Linear Regresion

TI linear aproximation of a casual relationship between two or more variables.

---

process

GET SAMPLE DATA → Design model that works for that sample → Make predictions for whole population

---

Types of variable - Dependant (predicted) - $Y = F(x_1, x_2 \dots x_k)$

Independent $x_1, x_2 \dots x_k$

TYPES OF REGRESIONS - SIMPLE REGRESION

$y = \beta_0 + \beta_1 x_1 + \varepsilon$

* y = dependant variable
  x = independant variable
  β_1 = quantify the dependent effect
  β_0 = constant base
  ε = error (on average is 0)

Ej: income (y) depends on years of education (x)

In USA $\beta_1 \approx 5000$ β_0 = minimun wage

Equation

$$\hat{y} = b_0 + b_1 x_1$$

* when we had a hat $\hat{y}$ it means is an estimate or predicted value

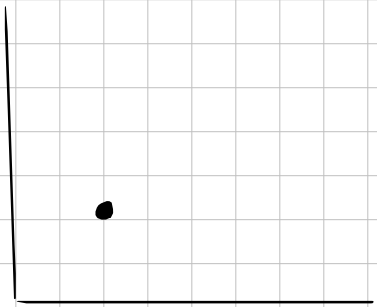
CORRELATION vs REGRESSION

□ Correlation do not imply causation □

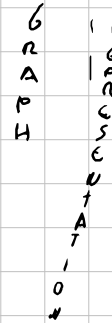
CORRELATION

- Imply relationship
- Variables move together
- $p(x,y) = p(y,x)$

SINGLE POINT



GRAPH



REGRESSION

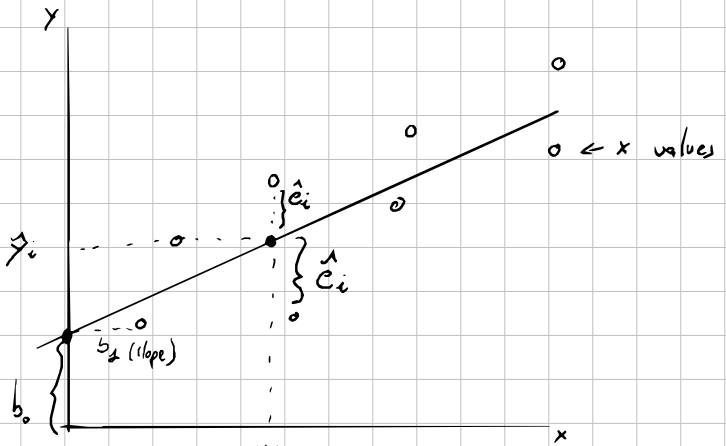
- One variable affect the other
- Instead of move together is more a cause and effect situation
- $P(x,y) \neq p(y,x)$ Eg. Income and education

education usually mean more income but not the other way

LINE

GEOMETRICAL REPRESENTATION OF LINEAR REGRESION

$$\hat{y} = b_0 + b_1 x_1$$



Decomposition of Variability

- **SST / TSS (Sum of Squares Total)**
  - Sum of the total variablity of the dataset

$$\sum_{i=1}^{n} (y_i - \bar{y})^2$$

- **SSR - Sum of Squares Regression - ESS**
  - Measure of how well the line fit the data

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

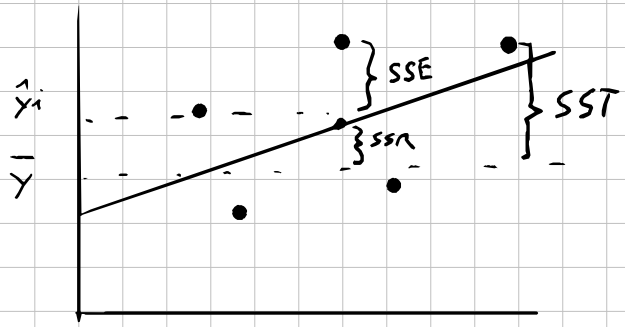
  - * If SSR = SST it means your regresion model is perfect and capture all the observed variability

- **SSE - Sum of Squares Error - RSS**
  - Measure the unexplained variability by the regresion

$$\sum_{i=1}^{n} e_i^2$$

Total Variability = Explained Variability + Unexplained Variability

$$SST = SSR + SSE$$



What is the OLS?