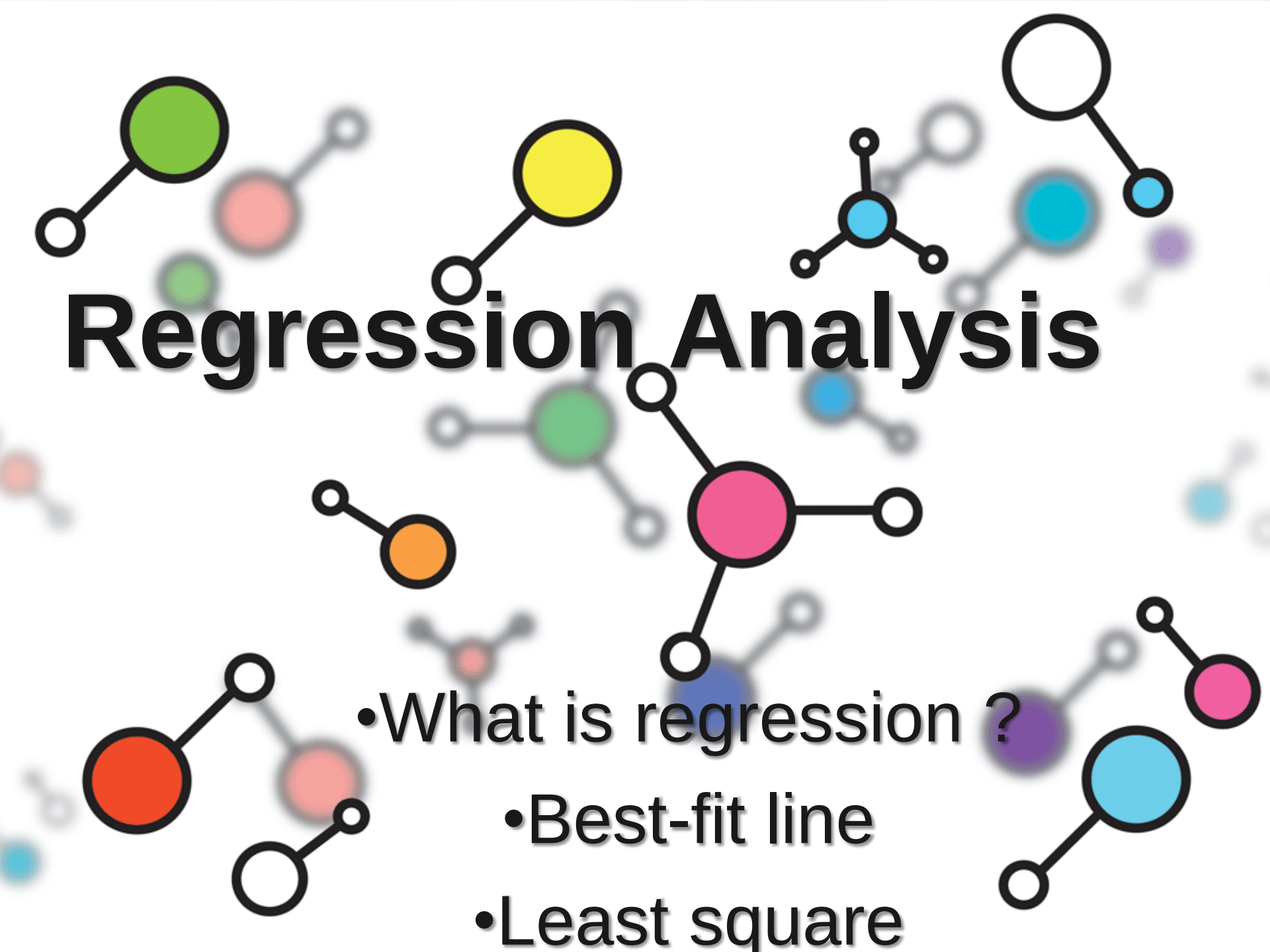
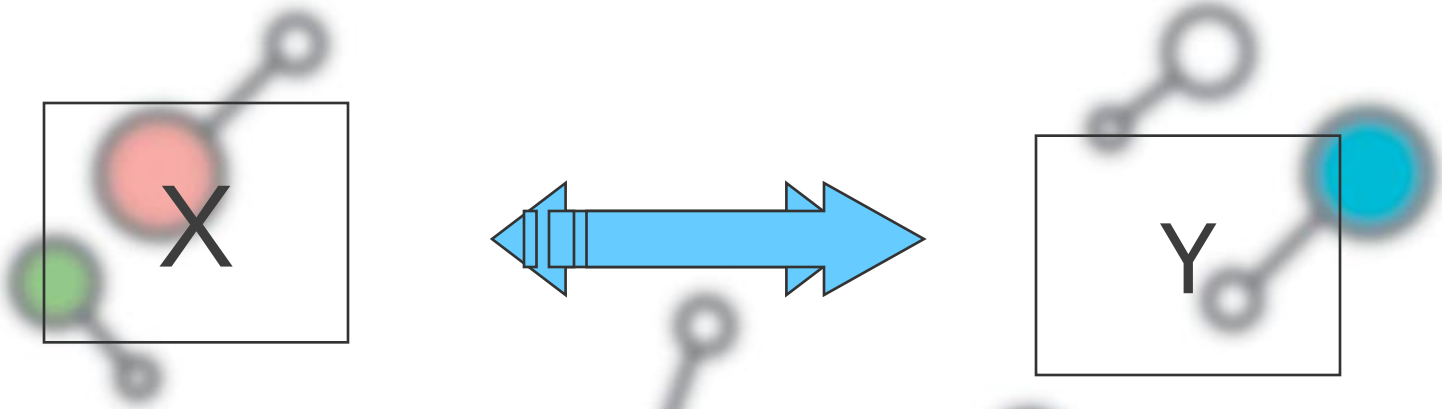


Regression Analysis

- What is regression ?
- Best-fit line
- Least square

What is regression ?

- **Regression models are statistical models which describe the variation in one (or more) variable(s) when one or more other variable(s) vary.**
- **Inference based on such models is known as regression analysis.**



- Predictor
- independent var

- Respons var
- Dependen var

Regression Analysis

- Can we say that X predict Y ?

→ Used to investigated how dependent variable can be predicted by independent variable as a partial or whole of the variable

Simple linear regression models

- Suppose that we have a response variable and an explanatory variable , then the simple linear regression model for on is given by

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

where β_0, β_1 are unknow parameters

ε_i are independent random variable with 0 mean and constant variance for all $i = 1, \cdots, n$

β_0, β_1 : unknow parameters,

non random parameters, the intercep and the slop

$\Rightarrow$ regression parameter/coeff

$h(x) = \beta_0 + \beta_1 x$ called the regression line

/ linear predictor

Kind of Regression Analysis

I. Linier Regression

- Simple Linier Regression → $\hat{Y} = a + bX$
- Multiple Linier Regression → $\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3$

II. Non Linier Regression

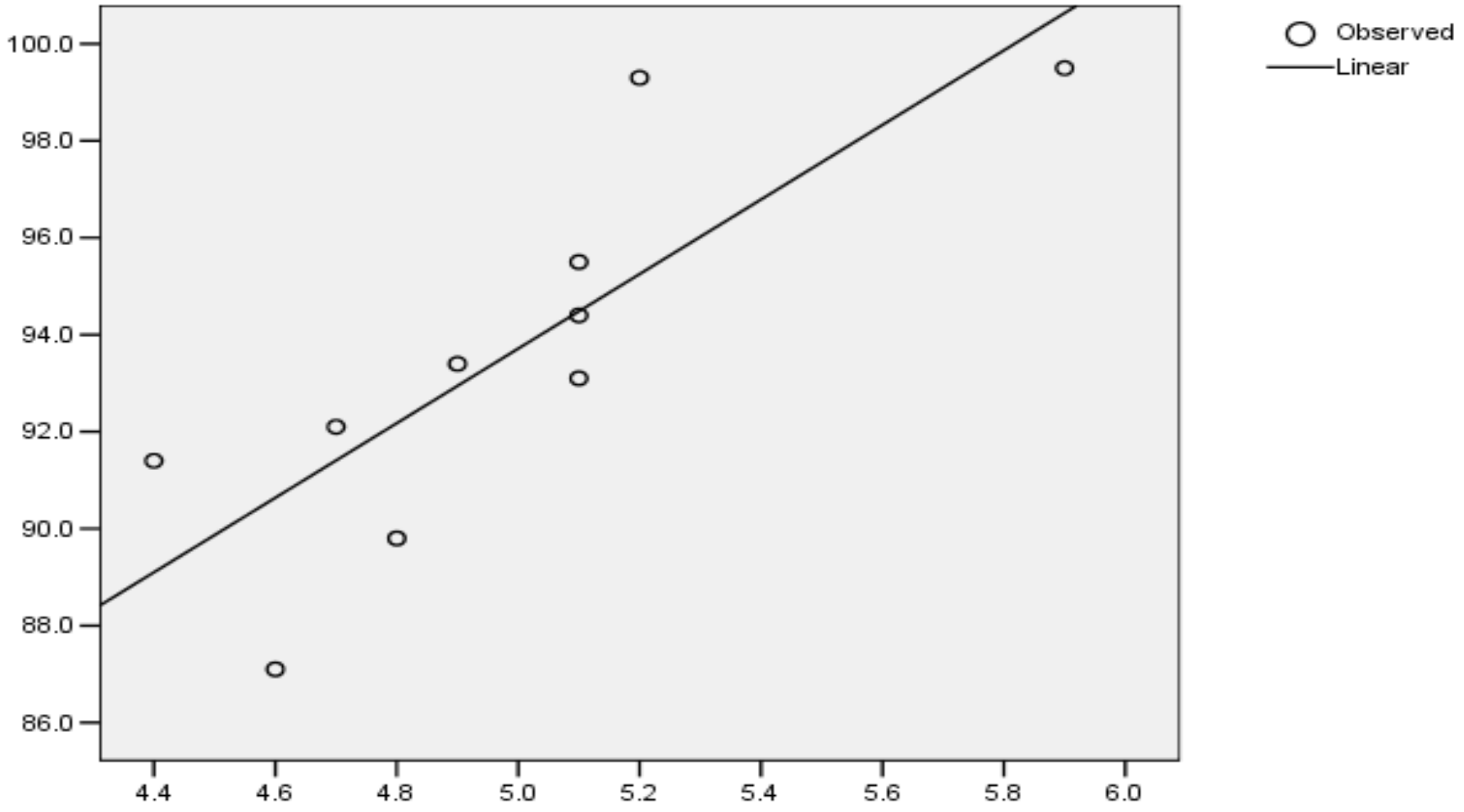
- Quadratic Regression → $\hat{Y} = a + bX + cX^2$
- Qubic Regression → $\hat{Y} = a + bX^2$

$\hat{Y} = a + bX + cX^2 + dX^3$

$\hat{Y} = a + bX^2 + cX^3$

$\hat{Y} = a + bX^3$

Simple Linier Regression



- X_i independent variable in the-*i th*
- Y_i dependent variable in the -*i*

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1,2,\ldots,n$$

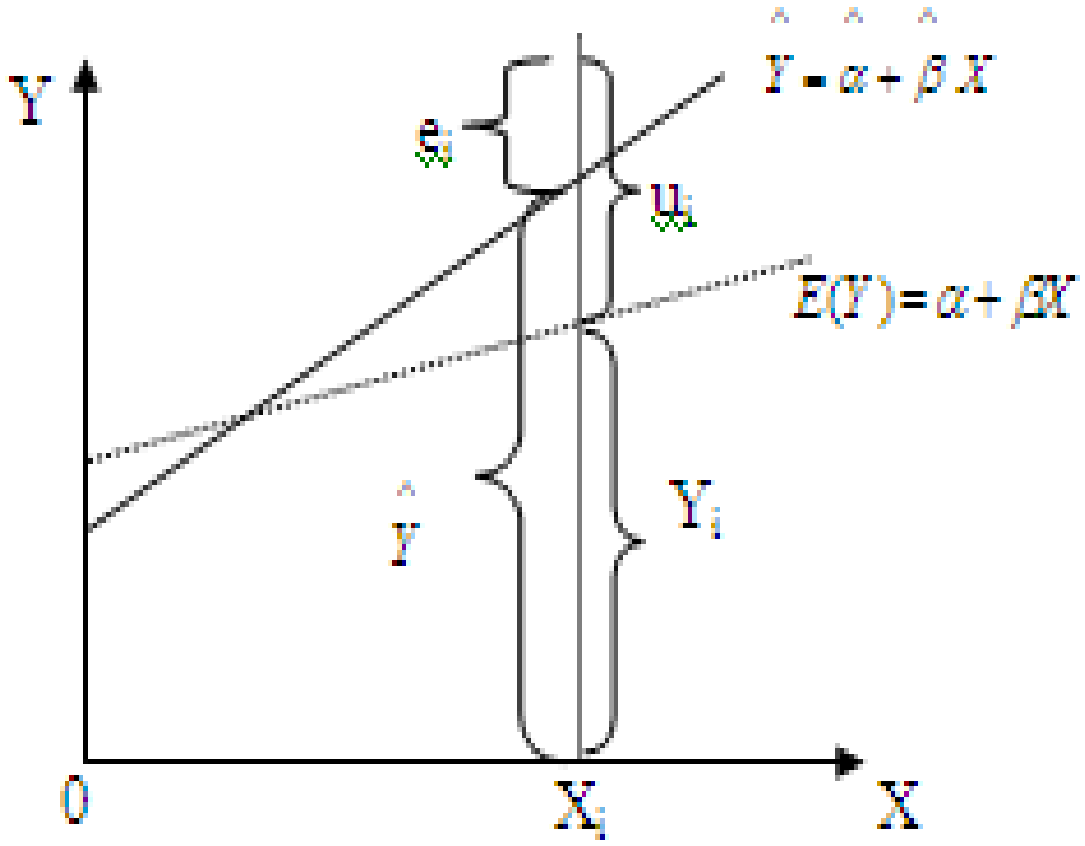
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X$$

$$\hat{Y}_i = a + bX$$

$\hat{\alpha}, \hat{\beta}$ atau a, b unknown parameter

ε_i random error with asumtion

$$E[\varepsilon_i] = 0 \qquad Var(\varepsilon_i) = \sigma^2$$



- **FROM SIMPLE LINIER REGRESSION WE'LL FIND**

$$e_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i)$$

AND

$$D = \sum e_i^{\,2} = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2$$

- **Deviate D as partial to a and b !!!!**

- So we'll get :

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$a = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n\sum X^2 - (\sum X)^2}$$

- **The test of regression model**
  - **partial (per- coefisient) → t- Test**
  - **Whole → F-Test (Anova)**

- **how well a fitted line describes the variation in data: the coefficient of determination ; R^2 % variation of Y can be explain by X)**

Alternatively…

y	x	xy	x^2	y^2
.	.	.	.	.
.	.	.	.	.
Σy	Σx	Σxy	Σx^2	Σy^2

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$a = \bar{y} - b\bar{x}$$

$$\bar{y} = \frac{\sum y}{n} \qquad \bar{x} = \frac{\sum x}{n}$$

Example

Find the estimation regression function from the data bellow:

Mat (X)	Fis (Y)	XY	X2	Y2
60	80	4800	3600	6400
45	69	3105	2025	4761
50	71	3550	2500	5041
60	85	5100	3600	7225
50	80	4000	2500	6400
65	82	5330	4225	6724
60	89	5340	3600	7921
65	93	6045	4225	8649
50	76	3800	2500	5776
65	86	5590	4225	7396
45	71	3195	2025	5041
50	69	3450	2500	4761
665	951	53305	37525	76095

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$= \frac{53305 - \frac{(665)(951)}{12}}{37525 - \frac{(665)^2}{12}} = 0.8972$$

$$a = \bar{y} - b_1 \bar{x} = 29.53$$

So:

$$\hat{Y}_i = 29.5294 + 0.8972 X_i$$

ANOVA APPROACH TO REGRESSION ANALYSIS

Look at here:

$$\underbrace{(y_i - \bar{y})}_{\text{variation}} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{regression}} + \underbrace{(y_i - \hat{y}_i)}_{\text{error}}$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SS_T} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SS_R} + \underbrace{2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)}_{=0} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SS_E}$$

ANOVA TABLE

Source Of Variation	SS	df	MS	F
Regression	SSR= $b^2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$	1	MSR=SSR/1	F=MSR/MSE
Error	SSE= SST-SSR	n-2	MSE=SSE/n-2	Ftabel F(alpha, 1,n-2)
Total	SST= $\sum_{i=1}^{n} y_i^2-\frac{\left(\sum y_i\right)^2}{n}$	n-1		

ANOVA TABLE:

Variation Source	SS	df	MS	F
Regression	SSR= $b^2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$	1	MSR=SSR/1	F=RKR/RKS
Error	SSE= SST-SSR	n-2	SSE=SSE/n-2	Ftable F(α, 1,n-2)
Total	SST= $\sum_{i=1}^{n}y_i^2-\frac{\left(\sum y_i\right)^2}{n}$	n-1		

Previously Ex

1. Hypothesis :

H_0 : X and Y linier relation isn't significant

H_1 : X and Y linier relation is significant

2. α=5%
3. Arrange ANOVA table
4. Conc : Reject H0 if F> F table

SV	SS	df	MS	F
Regression	541.193	1	541.193	29.04
Error	186.557	12-2=10	18.6557	Ftable F(α, 1,n-2)
Total	728.25	12-1=11		

Reject H0 because F=29.04>Ftable=4.96

So linier regression X and Y is significant

Coef Regression Sig Test

1. Hypothesis

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

2. Take any α
3. Conc : Reject H0 if t> t table

$$t = \frac{b}{s_b}$$

$$s_b = \sqrt{\frac{s_{y.x}^2}{c}}, \quad c = \sum x^2 - \frac{\left(\sum x\right)^2}{n} \qquad s_{y.x} = \sqrt{\frac{SS_E}{n-2}}$$

Pre Ex

1. Hypothesis

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

2. α=5%

3. Conc : Reject H0 if t> t table=t (α/2, n-2)

$$b = 0.8972$$

$$s_b = 0.166504$$

$$t = \frac{0.8972}{0.166504} = 5.388$$

Because of t=5.388>t table 2.228 then H0 is rejected so coef b is significant.

$$\sum x^2 = 37525, \quad \left(\sum x\right)^2 = (665)^2 = 442225$$

$$s_b = \sqrt{\frac{s_{y.x}^2}{c}}, \quad c = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 37525 - \frac{442225}{12} = 672.9167$$

$$s_b = \sqrt{\frac{s_{y.x}^2}{c}} = \sqrt{\frac{18.6557}{672.9167}} = 0.166504$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.862[a]	.744	.718	4.319

a. Predictors: (Constant), matematik

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	541.693	1	541.693	29.036	.000[a]
	Residual	186.557	10	18.656		
	Total	728.250	11			

a. Predictors: (Constant), matematik

b. Dependent Variable: fisika

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	29.529	9.311		3.171	.010		
	matematik	.897	.167	.862	5.389	.000	1.000	1.000

a. Dependent Variable: fisika