Supervised Learning Setup

Note: I've covered some derivations but this will help in the long run.

* If this book seems hard to approach, start from this part. I tried to simplify as much as I could. *

Supervised Learning: Think of this as a child who is walking by his father in a garden. His father is feeding him about properties about the flowers in the garden and after enough knowledge, the kid will be able to identify flowers all by himself. This is supervised learning.

Now, there is a thing called training set.

$$\{(x^1, y^1), (x^2, y^2), \ldots (x^n, y^n)\}$$

* Let's assume the most commonly used stat data which is house data.

(Suppose we have a dataset of house size vs house prize)

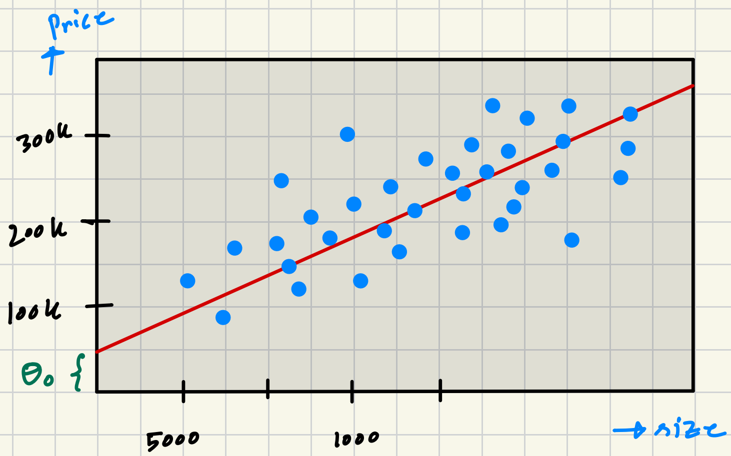
* Now, our training set is really important, it is a part of a large population. We care about new xs so that we can predict new ys.

* If 'y' is discrete, it is a classification problem.
(Ex: Is this a Rose? Is this a tulip)
If 'y' is continuous, it is a regression problem.

* Now. in the housing dataset, we have price vs area and our target is to find the price of a given lot area.
We can call this a hypothesis (function)
& we can plot this as

$$h(x) = \theta_0 + \theta_1(x)$$

{Following this notation is a bit hard for me; so, later I used $f(x) = \beta_0 + \beta_1(x)$}

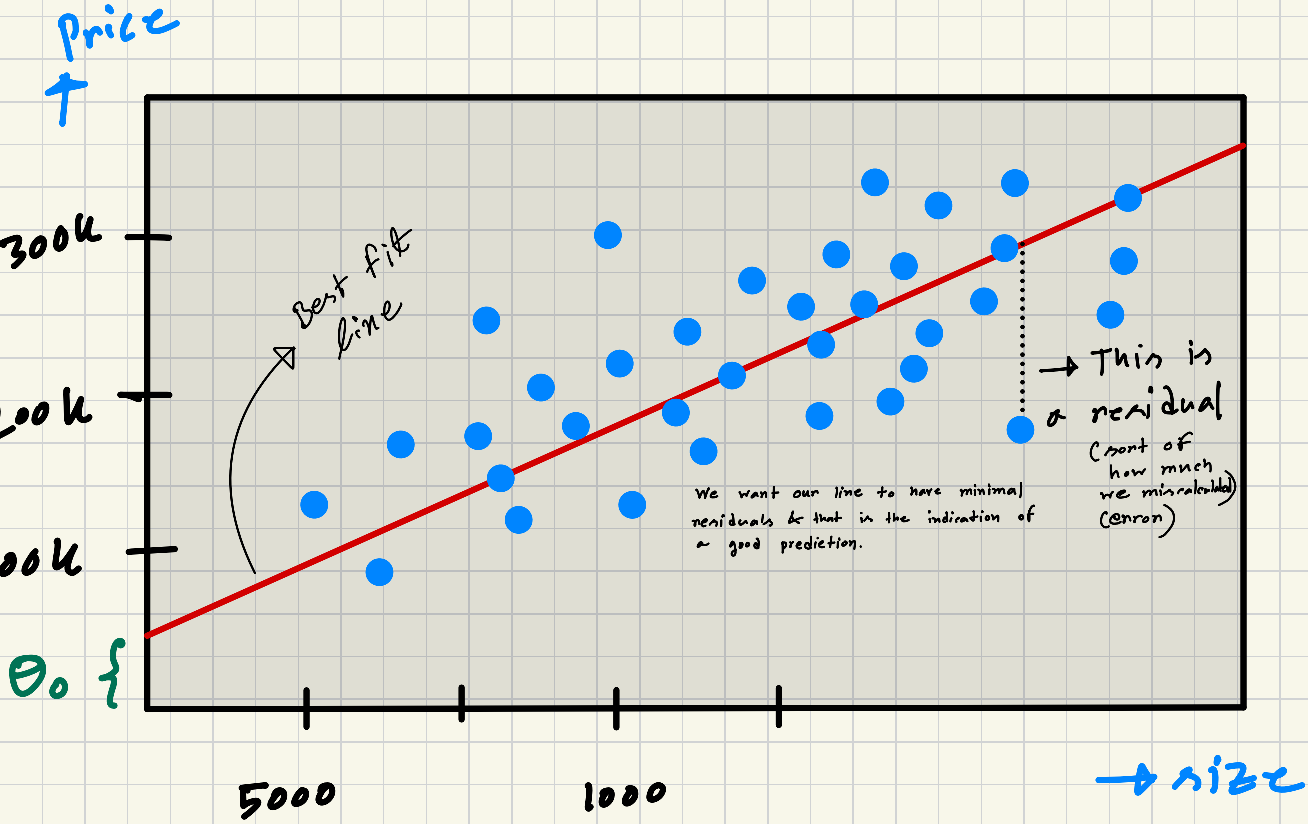


These are the Xs

This is y

Size	Bedroom Count	lot size			Price
2200	4	45k			400k
2500	3	30k			900k

How do we find an example of something good?



We'll use sum of the squares of the residuals & this is for historical purposes only. (You can use absolutes)

$$h_\theta(x) = \sum_{j=0}^{d} \theta_i x_j \qquad h_\theta(x) \approx y$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(h_\theta(x^i) - y^i\right)^2$$

Linear Regression

- ① **Simple Linear Regression :** One input column vs one output column.
- ② **Multiple LR :** Multiple input columns.
- ③ **Polynomial LR :** For non linear data.

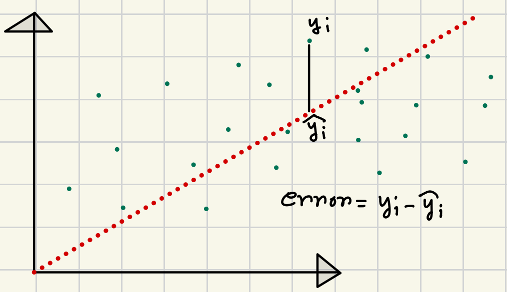
* Simple Linear Regression :

The main idea was finding the slope & intercept. There are two ways to do it

1) Closed Form Solution { We use direct formula } [OLS]
2) Non closed form { kind of derivation } [Gradient Descent]
   ↳ Better for higher dimensions.

OLS

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



1) We have sort of linear points.
2) We want to find the best fit line.
3) We find the sum of the square of the residuals.

$$RSS = e_1^2 + e_2^2 + e_3^2 + \dots + \dots e_n^2$$

$$\therefore E = e_1^2 + e_2^2 + \dots e_n^2$$

$$E = \sum_{i=1}^{n} d_i^2$$

But why square & why not modulus ?

→ Mod's graph is not differentiable.

But, the error we're finding it is in terms of y.

$\therefore d_i = y_i - \hat{y}_i$

$\hat{y}_i = m x_i + b$

Total error

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} (y_i - m x_i - b)^2$$

Average error

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

Now, our target is to minimize the error function & for that we need derivation.

$$\therefore E_{(m,b)} = \sum_{i=1}^{n} (y_i - m x_i - b)^2$$

Now,

$$\frac{\delta}{\delta b} E = 0$$

$$\Rightarrow \frac{\delta}{\delta b} E = \sum_{i=1}^{n} \frac{\delta}{\delta b} (y_i - m x_i - b)^2 = 0$$
$$= \sum_{i=1}^{n} \left(2 (y_i - m x_i - b) \times -1 \right) = 0$$
$$= \sum_{i=1}^{n} (y_i - m x_i - b) = 0$$
$$= \sum y_i - \sum m x_i - \sum b = 0$$
$$= \frac{\sum y_i}{n} - \frac{\sum m x_i}{n} - \frac{\sum b}{n} = 0$$
$$= \bar{y} - m\bar{x} - b = 0$$

$$\frac{\delta}{\delta m} E = \frac{\delta}{\delta m} \sum (y_i - m x_i - \bar{y} + m\bar{x})^2$$
$$= \sum \left(2 (y_i - m x_i - \bar{y} + m\bar{x}) \times (-x_i + \bar{x})\right)$$
$$= \sum \left[(y_i - \bar{y}) - m(x_i - \bar{x})\right](x_i - \bar{x}) = 0$$
$$= \sum \left[(y_i - \bar{y})(x_i - \bar{x})\right] = m \sum (x_i - \bar{x})^2$$
$$\therefore m = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})}$$

Multiple Linear Regression

→ When multiple input columns exist.

Let's assume a dataset

x_1	x_2	y	
cgpa	iq	salary	

$$y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n$$

$$y = B_0 + B_1X_1 + B_2X_3$$

$$y = B_0 + \sum_{i=1}^{n} B_iX_i$$

* Mathematical Proof:

* I'll work with a dataset that has 9 input columns & one output column.

$$\hat{y} = B_0 + B_1X_1 + B_2X_2 + B_3X_3$$

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} B_0 & B_1X_{11} & B_2X_{12} & B_3X_{13} & B_mX_{1m} \\ B_0 & B_1X_{21} & B_2X_{22} & B_3X_{23} & B_mX_{2m} \\ \vdots & & & & \\ B_0 & B_1X_{n1} & B_2X_{n2} & B_3X_{n3} & B_nX_{nm} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & X_{1m} \\ 1 & X_{21} & X_{22} & X_{23} & X_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & X_{nm} \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_m \end{bmatrix}$$

∴ Prediction Matrix = Unknown matrix × Co-efficient matrix

Error Matrix = Actual Y matrix − Predicted Y matrix

$$Mat(e) = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_3 - \hat{y}_3 \end{bmatrix}$$

Now, $E = e^T e$

$$\Rightarrow \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \begin{pmatrix} y_1 - \hat{y}_1 & y_2 - \hat{y}_2 & \cdots & y_n - \hat{y}_n \end{pmatrix} \begin{pmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{pmatrix}$$

$(1 \times n)(n \times 1) = 1 \times 1$

Now,

Formulae

$(A \pm B)^T = A^T \pm B^T$

$(AB)^T = B^T A^T$

matrix diff formulae

$y = A^T X A$

$\frac{dy}{dA} = 2XA^T$

$$E = e^T e = (Y - \hat{Y})^T (Y - \hat{Y})$$
$$= (Y^T - \hat{Y}^T)(Y - \hat{Y})$$
$$= (Y^T - (XB)^T)(Y - XB)$$
$$= Y^T Y - \underbrace{Y^T X B - (XB)^T Y}_{\text{Equal}} + (XB)^T (XB)$$

$$E = Y^T Y - 2Y^T X B + (XB)^T (XB)$$

↳ Loss function

Now, we need to find its lowest point using differentiation

$$\frac{dE}{dB} = \frac{d}{dB}\left[Y^T Y - 2Y^T X B + B^T X^T X B\right] = 0$$

$$= 0 - 2Y^T X + 2X^T X B^T$$

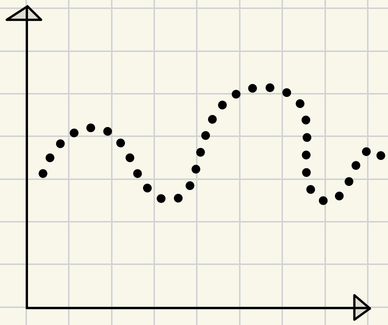
$$((A^T A)^{-1})^T = A^T A^{-1}$$

Now, $\cancel{2} X^T X B^T = \cancel{2} Y^T X$

$B^T = Y^T X [X^T X]^{-1} \Rightarrow B = [(X^T X)^{-1}]^T X^T Y$

$\therefore B = (X^T X)^{-1} X^T Y$

Polynomial Linear Regression



$$y = ax^2 + bx + c$$

$$\hat{y} = \hat{a}x^2 + \hat{b}x + c$$

$$e_i = (y_i - \hat{y}_i)$$

$$RSS = e_1^2 + e_2^2 + e_3^2 + \dots\dots + e_n^2$$

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

$$= \sum (ax^2 + bx + c - \hat{a}x^2 - bx - c)$$

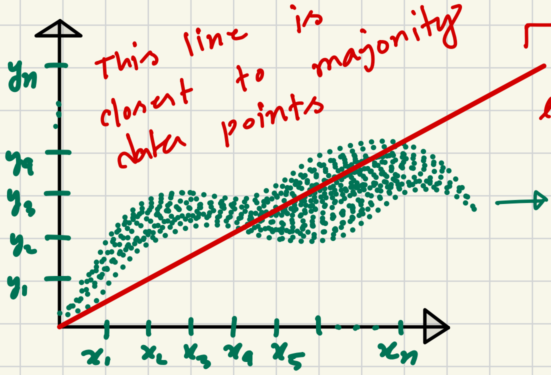
$$y = \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \theta_3 x^3 + \dots\dots \theta_n x^n$$

Q: Find the best fit second degree polynomial for the given data $\{(1,3), (2,4), (3,8)\}$

1 → We had 200 datapoints. (of two dimensional data)

→ We assumed some relationship $Y \approx B_0 + B_1 X$

→ Data points were $(x_1, y_1), (x_2, y_2), (x_3, y_3) \ldots\ldots (x_n, y_n)$



→ We actually find a $\hat{B}_0$ & $\hat{B}_1$ that creates a line.

→ Suppose, this is how the data is placed.

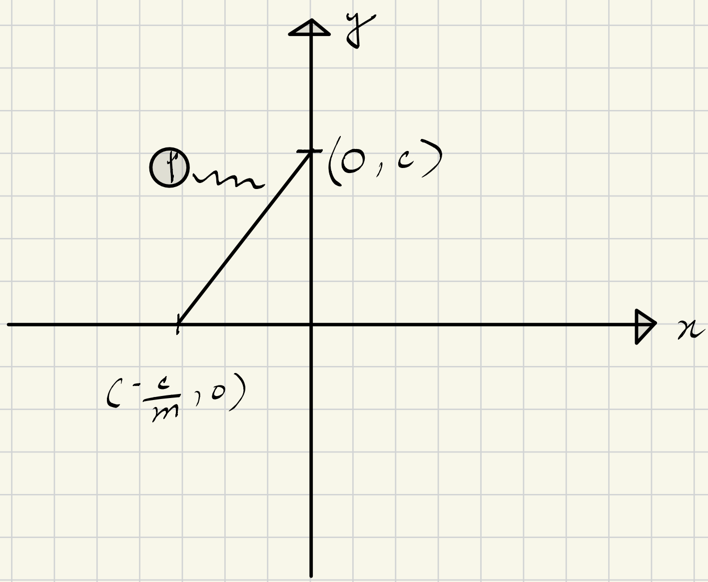
Note: The simple linear equation is $y = mx + c$ where m is the slope and c is the y intercept.

If we want to plot this line, we'll have to

$$y = mx + c$$

$$\Rightarrow mx - y = -c \Rightarrow \frac{mx}{-c} + \frac{y}{c} = 1 \Rightarrow \frac{x}{-c/m} + \frac{y}{c} = 1$$

→ 1



$$\sum_{i=0}^{n} 3i-1 = (3\times 0-1) + (3\times 1-1) + (3\times 2-1) + \cdots + (3\times n-1)$$

$n \to$ End value

$i=0 \to$ Stant value

Now, $\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$ predicted value of y based on the ith value of x.

y_i = Obsend ith response

Residue $e_i = y_i - \hat{y}_i$

$\therefore$ RSS (Residual Sum of Squares) $= e_1^2 + e_2^2 + \ldots + e_n^2$

We try to minimize RSS through $\hat{B}_0$ & $\hat{B}_1$

$$\left.\begin{aligned} \hat{B}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n}(x_i - \bar{x})^2} \\ \hat{B}_0 &= \bar{y} - \hat{B}_1 x \end{aligned}\right\} 3.4$$

$\to$ **We can use this as formulas foR the slope & Intercept. This $\hat{B}_1$ & $\hat{B}_0$ give the smallest RSS.**

* Assessing Precision

Standard error $\Rightarrow$

$$SE(\hat{B}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$ → Standard error for slope

$$SE(\hat{B}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{n\sum_{i=1}(x_i - \bar{x})^2}\right]$$

→ Standard error for intercept

$$\sigma^2 = Var(E)$$