

Example: Soft drink bottling company wanted to control the variance by not allowing the variance to exceed 0.0004. Does the sample of size 28 with a variance 0.0010 indicate that their bottling process is out of control? Use $\alpha = 0.05$.

Solution:

0.0004: specification set by company

0.001: process variance (it must be sufficiently small)

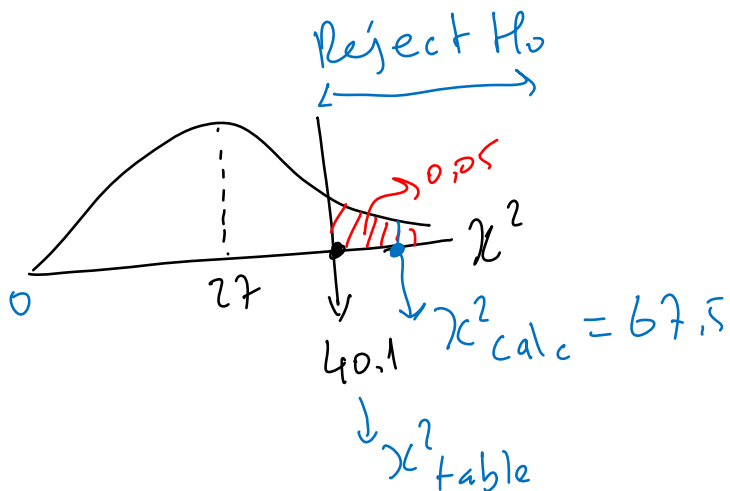
$\times H_0: \sigma^2 = 0.0004$ (machine is not out of control)

$\vee H_A: \sigma^2 > 0.0004$ (" " " out of control)

$\alpha = 0.05$, d.f. = 27

$$\chi^2_{\text{calc}} = \frac{(n-1) \cdot s^2}{\sigma^2} = \frac{27 \times 0.001}{0.0004} = 67.5$$

$$\chi^2_{\text{table}}(27, 0.05) = 40.1$$



Decision: Reject H_0

Conclusion: The bottling process is out of control (problem!!!).

Example: One of the factors used in determining the usefulness of a particular exam as a measure of students' ability is the amount of "spread" that occurs in the grades. A set of test results is of little value if the range of the grades is very small. However, if the **range of grades is quite large**, there is a definite difference in the scores achieved by the "better" students and the scores achieved by the "poorer" students.

On an exam with a total of 100 points, it has been claimed that a Standard deviation of 12 points is desirable. An instructor gave an exam to his class. There were 28 students and the Standard deviation of those 28 scores was found to be 10.5. Does the instructor have evidence at 0.05 level of significance that this exam does not have the specified Standard deviation ? (i.e., test whether it is a good exam)

Solution:

$$\left. \begin{array}{l} \checkmark H_0: \sigma = 12 \\ \times H_A: \sigma \neq 12 \end{array} \right\} \alpha/2 = 0,025$$

$$n = 28, \text{ d.f.} = n - 1 = 27, \quad S = 10.5, \quad \alpha = 0.05$$

$$\alpha/2 = 0.05/2 = 0.025$$

Critical values:

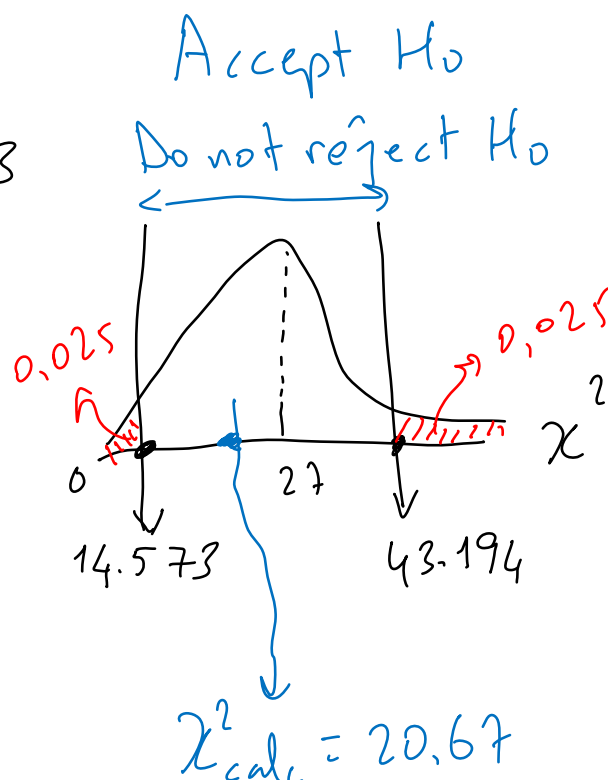
$$\chi^2_{\text{lower}} (27, 0.975) = 14.573$$

$\left. \begin{array}{l} \downarrow \\ \text{d.f.} \end{array} \right\} \left. \begin{array}{l} \downarrow \\ (1 - \frac{\alpha}{2}) \end{array} \right\}$

$$\chi^2_{\text{upper}} (27, 0.025) = 43.194$$

$\left. \begin{array}{l} \downarrow \\ \alpha/2 \end{array} \right\}$

$$\chi^2_{\text{calc}} = \frac{27 \times (10.5)^2}{(12)^2} = 20.67$$



Decision: Do not reject H_0

Conclusion: This exam has the specified st. dev. (12). i.e., it is a good exam.

Confidence Interval (CI)

$$\frac{(n-1) \cdot s^2}{\chi^2(\text{d.f.}, \alpha/2)} < \sigma^2 < \frac{(n-1) \cdot s^2}{\chi^2[\text{d.f.}, (1-\frac{\alpha}{2})]}$$

Example: Using the same data, calculate 95 % confidence interval for the estimate of the population variance and Standard deviation.

Solution:

$$\frac{27 \times (10.5)^2}{43.194} < \sigma^2 < \frac{27 \times (10.5)^2}{14.573}$$

$$68.9 < \sigma^2 < 203.9 \rightsquigarrow 95\% \text{ CI for } \sigma^2$$

* for st. dev. \Rightarrow take $\sqrt{\quad}$ both sides \Rightarrow

$$8.3 < \sigma < 14.3 \rightarrow 95\% \text{ CI for } \sigma.$$

b) Test of Goodness of Fit by χ^2 -Distribution

The data that we will be using in this technique will be enumerative, that is, the data used will result from counts of occurrences (experiments).

- This test compares any 3 or more groups or classes.
- The experimental and predicted (estimated, theoretical) data are compared.
- **This test will always be a one-tailed, upper-tailed test.**
- Reject the null hypothesis if χ^2_{calc} exceeds the χ^2_{table} value for (n-1) degrees of freedom and α given.

$$\chi^2_{\text{calc}} = \sum \frac{(O - E)^2}{E}$$

O: observed or experimental values

E: expected, theoretical data.

This test will always be a one-tailed, upper tailed test.

Example: There were 7 sections of a particular mathematic course. Students were scheduled to meet at various times with a variety of instructors. The following table shows the number of students who selected each of the 7 sections:

Section	1	2	3	4	5	6	7	Total
# of students	18	12	25	23	8	19	14	119

Do the data indicate that the students had a preference for certain sections? Or do the data indicate that each section was equally likely (by chance) to be chosen?

Use $\alpha = 0.05$.

Solution:

For 119 students to be equally distributed among each class \Rightarrow

$$\frac{119}{7} = 17 \text{ students/class} \rightarrow \text{expected or theoretical}$$

X H_0 : Students are equally distributed (no preference)

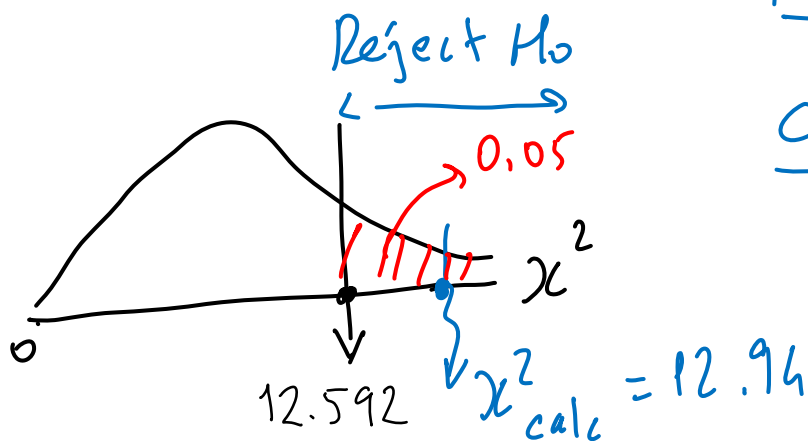
✓ H_A : " are not " (there is preference)

$$\chi^2_{\text{calc}} = \sum_{i=1}^N \frac{(O - E)^2}{E}$$

$$\chi^2_{\text{calc}} = \frac{(18-17)^2}{17} + \frac{(12-17)^2}{17} + \dots + \frac{(14-17)^2}{17} = 12.94$$

$$\chi^2_{\text{table}} (\text{d.f., } \alpha) \Rightarrow n = 7, \nu = 6$$

$$\chi^2_{\text{table}} (6, 0.05) = 12.592$$



Decision: Reject H_0

Conclusion: There seems to be a preference shown for certain sections.

Example: The data obtained from an experiment periodically are 1, 3, 6, 9 and 13. These data are fitted to a mathematical model (equation). The predicted values obtained from the model are 2, 4, 5, 6 and 9. Does the model fit (i.e., suitable) the experimental data? Use $\alpha = 0.05$.

Solution:

Observed	1	3	6	9	13
Expected	2	4	5	6	9

✓ H_0 : Data fit the model (i.e., model is suitable)

✗ H_A : " do not fit " " (" not ")

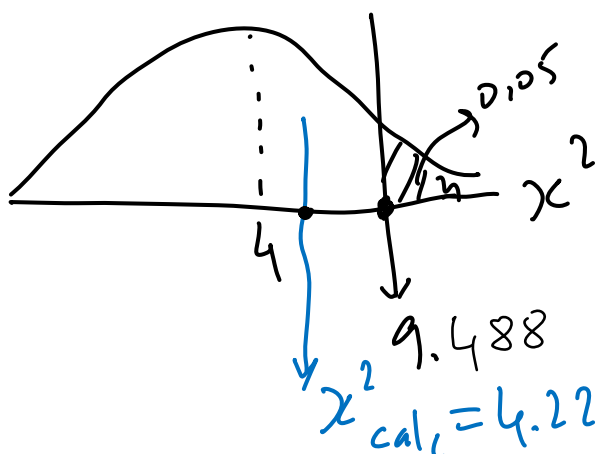
$$n = 5, \quad \nu = 4$$

$$\chi^2_{\text{calc}} = \frac{(1-2)^2}{2} + \frac{(3-4)^2}{4} + \frac{(6-5)^2}{5} + \frac{(9-6)^2}{6} + \frac{(13-9)^2}{9}$$

$$\chi^2_{\text{calc}} = 4.22$$

$$\chi^2_{\text{table}}(4, 0.05) = 9.488$$

Do not reject H_0



Decision: Do not reject H_0

Conclusion: The model is suitable for the data.

Homework 1: At a cereal filling plant, quality control engineers do not want the variance of the weights of 750 grams cereal boxes to exceed 100 gram^2 . A sample of 7 boxes of this type of cereal with a nominal weight of 750 grams had the following weights: 775, 780, 781, 795, 803, 810, 823.

Does this sample provide strong evidence that the true variance of the weights exceeds 100 gram^2 ? Use $\alpha = 0.05$.

Homework 2: A vending machine for beverages has five different choices. The salesman for the machine says that the types of beverages are **all equally favored** among people. Ali wonders if this is true so he observes what people get out of the machine and records the results in the table below. Is the salesman right ?

Use $\alpha = 0.05$.

Drink	Coke	Pepsi	Sprite	Orange	Ice Tea
Frequency	52	63	25	59	48

Linear Correlation and Linear Regression

A) Linear Correlation

The purpose of correlation analysis is to measure the strength of linear relationship between variables (say x, y).

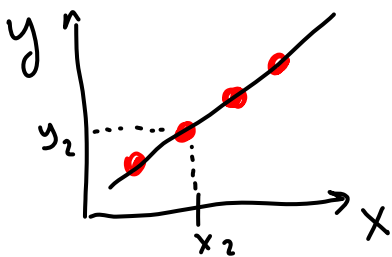
Correlation coefficient (r) is the measure of the strength of linear relationship between two variables.

- a) Perfect positive correlation
- b) " negative "
- c) No correlation

a) Perfect Positive Correlation

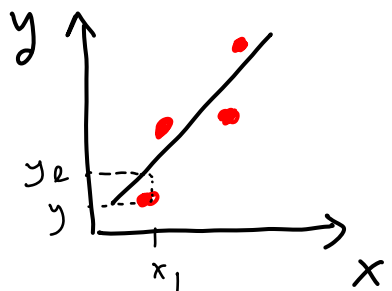
Two variables are said to have perfect correlation if;

1. for each unit increase in one variable there is a fixed increase in the other,
2. for each unit decrease in one variable there is a fixed decrease in the other.

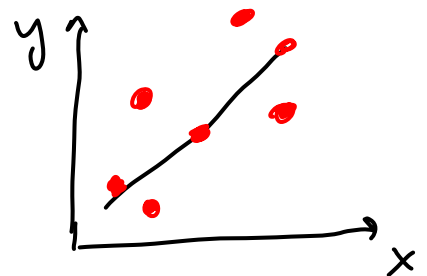


Perfect positive correlation

— predicted
● experimental



High positive correlation



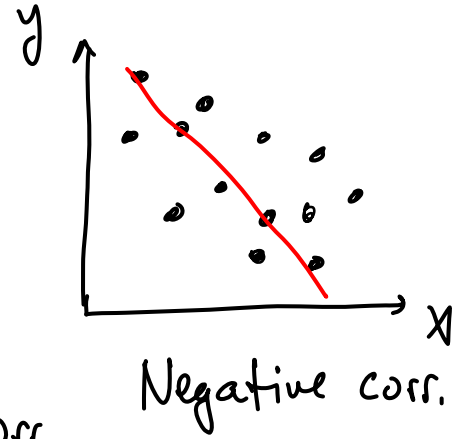
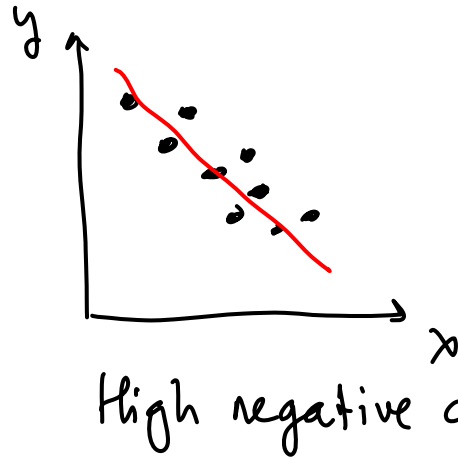
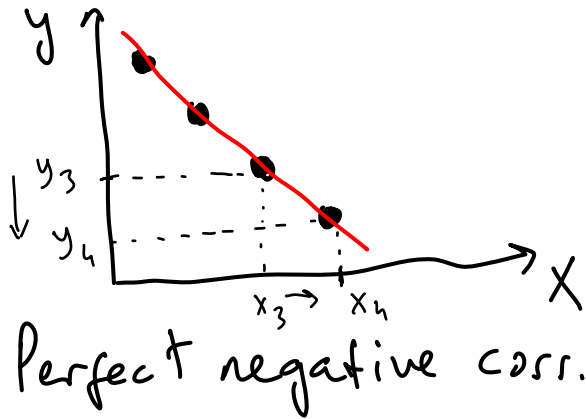
Positive Correlation

b) Perfect Negative Correlation

A distribution will show negative correlation if;

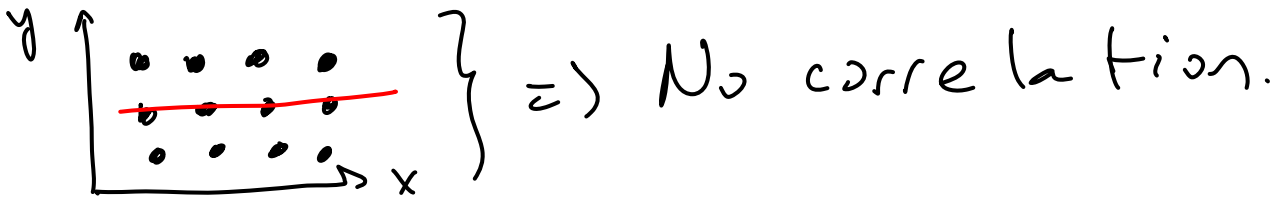
1. for each unit increase in one variable there is a fixed decrease in the other,
2. for each unit decrease in one variable there is a fixed increase in the other.

● experimental
— predicted



c) No Correlation

For each unit change (increase or decrease) in one variable, the other variable tends to stay about the same value.



① Correlation coefficient (r) changes between -1 and $+1$. $-1 \leq r \leq +1$.

If $r = +1 \Rightarrow$ perfect positive correlation.

$r = -1 \Rightarrow$ " negative " "

$r = 0 \Rightarrow$ no linear correlation.

If r is close to $+1 \Rightarrow$ the correlation may be high.
 r " " " $-1 \Rightarrow$ " " " " " "
 r " " " zero \Rightarrow " " " " very low.

The r Value for a Sample

$$r = \frac{\sum(x - \bar{x}) \cdot (y - \bar{y})}{(n-1) \cdot S_x \cdot S_y} \quad \text{OR alternatively,}$$

$$r = \frac{n \cdot (\sum x \cdot y) - (\sum x) \cdot (\sum y)}{\left(\sqrt{n \cdot (\sum x^2) - (\sum x)^2} \right) \cdot \left(\sqrt{n \cdot (\sum y^2) - (\sum y)^2} \right)}$$

$$S_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}, \quad S_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

\bar{x} and \bar{y} : means of x and y variables

S_x and S_y : st. dev. of x and y variables.

n : sample size.

Example: Using correlation coefficients, determine which of the following distributions show a stronger correlation between the variables.

Male teenagers and agility		Female teenagers and agility	
Age (x)	Agility score (y)	Age (x)	Agility score (y)
13	4	13	4
14	5	14	7
15	7	15	5
16	7	16	6
17	8	17	8

Solution:

Find r value for male \Rightarrow

$\frac{x}{}$	$\frac{y}{}$	$\frac{x \cdot y}{}$	$\frac{x^2}{}$	$\frac{y^2}{}$
13	4	52	169	16
⋮	⋮	⋮	⋮	⋮
17	8	136	289	64
$\frac{\sum x = 75}{}$	$\frac{\sum y = 31}{}$	$\frac{\sum (x \cdot y) = 475}{}$	$\frac{\sum x^2 = 1135}{}$	$\frac{\sum y^2 = 203}{}$

$$r_{\text{Male}} = \frac{5(475) - 75(31)}{(\sqrt{5(1135) - (75)^2}) \times (\sqrt{5(203) - (31)^2})}$$

$$r_{\text{Male}} = 0.96$$

Similarly find for females \Rightarrow

$$r_{\text{Female}} = 0.70$$

Since $r_{\text{male}} > r_{\text{Females}} \Rightarrow$ there is a stronger relationship between age and agility among the boys than there is for the girls.

B) Linear Regression

Objectives in regression analysis is the making of predictions.

e.g., predicting the shelf-life of a vegetable oil at any temperature.

$T (^{\circ}\text{C})$	Shelf-life (month)	
20	→ 18	}
30	→ 12	
40	→ 8	
⋮	⋮	
⋮	⋮	

$\Rightarrow y = f(T) \Rightarrow y = -ax + b$

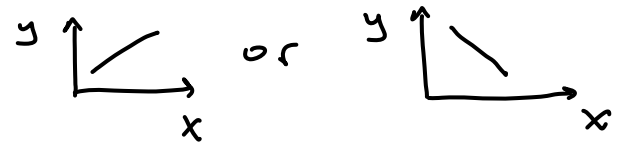
\downarrow
S.L

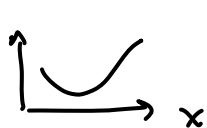
\downarrow
temp.


\downarrow
T

Always we want to find the best equation of the line of best fit to express the relationship of the two variables (Temperature, shelf-life, etc.)

* Some prediction equations:

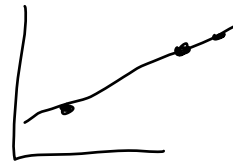
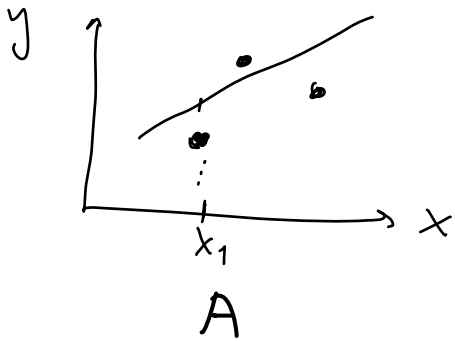
1) $y = b + ax \rightsquigarrow$ linear \rightarrow 

2) $y = a + bx + cx^2 \rightsquigarrow$ quadratic \rightsquigarrow 

3) $y = a \cdot (b^x) \rightsquigarrow$ exponential \rightsquigarrow 

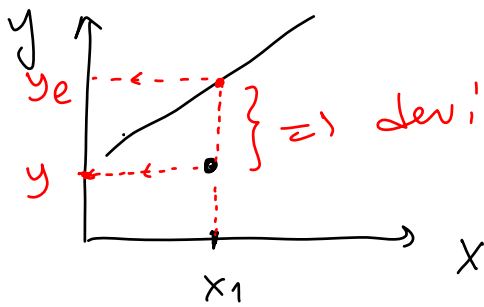
If a straight-line relationship seems appropriate, the best-fitting straight line is found by using the method of LEAST SQUARES.

This method results in a line which reduces the sum of the squared deviations of observed y -values from the regression line (predicted) to a **MINIMUM**.



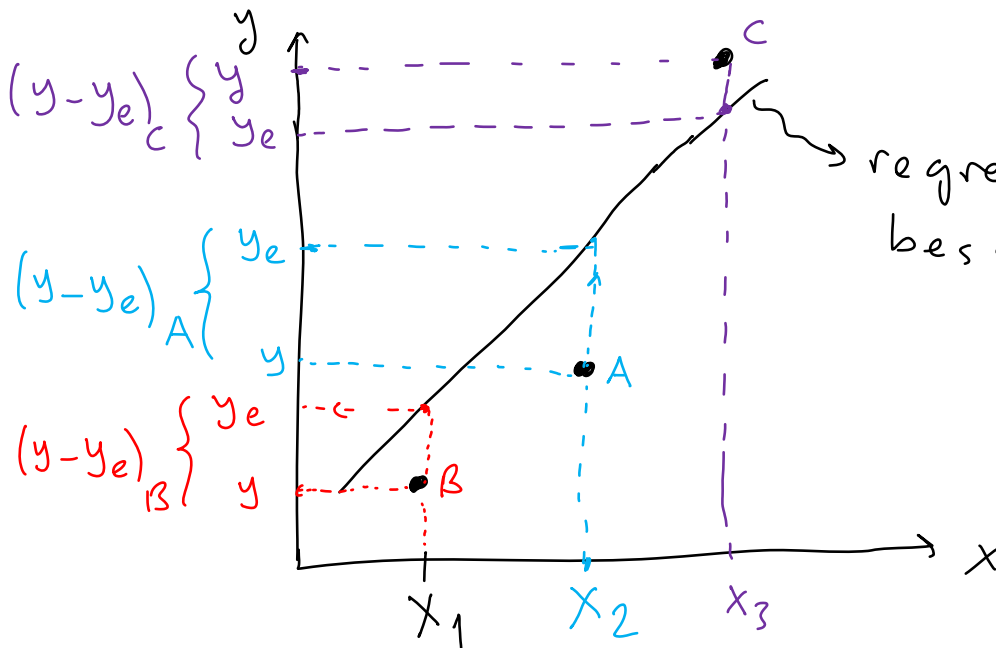
B \Rightarrow line B is better than A.

The **deviation** or **error of estimation** between observed (y) and predicted (y_e) values of y for a specified value of x is $(y - y_e)$.



\Rightarrow deviation = $y - y_e$ = error of estimation.

In the least squares method $\underbrace{\sum (y - y_e)^2}_{\text{sum of squares}}$ is a minimum.

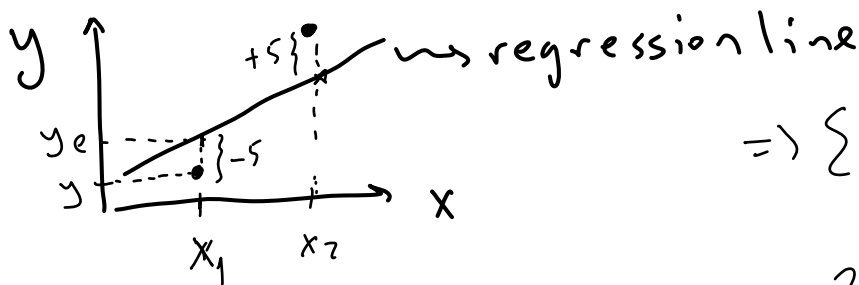


regression line or best fit line

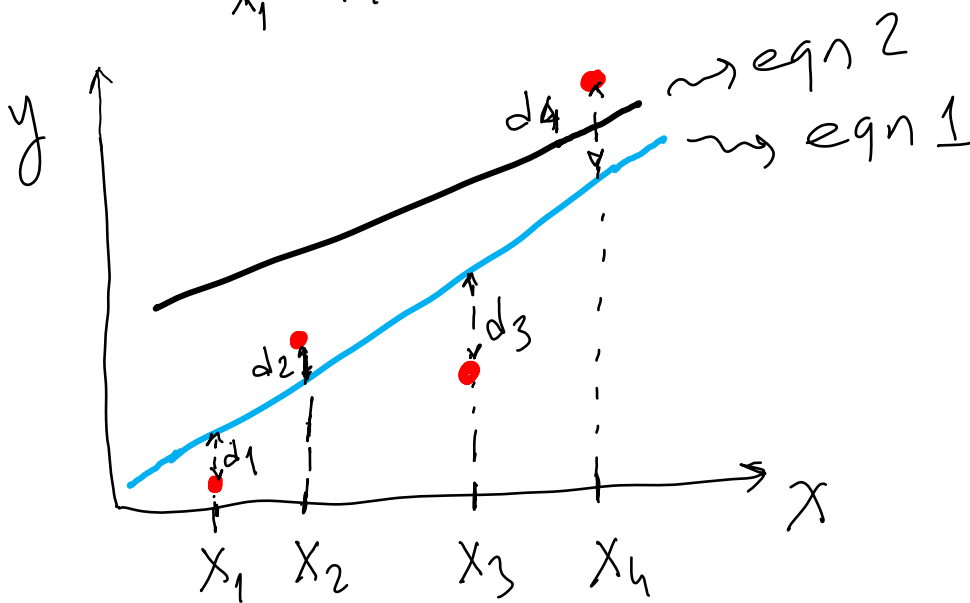
$$\sum (y - y_e)^2 = (y - y_e)_A^2 + (y - y_e)_B^2 + (y - y_e)_C^2$$

should be minimum (as small as possible)

If $\sum (y - y_e) = 0 \Rightarrow$ the (+) and (-) deviations just balance out.



$$\Rightarrow \sum (y - y_e) = -5 + 5 = 0$$

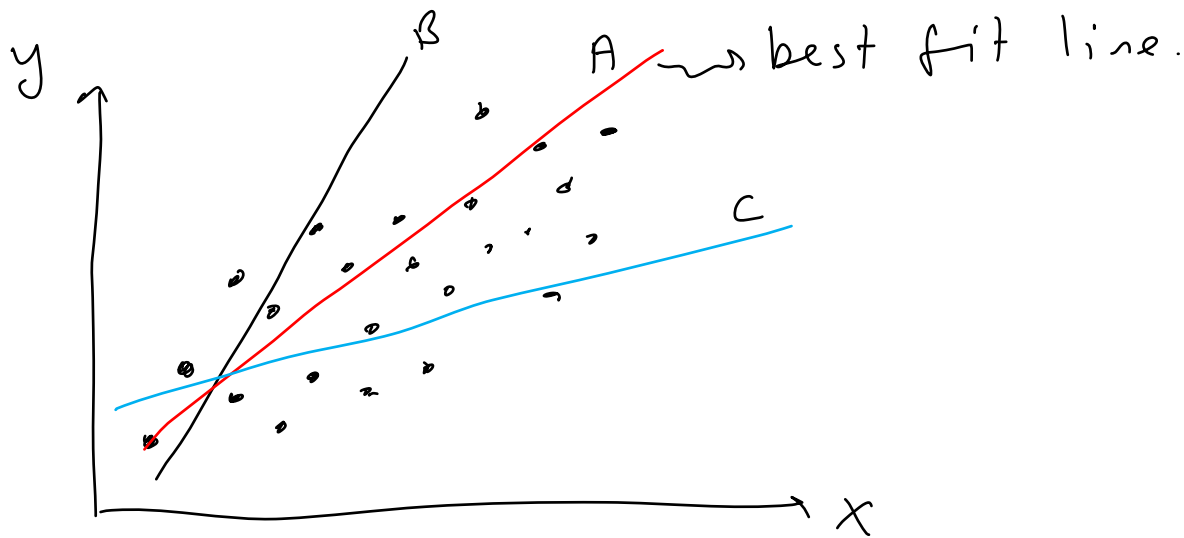


D for eqn 2 $>$ D for eqn 1 \Rightarrow

eqn 1 is the best eqn; because

$$\sum D = \sum (y - y_e)^2 \text{ is minimum.}$$

$y_e = b_0 + b_1 \cdot X \rightsquigarrow$ best fit line equation.



$$\sum (y - y_e)_A^2 < \sum (y - y_e)_C^2 < \sum (y - y_e)_B^2$$

The line A is the best fit line

IN SUMMARY:

- **The Least Squares Method:** It is some way putting a line in between experimental data points. It somehow minimizes the overall distance from the line to all the experimental data points. That is why we call it Least Squares.
- The line (best-fit line) should be as close as possible to the points.
- The best linear model (equation) will have the smallest value of D.

$$\sum D = \sum (y - y_e)^2$$

- Don't worry about positive and negative distances. We square them.