$$a = \frac{\overline{xy} - \bar{x}\bar{y}}{\left(\overline{x^2} - \bar{x}^2\right)}$$

$$b = \frac{\bar{y}\,\overline{x^2} - \bar{x}\cdot\overline{xy}}{\overline{x^2} - \bar{x}^2}$$

CLOSED FORM SOLUTION to the problem min(SSE)

possible in the case of S.L.R

For M.L.R : Numerical Methods "Gradient Descent Method"

Dependent

$y = a \cdot x + b$

$= \beta_1 x + \beta_0$

one independent variable

Simple Linear Regression

S.L.R.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$

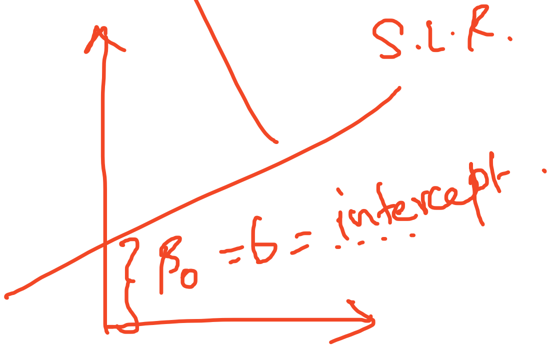
coefficients.

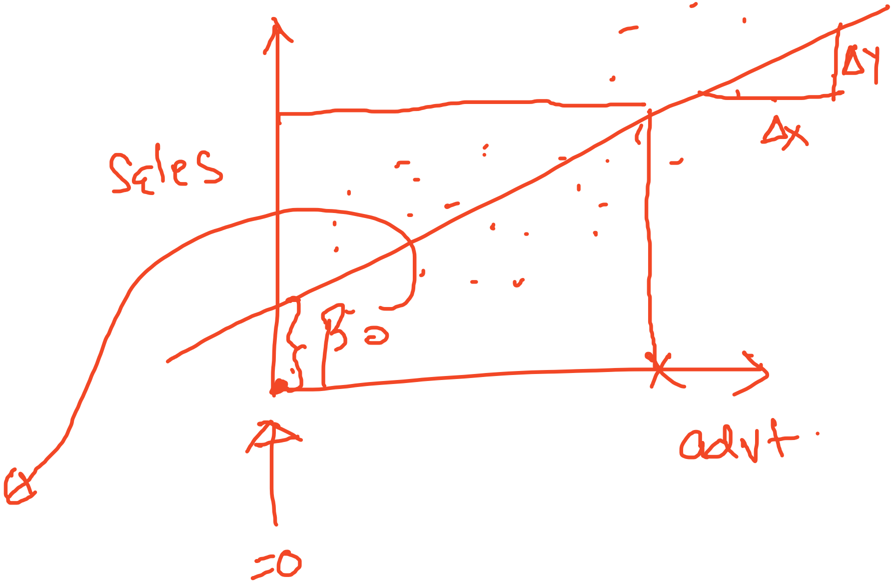
Multiple Indep. Vars $\Rightarrow$ M.L.R.

Multiple Linear Regression

$y = ax + b$

$y = \beta_0 + \beta_1 x$





β_0 = intercept = Net effect of all the unknown variables.

Can you guarantee that you have included ALL the 'independent variables while creating any ML model? Impossible ...

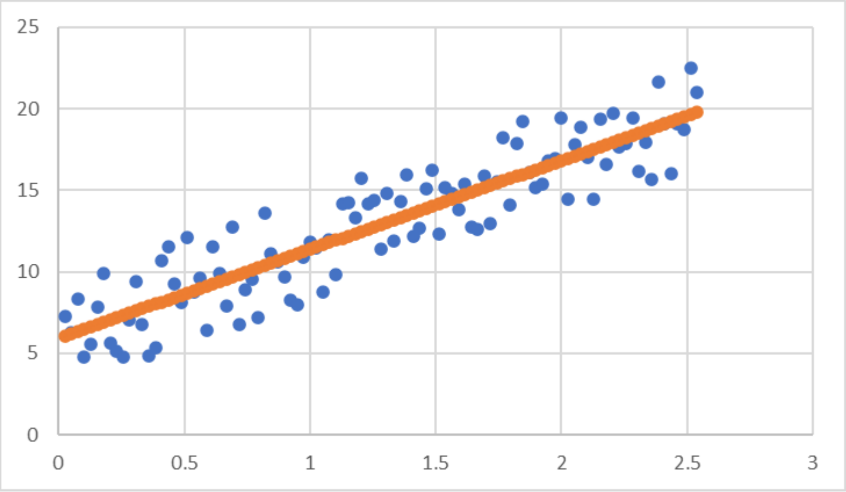
e_i → Regression Errors

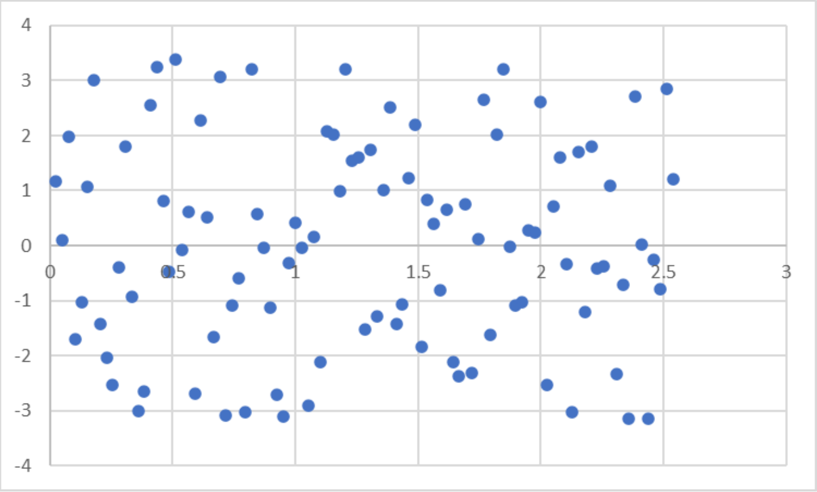
& the variants .... MAE, MSE, RMSE

→ How to interpret this (and the others) & how to use it?

Errors
- Absolute context → where/how can you use it?
- Relative context
  - → LR1 – 3.64
  - → LR2 – 2.55

You can use these values to select the better model



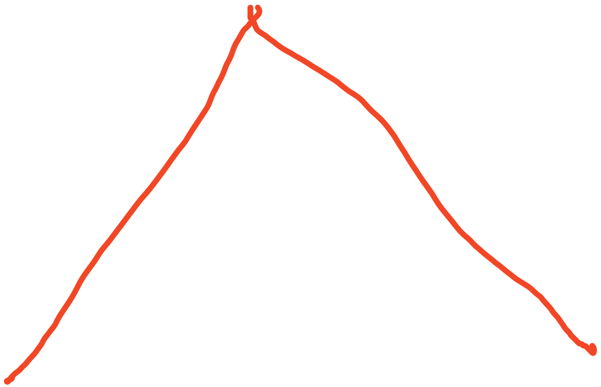


y_i v/s x_i and

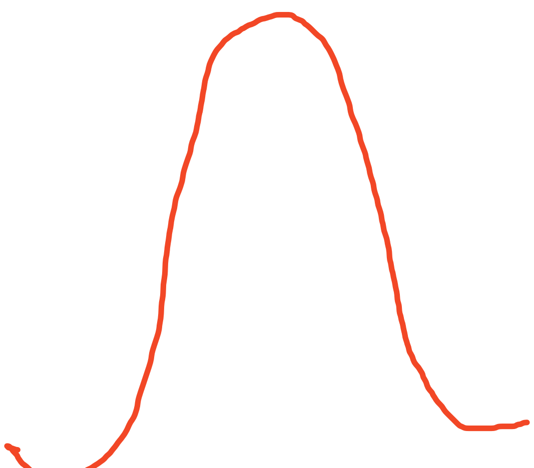
$\hat{y}_i$ v/s x_i

e_i v/s x_i

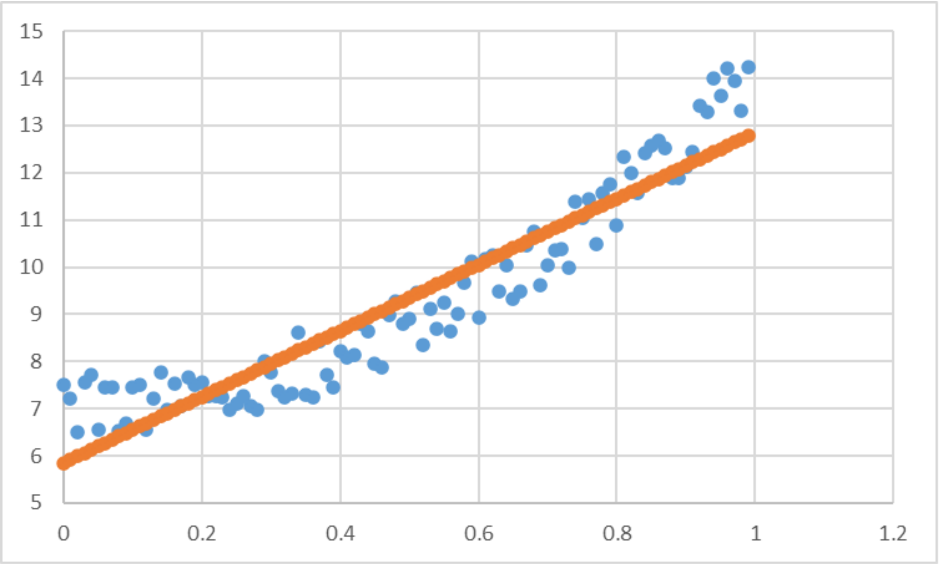
- Errors do not show any distinctive pattern.
- Histogram of the errors should indicate NORMAL DISTRIBUTION.



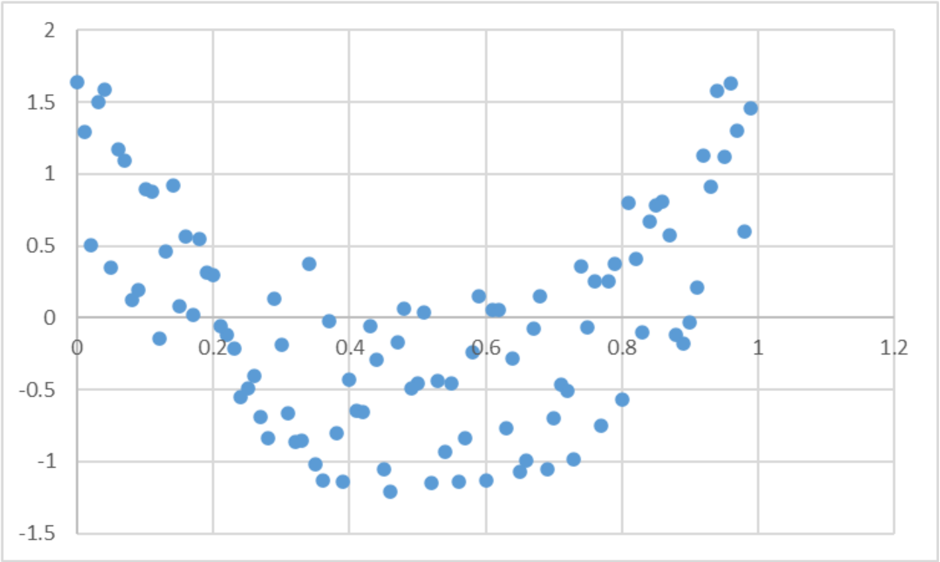
Why can't the histogram of errors take the shape of 'triangular distribution'?



Are there any mathematical tests that prove that a dataset (e_i) is following Normal distribution.



Example of a line "force fitted" on non-linear data.



The resulting scatter plot of the error values $\{e_i\}$

- Errors display a distinct pattern

$\Rightarrow$ the model has failed to pick-up the inherent pattern in the data.

OUTPUTS CREATED BY REGRESSION TOOLS / FUNCTIONS

What do these numbers indicate?

	A	B	C	D	E	F	G	H	I	J	K	L
1	y	x				SUMMARY OUTPUT						
2	7.238462	0.025641										
3	6.310256	0.051282				*Regression Statistics*						
4	8.315385	0.076923				Multiple R	0.906270151					
5	4.787179	0.102564				R Square (R^2)	0.821325586					
6	5.592308	0.128205				Adjusted R Square	0.819483582					
7	7.830769	0.153846				Standard Error	1.882513522					
8	9.902564	0.179487				Observations	99					
9	5.607692	0.205128										
10	5.146154	0.230769				ANOVA						
11	4.784615	0.25641					*df*	*SS*	*MS*	*F*	*Significance F*	
12	7.05641	0.282051				Regression	1	1580.159507	1580.16	445.8869	4.72338E-38	
13	9.394872	0.307692				Residual	97	343.7541446	3.543857			
14	6.8	0.333333				Total	98	1923.913652				
15	4.871795	0.358974										
16	5.376923	0.384615					*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	10.71538	0.410256			'b'	Intercept	5.922586409	0.381284372	15.53325	4.65E-28	5.165842474	6.679330343
18	11.55385	0.435897			'a'	x	5.452241187	0.258203842	21.11603	4.72E-38	4.939778036	5.964704338
19	9.258974	0.461538										
20	8.097436	0.487179										
21	12.10256	0.512821										

What is a good model?

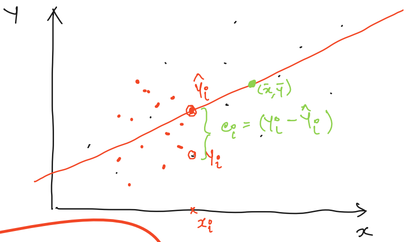
– One that explains most of the variations in the data.

$\sum (y_i - \bar{y})^2 = SST$ (SST = measure of total variation in the given dataset)

$\sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$

$\sum \left[(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)\cdot(\hat{y}_i - \bar{y}) \right]$

$\sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2\sum(\ )(\ )$



$SST = SSE + SSR + \underbrace{2\sum(\ )(\ )}_{0}$

SSR => total variation explained by the regression model
SSE => variation NOT explained by the model, attributed to random errors

Work this out and confirm for yourself using the data data set already with you

$SST = SSR + SSE + \text{ZERO}$

$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$

$1 = \underline{R^2} + \frac{SSE}{SST}$

$R^2 = \boxed{\text{COEFFICIENT OF DETERMINATION (C.O.D).}}$

= Square of the correlation coefficient 'r' between x & y.

$R^2 = 1 - \frac{SSE}{SST}$

R^2 is the square of the correlation coefficient 'r' } S.L.R

SELF STUDY
- CORRELATION
- CORRELATION COEFF

+ve correlation

-ve correlation.

$r = \frac{\sum (x_i - x)(y_i - y)}{(\ \ldots\ )}$ ~ Correlation.

So far, we have calculated SSE, MSE, RMSE, MAE, R2 as metrics reflecting the quality of Linear Regression. However, when we use built-in LR functionality, in tools like Excel, many more numbers are generated .. as shown below. What are they and how to interpret / use them?

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1580.159507	1580.16	445.8869	4.72338E-38
Residual	97	343.7541446	3.543857		
Total	98	1923.913652			

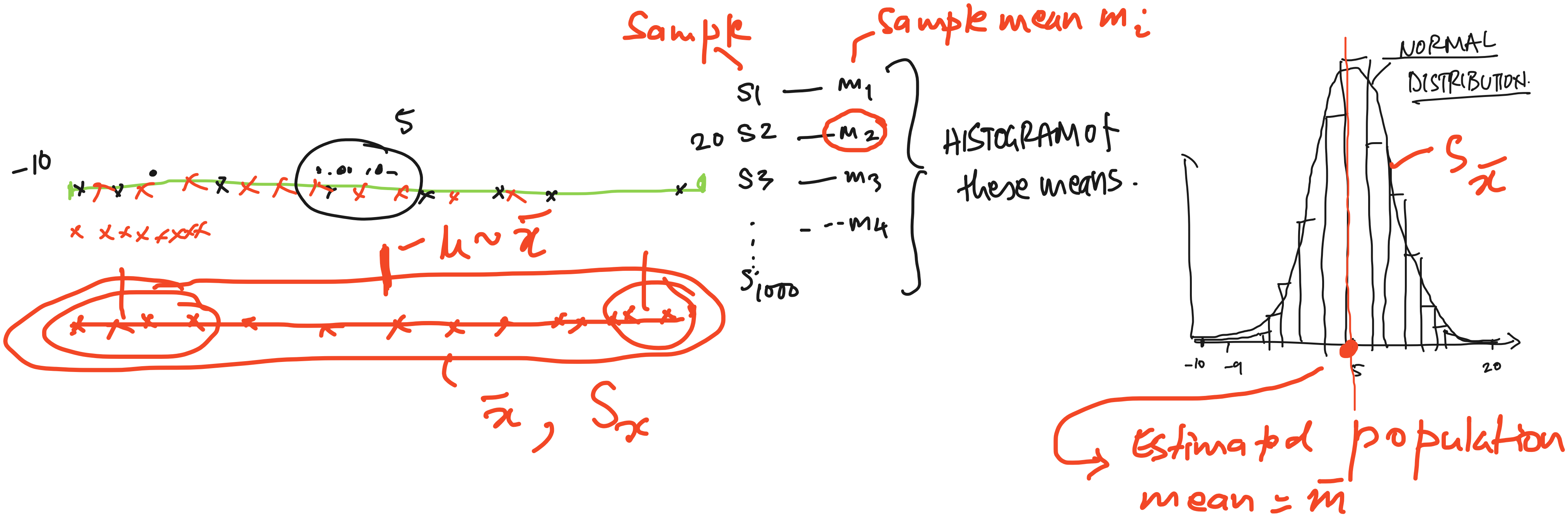
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.922586409	0.381284372	15.53325	4.65E-28	5.165842474	6.679330343
x	5.452241187	0.258203842	21.11603	4.72E-38	4.939778036	5.964704338

Regression Statistics	
Multiple R	0.906270151
R Square	0.821325586
Adjusted R Square	0.819483582
Standard Error	1.882513522
Observations	99

- To understand these numbers we have to go back to the basics of statistics.
- We need to start with the fact that the (y,x) data that we have is essentially a **sample** (in this case 1 sample of 99 observations)
- We have fitted an LR model using this sample. Therefore the calculated values of **a** and **b** are only an estimate of the population's **actual** a and b.
- Our aim, really, is to predict the value of **y** for an **x** that is not a part of the sample. That is, we need a model that is '**general** and which reflects the reality of the population, and not limited to the sample that we have.
- So we really need to know **how good** an estimate these calculated values (a, b) are. Are they really usable? How much confidence should we have on our calculations?
- This is where we need to understand the concepts, from statistics, of **sampling distributions** and **confidence intervals**

We conduct some 'thought' experiments, related to estimating the population mean from the sample mean:

- Assume that from a population we can take multiple **good**, **representative** samples, let's say **k** samples, each of size **n.** Let's call each sample as s_i
- Using each s_i, we calculate its mean and call it m_i
- Since our samples are **good, representative** samples of the population, they will result in means m_i that are close to each other (why ? try to reason this out)
- If we collect all the m_i and create a frequency table and a histogram, it's shape will be as shown below.



- We will observe that such a histogram indicates that the calculated means m_i tend to have Normal Distribution (as per the **Central Limit Theorem** - see next slide)
- This distribution is known as the **Sampling Distribution of the mean** or **Sampling Distribution of the sample mean** and it has the following properties:
    - The **Expected Value** (ie. mean) of such a distribution is very close to the population mean
    - The Standard Deviation of this distribution - known as the **Standard Error**, and denoted by **S_xbar** - is related to **sigma**, the population's standard deviation in the following way:

$$S_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$ where n = size of the sample.

    - Implication of this formula: For a given population, with a given sigma, S_xbar reduces with increase in the sample size **n**. This, in turn, indicates less uncertainty in estimating the true value of the population mean.
    - This appeals to our common sense that **as the sample sizes increase, our analysis becomes more accurate** or, conversely, **smaller sample sizes result in more uncertainty or inaccuracy in our predicted results**

So - given 100 observations, does it make sense to treat it as 1 sample of size 100, or 10 samples of size 10?

The Central Limit Theorem

The Central Limit Theorem (CLT) is a fundamental concept in statistics that describes the distribution of sample means for a sufficiently large sample, regardless of the shape of the original population distribution.

Central Limit Theorem:

For a random sample of size n drawn from any population with a finite mean μ and a finite standard deviation σ, the distribution of the sample means will approach a normal distribution as n becomes sufficiently large. Specifically, as n approaches infinity, the distribution of the sample means will have a mean equal to the population mean ($\mu_{\bar{X}} = \mu$) and a standard deviation equal to the population standard deviation divided by the square root of the sample size ($\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$).

The Central Limit Theorem is particularly powerful because it allows statisticians to make inferences about population parameters based on the distribution of sample means, **even when the original population distribution is unknown or not normally distributed**. This theorem forms the basis for many statistical techniques and hypothesis tests that rely on the normal distribution.