세상을 이해하는 통계학의 렌즈

REGRESSION

2020학년도 2학기

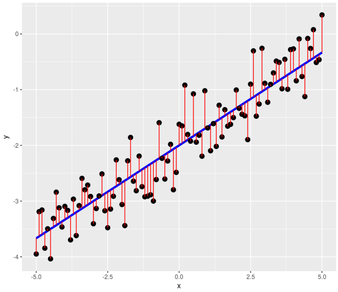
Ordinary Least Squares Regression

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

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

- 키워드: OLS 회귀분석, L2 손실함수, Objective Function, Argument, Minimize, a와 b의 함수, Mean Squared Error, 중심선(f(X))을 기준으로 데이터들이 흩어져 있는 정도, y= ax+b라는 가정.

OLS Regression

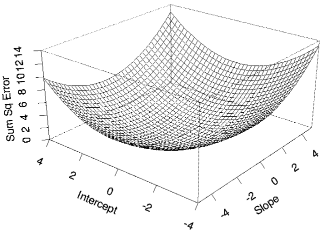


$$\underset{a,\ b}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

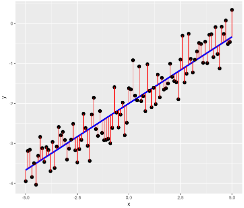
$$y = a^* x + b^*$$

OLS Regression

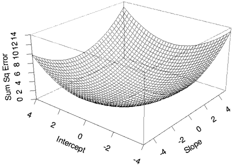
- Visualization of Quadratic Loss given a(slope) and b(intercept).



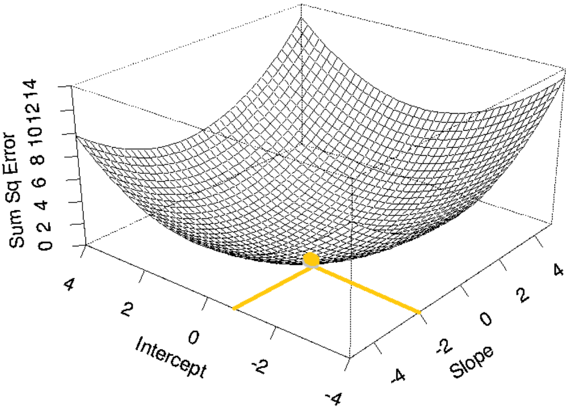
OLS Regression



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



OLS Regression



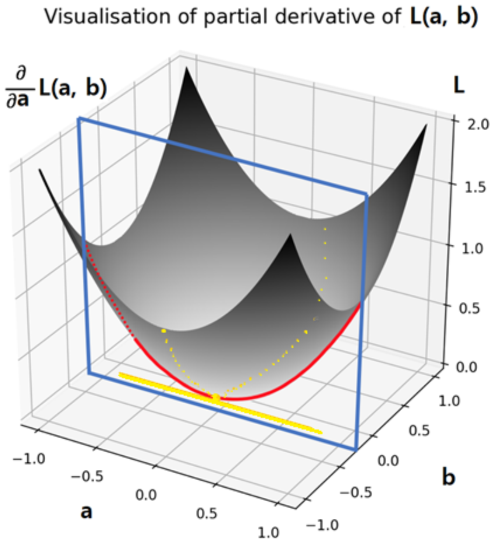
OLS Regression

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

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

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

OLS Regression



OLS Regression

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

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

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

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

OLS Regression

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

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

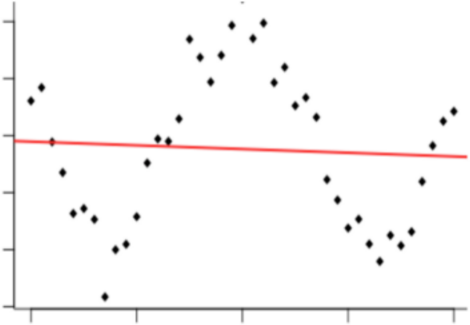
$\bar{x}$: x의 평균
$\bar{y}$: y의 평균

$$f(\mathbf{x}) = a^*x + b^*$$

Pearson Correlation Coefficient

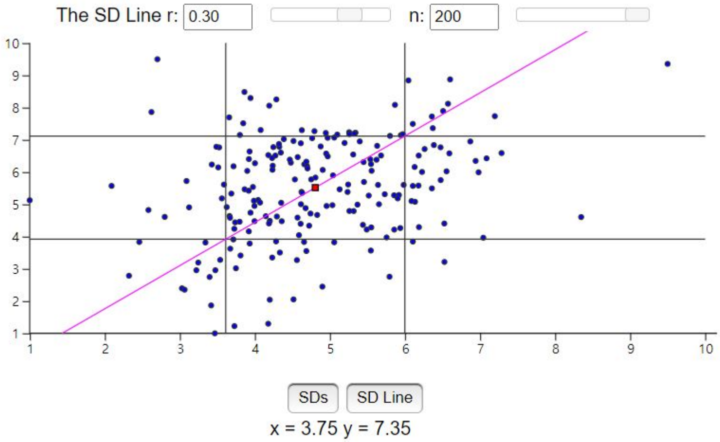
$$r_{XY} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

Pearson Correlation Coefficient



$r = -0.028$

Standard Deviation Line



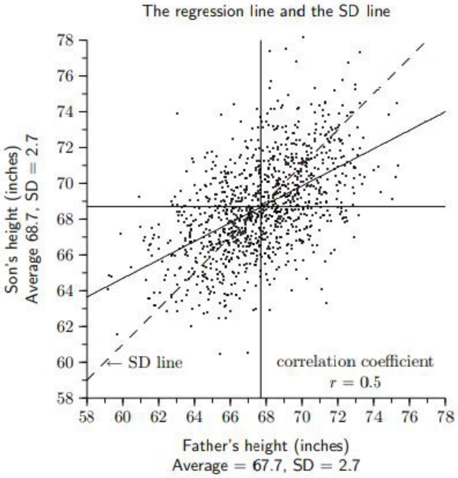
OLS Regression

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

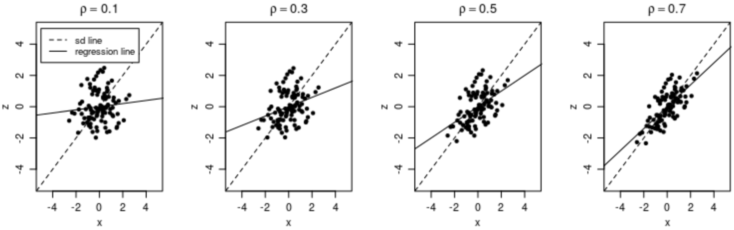
$\bar{x}$: x의 평균
$\bar{y}$: y의 평균

$$f(\mathbf{x}) = a^* x + b^*$$

Meaning of "a(slope)" in OLS Regression Line



Meaning of "a(slope)" in OLS Regression Line



OLS Regression

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

$\bar{x}$: x의 평균
$\bar{y}$: y의 평균

$$f(\mathbf{x}) = a^* x + b^*$$

OLS Regression

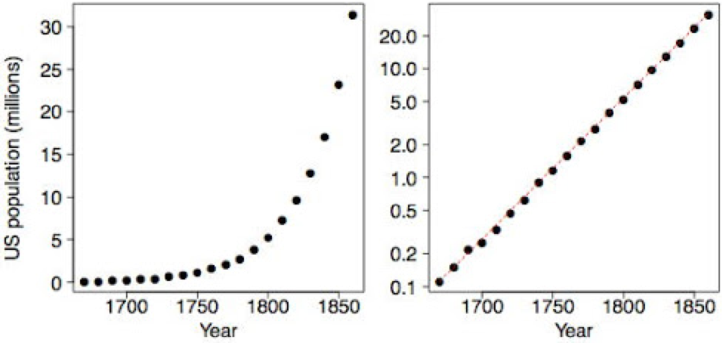
$$y = a^* x + b^*$$

$$\log y = a^* x + b^*$$

$$y = a^* \log x + b^*$$

$$\log y = a^* \log x + b^*$$

Log Transformation for Linearity



OLS Regression

$$y = a^* x + b^*$$

$$\log y = a^* x + b^*$$

$$y = a^* \log x + b^*$$

$$\log y = a^* \log x + b^*$$