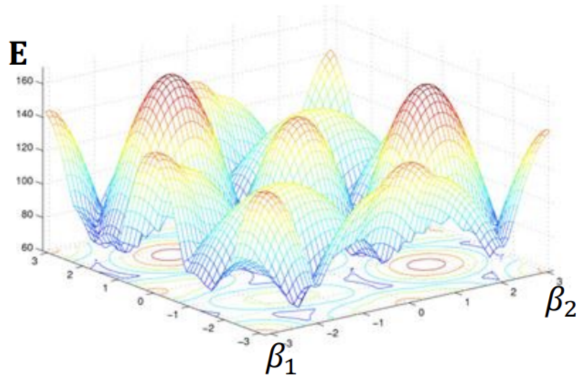


< Multiple inputs >

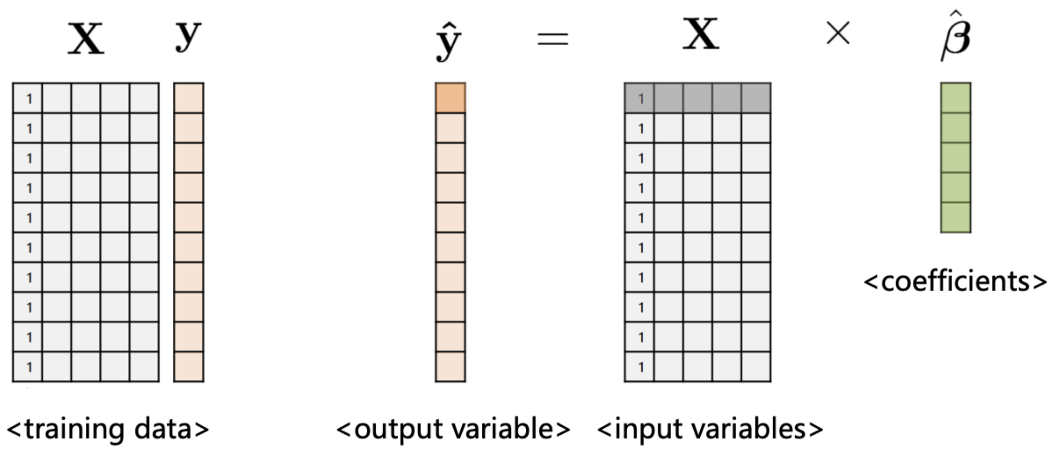
How to solve the optimization problem

- **Cost function E 가 간단한 경우:**
  - E 의 s도함수가 0이 되는 지점을 계산하여 최적의 매개변수를 직접 찾음.
  - 즉, $\frac{\partial E}{\partial \beta_j} = 0$ 형태의 방정식을 푸는 것.
- **Cost function E 가 복잡한 경우:**
  - 여러 모델 매개변수가 있거나, 매개변수들이 서로 연관되어 있을 때 직접 해를 찾을 수 없음.
  - 이 경우, 경사하강법(Gradient Descent)과 같은 수치적 최적화 기법을 사용.
  - 경사하강법의 기본 아이디어는 $\beta_j := \beta_j - \alpha \frac{\partial E}{\partial \beta_j}$와 같이 매개변수를 반복적으로 업데이트하는 것. 여기서 α 는 학습률.





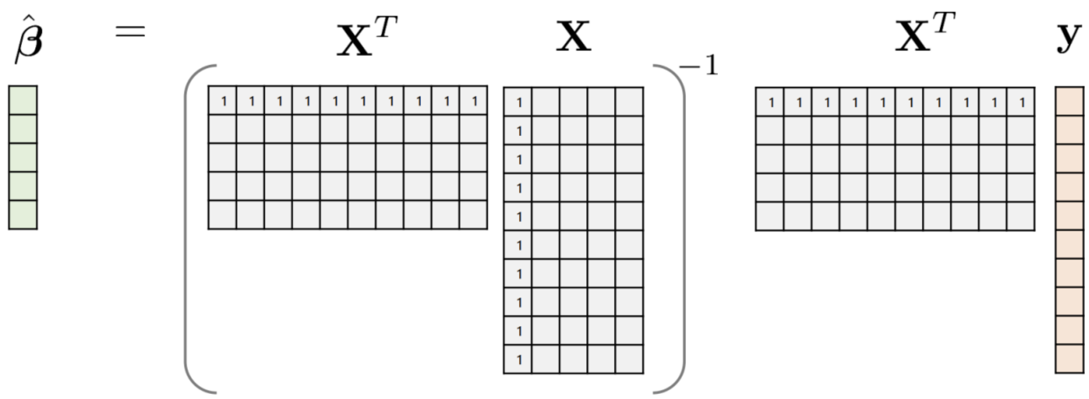
- **데이터 표현**:
  - $\mathbf{X}$: n 개의 관측값과 d 개의 독립 변수를 가진 $n \times (d+1)$ 행렬
    - 여기서 "+1"은 절편 항을 위한 것으로, 보통 $\mathbf{X}$ 행렬에 1로 이루어진 열을 추가한다.
  - $\mathbf{y}$: 종속 변수를 나타내는 $n \times 1$ 벡터이다.



  - $\hat{\beta}$: 절편과 기울기 계수를 포함한 $(d+1) \times 1$ 벡터로, 회귀 계수의 추정치이다.
- **목적 함수**:
  - 최소화하려는 목적 함수는 $E(\mathbf{X}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$ 로, 관측값과 예측값 사이의 차이를 제곱한 합이다.
- **목적 함수의 도함수**:
  - $E(\mathbf{X})$ 를 $\hat{\beta}$ 에 대해 미분하고 0으로 설정하면 정규 방정식을 얻는다

$$-\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$$

- $\hat{\beta}$ **의 해결**:
  - 정규 방정식을 재정렬하면 $\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\hat{\beta} = 0$ 이 되고, 이를 간소화하면 $\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{y}$ 가 된다.
  - $\mathbf{X}^T\mathbf{X}$ 가 가역일 때, $\hat{\beta}$ 에 대한 해는 $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ 가 되어 회귀 계수의 유일하고 명시적인 해를 얻을 수 있다.



Regression Error Measures

회귀 오류 측정의 목적:

- 값 예측의 정확성 측정.
- 예측된 값이 실제 알려진 값(즉, ground truth)에서 얼마나 벗어나 있는지를 측정.

손실 함수(Loss Function):

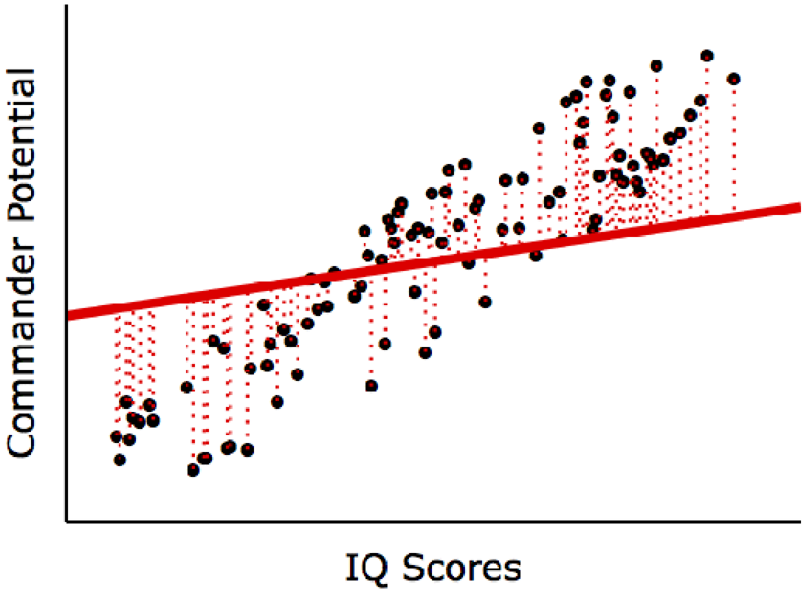
- 실제 값 y_i 와 예측된 값 y_i' 사이의 오류를 측정.
- **절대 오류(Absolute Error):** $|y_i - y_i'|$
- **제곱 오류(Squared Error):** $(y_i - y_i')^2$

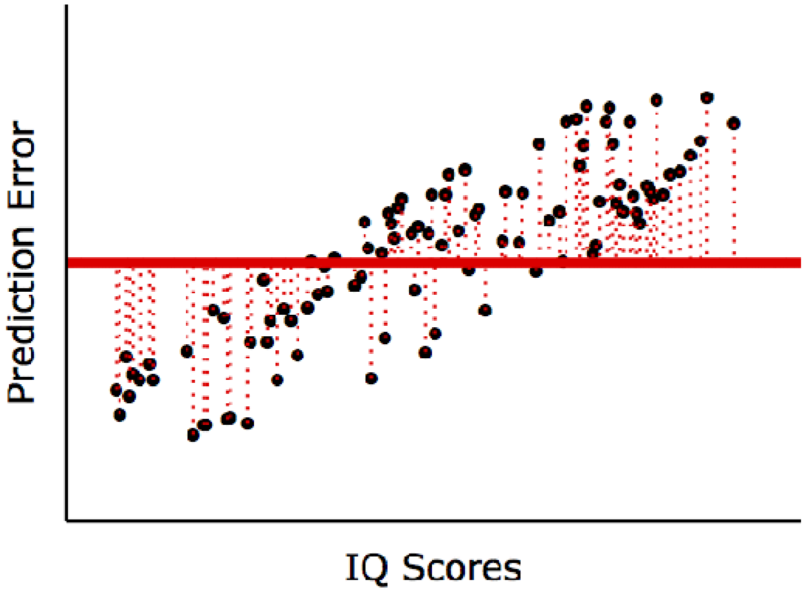
테스트 오류(Test Error, 일반화 오류):

- Test Sets 에 대한 평균 손실.
- **평균 절대 오류(Mean Absolute Error, MAE)**
  - $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'|$
- **평균 제곱 오류(Mean Squared Error, MSE)**
  - $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2$
  - 평균 제곱 오류는 이상치(outliers)에 영향을 많이 받을 수 있음
- **제곱근 평균 제곱 오류(Root Mean Squared Error, RMSE)**
  - $\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2}$
  - 이상치의 영향을 완화하고, 예측된 양과 동일한 크기를 얻기 위해 자주 사용

MAE는 모든 오류를 동일하게 취급하는 반면, MSE와 RMSE는 큰 오류에 더 많은 가중치를 부여합니다. 따라서, 이상치가 중요한 역할을 하는 데이터셋에서는 MSE나 RMSE를 사용하는 것이 좋습니다.

https://annalyzin.files.wordpress.com/2016/01/regression-residual-simulation-tutorial2.gif?w=561&h=842


Commander Potential
IQ Scores


Prediction Error
IQ Scores