Linear Regression

- Goal of regression is to estimate the impact of the variation in X (independant variables) on central tendency of Y (outcome variable).
- There are many measures of central tendency, for now let us consider the mean:

$$\text{Population: } \mu = E(Y)$$

$$\text{Sample: } X_1, X_2, \dots, X_n \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Although the above definitions are intuitive, an equivalent definition of the mean is as follows:

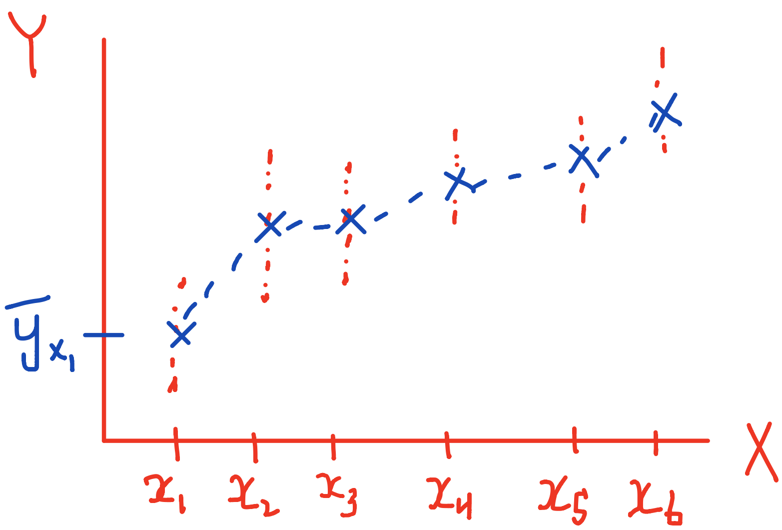
$$\text{Population: } \min_{\mu} E\left[(Y-\mu)^2\right] \Rightarrow \mu = E(Y)$$

$$\text{Sample: } \min_{\mu} \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The relationship between X and Y can be summarized by the conditional expectation function E(Y|X). The CEF is interesting because it measures the mean tendency of Y|X (Y given X).

$$\min_{\mu(x)} E\left[(y-\mu(x))^2 \mid X\right] \Rightarrow \mu(x) = E(Y|X)$$

- The CEF is a function of X and represents the average of Y for a given value of X.



$$\underline{\text{CEF:}}\ E[Y|X=x]$$

$$\underbrace{\bar{y}_{x_1} = \frac{1}{n_{x_1}} \sum_{j:X=X_1} y_j}_{\text{Avg. of } Y \text{ at } X=X_1}$$

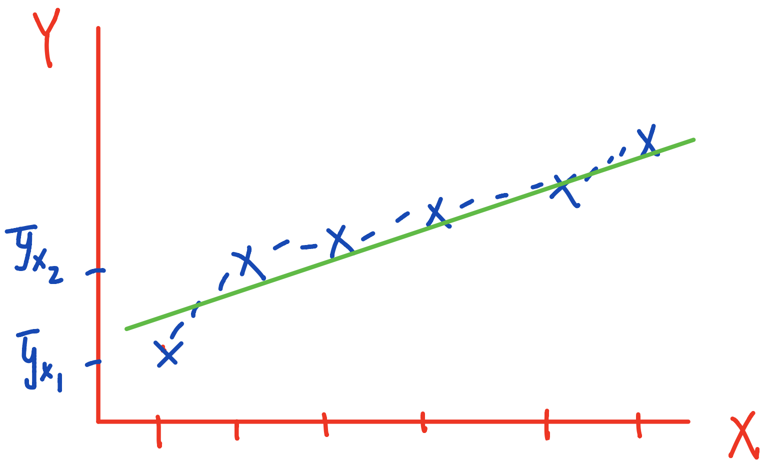
- It can be shown that Y can be decomposed into the CEF plus an orthogonal (unrelated to X) error term.

$$\underbrace{Y_i}_{\text{Outcome}} = \underbrace{E[Y_i|X_i]}_{\text{CEF}} + \underbrace{\varepsilon_i}_{\text{Error}}, \quad \underbrace{E[\varepsilon_i|X_i]=0}_{\text{Mean Independance}}$$

- Linear regression imposes parametric assumption on the CEF.

$$E[Y_i|X_i] = \beta X_i \Rightarrow Y_i = \beta X_i + \varepsilon_i$$

- We are essentially approximating the CEF with the standard linear regression model.



OLS: $\hat{Y}_i = \hat{\beta} X_i$

CEF: $\{\bar{y}_{x_1}, \bar{y}_{x_2}, \dots\}$

- We can estimate the linear regression parameter as follows:

Since $\mu(x) = \beta X \Rightarrow \min_{\beta} E\left[(y - X\beta)^2\right]$

$$\Rightarrow \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \Rightarrow \hat{\beta} = \frac{S_{xy}}{S_x S_y}$$

where S_{xy} = sample cov(X, Y), S_x & S_y are stdev.

- We estimate the linear regression parameter above by minimizing the sum of square errors. This uses the square loss function (also known as L2 loss).
- As discussed above, using the square loss will result in a model that approximates the CEF. Using a different loss function will result in a different interpretation of the model.
- If we have multiple regressors x1, x2, ..., xk, all the math above can be extended using matrices. Now the parameter vector can be estimated as follows:

Since $\mu(X_1, \dots, X_k) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k = X\beta$

$$\Rightarrow \min_{\beta} \frac{1}{n} (Y - X'\beta)'(Y - X'\beta) \Rightarrow \hat{\beta} = (X'X)^{-1} X' y$$

- Note above that above dimensions are: dim(beta) = (k+1)x1, dim(X) = n x (k+1), and dim(y) = n x 1