세상을 이해하는 통계학의 렌즈

REGRESSION

2020학년도 2학기

앞으로 배울 것

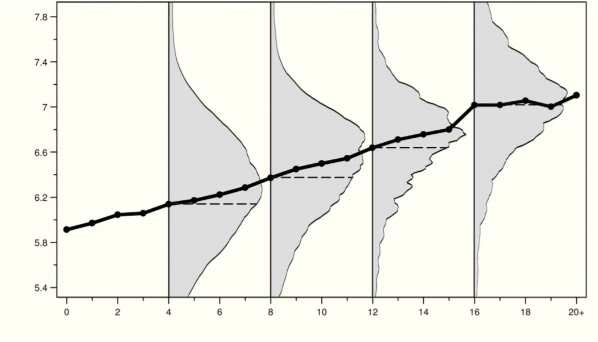
- 여러 종류의 Regression(회귀분석) 방법들 중 **일부**
  - OLS Regression
  - 2SLS Regression
  - Quantile Regression
  - Logistic Regression
  - Polynomial Regression
  - Lasso Regression
  - Ridge Regression
  - Elastic Net Regression
  - Deming Regression
  - $\cdots$ and so on.

오늘 배울 것

- Regression
  - **Function** of Central Tendency (대표 함수)
  - **Function** of Dispersion (대표 함수를 기준으로 흩어진 정도)
  - Setting the Former with the Latter (후자를 가지고 역으로 전자를 만들기)
- OLS Regression (최소제곱 회귀분석)
  - $y = ax + b$ (대표함수의 형태를 가정)
  - L_2 Loss (흩어진 정도를 측정하는 함수를 선정)

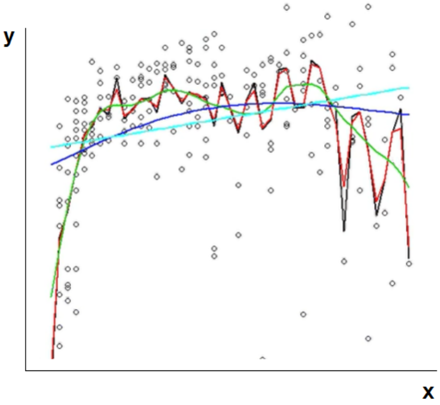
Regression

- **Function** of Central Tendency: (이상적인) 대표함수



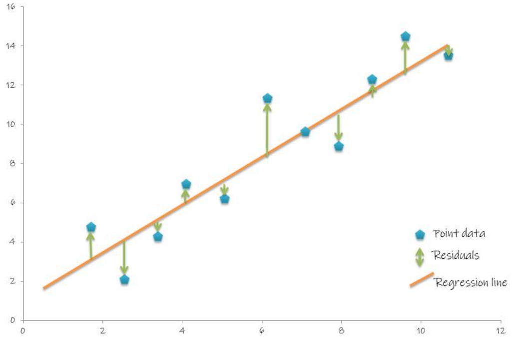
Regression

- Approximating **Function** of Central Tendency:
  관측치로 근사해보는 대표함수



Regression

- **Function** of Dispersion:
  대표함수(투입)를 기준으로 흩어진 정도(산출)



Regression

Function of Dispersion

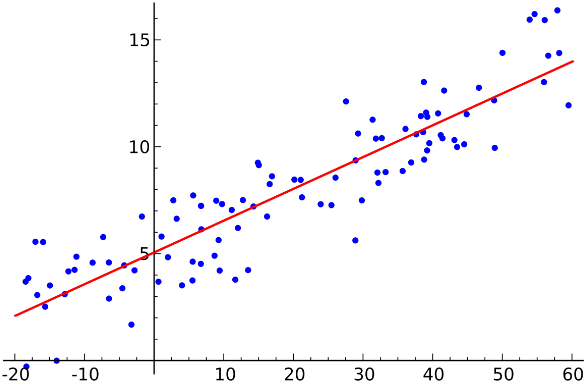
- Loss Functions (참고로 $f(x_i)$: function of central tendency)
  - $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ (Mean Squared Error, Quadratic loss, L_2 Loss)
  - $\frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)|$ (Mean Absolute Error, L_1 Loss)
  - $\sum_{y_i < f(x_i)} (\gamma - 1) \cdot \{y_i - f(x_i)\} + \sum_{y_i \geq f(x_i)} \gamma \cdot \{y_i - f(x_i)\}, \gamma \in (0,1)$
    (Quantile Loss) : '분위수'를 기준으로 |편차|들을 가중 평균
  - $\sum_{i=1}^{n} log[cosh\{f(x_i) - y_i\}]$
    (Log-Cosh Loss) : 미분이 가능하도록 만든 MAE의 업그레이드 버전
  - $\cdots$ and so on.

Regression

- 중심선을 잘 그리는 법: 손실함수(Loss Function)를 활용하기
  - Plotting the Function of Central Tendency
    by Minimizing the Function of Dispersion

OLS Regression

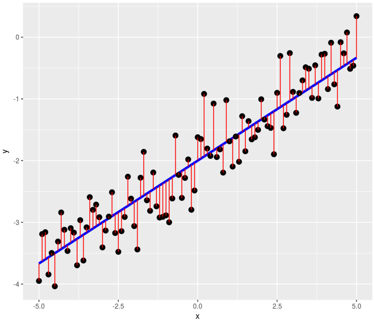
- OLS(Ordinary Least Squares) Regression
  - $f(x) = ax + b$
  - L_2 Loss



Ordinary Least Squares Regression

$$\underset{a,\ b}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

$$f(x_i) = ax_i + b$$



OLS Regression

$$L(a,b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

$$\frac{\partial L(a,b)}{\partial a}|_{a=a^*} = 0$$

$$\frac{\partial L(a,b)}{\partial b}|_{b=b^*} = 0$$

OLS Regression

$$a^* = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

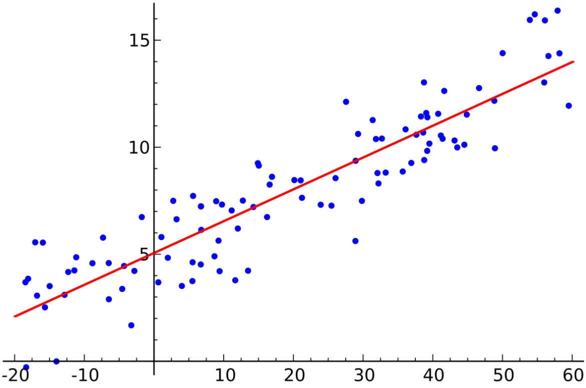
$$b^* = \bar{y} - a^* \bar{x}$$

OLS Regression

$$f(\mathbf{x}) = [\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]\mathbf{x} + [\bar{y} - a^*\bar{x}]$$

OLS Regression

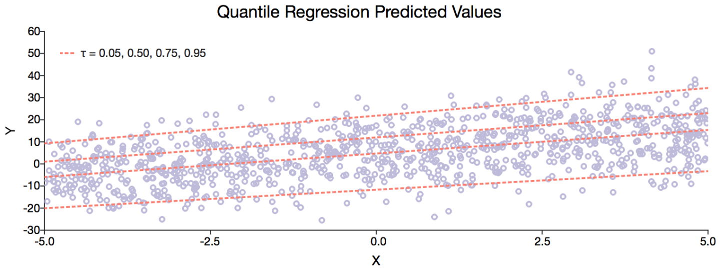
- OLS Regression
  - $f(x) = a^* x + b^*$
  - L_2 Loss



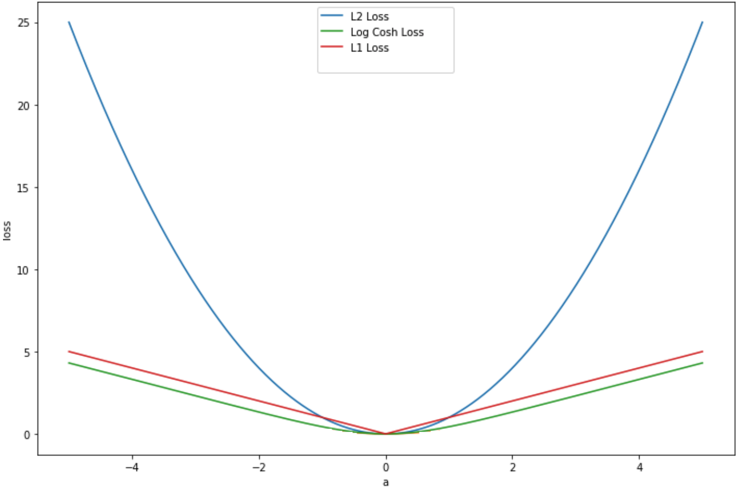
Regression

- **Function** of Dispersion
  - Loss Functions (참고로 $f(x_i)$: function of central tendency)
    - $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ (Mean Squared Error, Quadratic loss, L_2 Loss)
    - $\frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)|$ (Mean Absolute Error, L_1 Loss)
    - $\sum_{y_i < f(x_i)} (\gamma - 1)\cdot\{y_i - f(x_i)\} + \sum_{y_i \geq f(x_i)} \gamma \cdot \{y_i - f(x_i)\}, \gamma \in (0, 1)$
      (Quantile Loss) : '분위수'를 기준으로 |편차|들을 가중 평균
    - $\sum_{i=1}^{n} log[cosh\{f(x_i) - y_i\}]$
      (Log-Cosh Loss) : 미분이 가능하도록 만든 MAE의 업그레이드 버전
    - $\cdots$ and so on.

Quantile Regression



Log-cosh Regression



Log-cosh Regression