Linear Regression and Correlation

- Explanatory and Response Variables are Numeric
- Relationship between the mean of the response variable and the level of the explanatory variable assumed to be approximately linear (straight line)
- Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma)$$

  - $\beta_1 > 0 \Rightarrow$ Positive Association
  - $\beta_1 < 0 \Rightarrow$ Negative Association
  - $\beta_1 = 0 \Rightarrow$ No Association

Least Squares Estimation of β_0, β_1

- $\forall$ $\beta_0 \equiv$ Mean response when x=0 (y-intercept)
- $\forall$ $\beta_1 \equiv$ Change in mean response when x increases by 1 unit (slope)
- β_0, β_1 are unknown parameters (like μ)
- $\beta_0+\beta_1 x \equiv$ Mean response when explanatory variable takes on the value x
- Goal: Choose values (estimates) that minimize the sum of squared errors (*SSE*) of observed values to the straight-line:

$$\hat{y} = b_0 + b_1 x \qquad SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2$$

Example - Pharmacodynamics of LSD

- Response (y) - Math score (mean among 5 volunteers)
- Predictor (x) - LSD tissue concentration (mean of 5 volunteers)
- Raw Data and scatterplot of Score vs LSD concentration:

Score (y)	LSD Conc (x)
78.93	1.17
58.20	2.97
67.47	3.26
37.47	4.69
45.65	5.83
32.92	6.00
29.97	6.41

Source: Wagner, et al (1968)