

## **Linear Regression: Least Squares Concept**

### **Regression:**

In real life, we may try to **determine the relationship between one quantity  $x$  and another quantity  $y$** . For example,  $x$  can be a person's height and  $y$  can be that person's shoe size. If we record  $(x, y)$  on a coordinate system, we may be able to find that the higher person has the larger shoe size. By some statistics methods, we will be able to estimate the shoe size for a person if his height is known.

Based on existing data, we can try to draw a curve line to represent the data set. After that, we can predict future information according to the curve line we drew. **The process of collecting data, making relationship among data and predicting some information is called regression.**

**The Correlation Coefficient:** The strength and direction of the relationship between  $x$  and  $y$  are measured using the **correlation coefficient,  $r$** .  $-1 \leq r \leq 1$ .

If  $r \approx 0$ , then there is a weak relationship.

If  $r \approx 1$  or  $r \approx -1$ , then there is a strong relationship.

$$r = \frac{S_{xy}}{S_x S_y} \text{ where } S_{xy} = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)}{n-1}, S_x = \text{std. dev. of the } x\text{'s}, S_y = \text{std. dev. of the } y\text{'s}.$$

### **Example:**

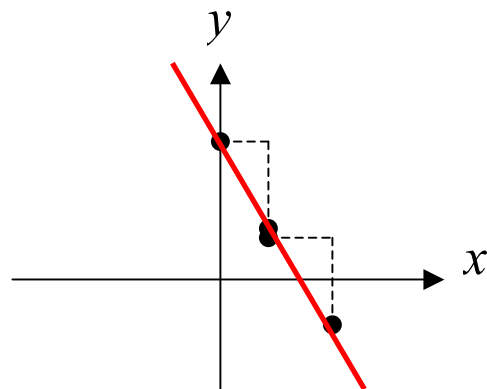
**Review** for a **linear equation** and a line:

Use the **slope and y-intercept** method to sketch the line for the equation,  $y = 3 - 2x$

From the equation, we got the following information:

$$\begin{aligned} \text{slope} &= -2 \\ \text{y-intercept} &= 3. \end{aligned}$$

Therefore, we will first locate y-intercept on the coordinate system. Then, use slope information to run 1 unit to right and fall 2 units down.



## Linear Regression:

The equation of the **least squares line** is  $y = b_0 + b_1 x$

where the **slope**  $b_1$  of the line is given by  $b_1 = \frac{SS(xy)}{SS(x)} = \frac{(\sum xy) - \frac{(\sum x)(\sum y)}{n}}{(\sum x^2) - \frac{(\sum x)^2}{n}}$

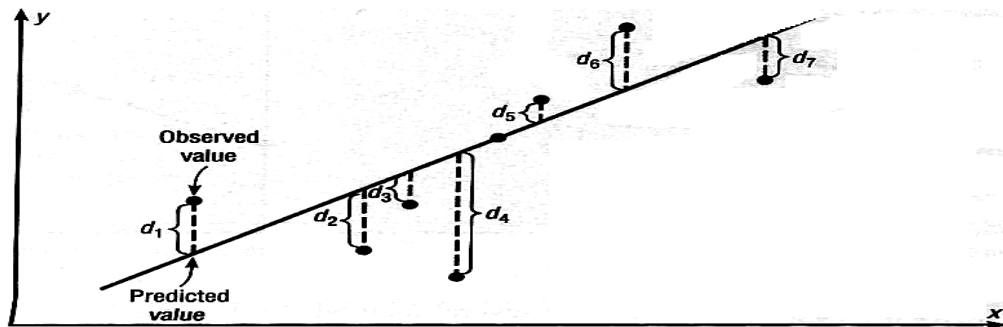
and the **y-intercept**  $b_0$  of the line is given by  $b_0 = \bar{y} - b_1 \bar{x}$

**Note:** The method of **least squares** is finding the value of  $b_0$  and  $b_1$  as

$$Q = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2 \text{ has minimum value.}$$

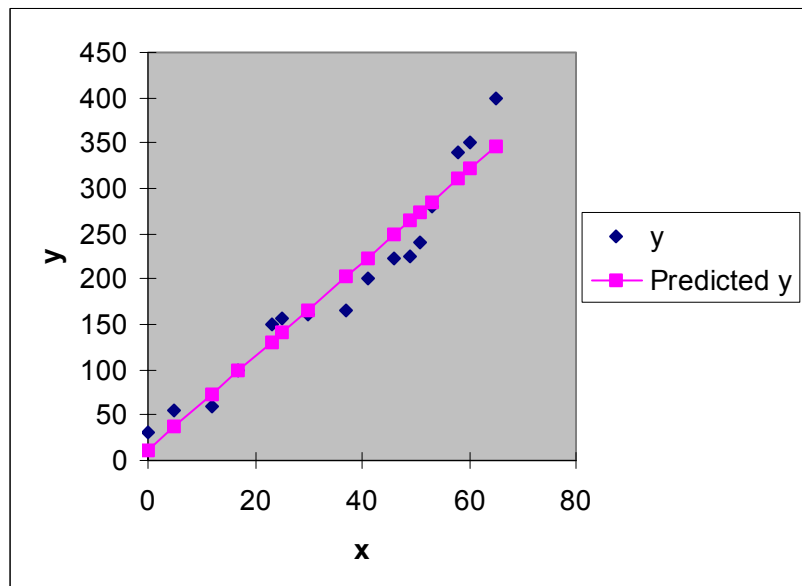
**Example:** (Note: This is copied from Bluman's Elementary Statistics)

**Line of Best Fit for a Set of Data Points:**

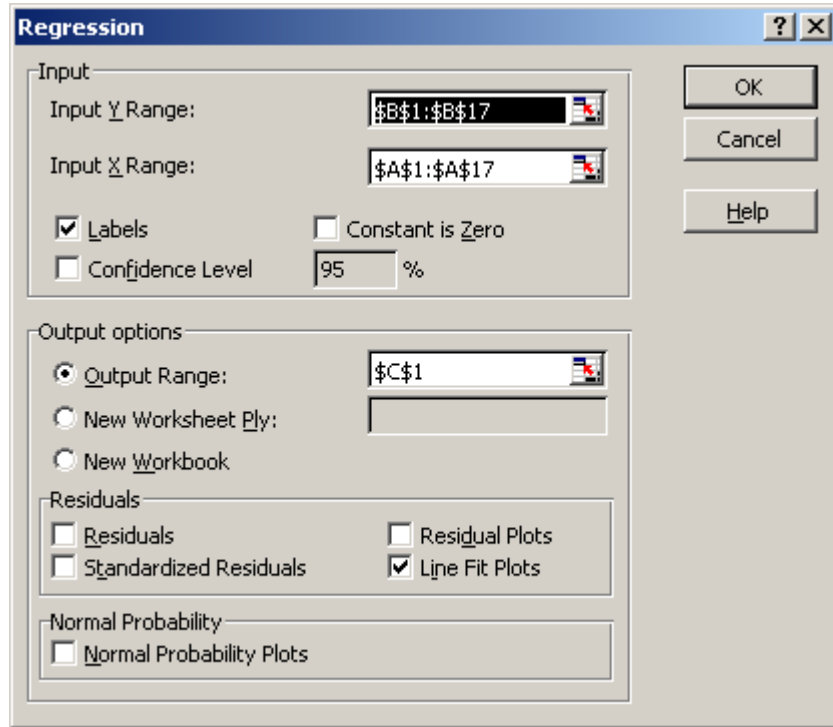
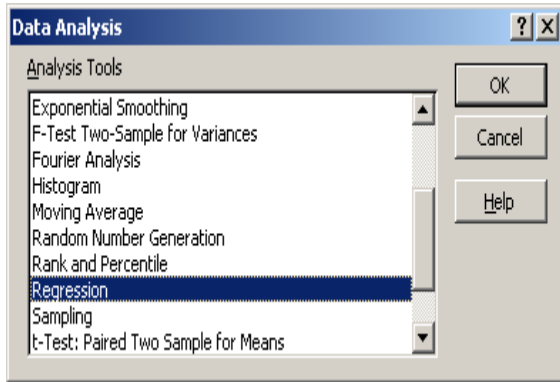


Example: Enter the following data:

x	y
0	30
5	55
12	60
17	100
23	150
25	157
30	160
37	165
41	200
46	222
49	226
51	240
53	280
58	339
60	350
65	400



1. Click **Data Analysis** from **Tool** menu.
2. In **Data Analysis**, select **Regression** and click the [OK] button.
3. In the **Regression** dialog box, select the **Input Y Range**, **Input X Range**, **Labels**, **Output Range**, and **Line Fit Plots** and click the [OK] button.  
**Note:** *Input X Range*: independent variable range *Input Y Range*: dependent variable
4. Double click on predict points to connect points as a straight line.



The results are shown below.

Regression Statistics	
Multiple R	0.967507
<b>R Square</b>	<b>0.936071</b>
Adjusted R Square	0.931504
Standard Error	28.46243
Observations	16

**R-square is closed to 1, GOOD FITTED**

**p-value > 0.05**, do not reject the null hypothesis that intercept=0

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	11.28528	14.72582	0.76636	<b>0.456191</b>	-20.2985	42.86905	-20.2985	42.86905
x	5.163349	0.360631	14.31755	9.41E-10	4.389872	5.936826	4.389872	5.936826

The regression line is  $y = 5.163349x + 11.28528$

For a linear regression, R-Square =  $r^2$  where  $r$  is a **correlation coefficient**