

WHAT IS THE LEAST SQUARES LINE?

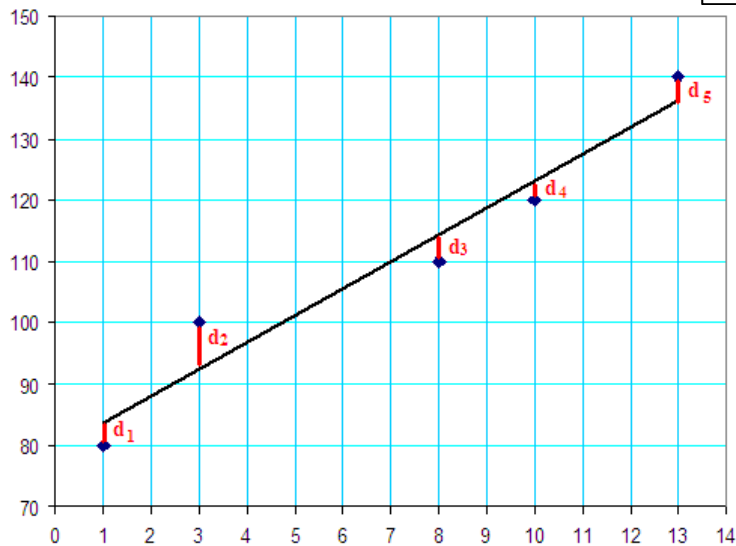
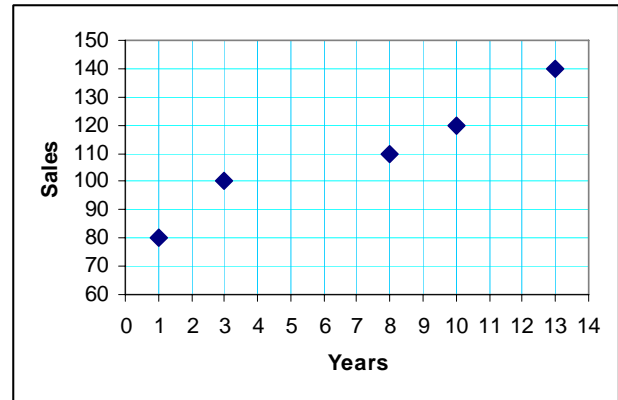
Problem:

A sales manager noticed that the annual sales of his employees increase with years of experience. To estimate the annual sales for his potential new sales person he collected data concerning annual sales and years of experience of his current employees: We'll use his data to create a formula that will help him estimate annual sales based on years of experience.

Years of experience	1	3	8	10	13
Annual sales in thousands	80	100	110	120	140

Work:

Figure to the right shows scatter graph of his data. Each point represents data on one single current employee. For example: the first employee has 1 year of experience and made 80 thousands in sales, so he is represented by the point with coordinates (1,80).



We'll create equation of the **least squares line**, which is also called **best-fit line** or **regression line**. The line is passing in between our points while the sum of the squares of the vertical distances from the data points to the line is as small as possible. The picture to the left shows the least square line passing in between our points, and the distances d_1 , d_2 , d_3 , d_4 , and d_5 . The equation of the least square line is found by minimizing the sum:

$$(d_1)^2 + (d_2)^2 + (d_3)^2 + (d_4)^2 + (d_5)^2$$

The procedure will be omitted in this paper. The final result of the minimization are formulas that let us calculate coefficients a and b for equation of the least square line $y = ax + b$. This equation may be used for the prediction of sales.

FORMULAS FOR COEFFICIENTS OF THE LEAST SQUARES LINE:

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

In this formula n stands for number of observed cases. In our case that is 5 employees. So: $n = 5$

Symbol \sum is called "sigma" and stands for the "sum of all". In our case:

$$\sum x = 1 + 3 + 8 + 10 + 13 = 35 \quad (\text{sum of all } x \text{ values})$$

$$\sum y = 80 + 100 + 110 + 110 + 120 + 140 = 550 \quad (\text{sum of all } y \text{ values})$$

$$\sum x^2 = 1^2 + 3^2 + 8^2 + 10^2 + 13^2 = 343 \quad (\text{sum of squares of } x \text{ values})$$

$$\sum xy = 1 \cdot 80 + 3 \cdot 100 + 8 \cdot 110 + 10 \cdot 120 + 13 \cdot 140 = 4280 \quad (\text{sum of } xy \text{ products})$$

We are now ready to use results of our calculations in the formula:

$$a = \frac{5 \cdot 4280 - 35 \cdot 550}{5 \cdot 343 - 35^2} = \frac{2150}{490} \approx 4.388 \quad b = \frac{550 - 4.388 \cdot 35}{5} = \frac{396.42}{5} \approx 79.284$$

That means that the equation of the least square line is $y = 4.388x + 79.284$

Conclusion:

We created the formula $y = 4.388x + 79.284$ that may be used to predict sales in thousands for a future employee. For example, formula may predict that an employee with $x=15$ years of experience will generate:
 $y = 4.388 \cdot 15 + 79.284 \approx 145$ thousands in sales per year.

Is our prediction reliable?

Once an equation is found for the least square line, we need to have some way of judging just how good the equation is for predictive purposes. In order to have a quantitative basis for confidence in our predictions, we need to calculate **coefficient of correlation**, denoted r . It may be calculated using the following formula:

<p>FORMULA FOR COEFFICIENT OF CORRELATION</p> $r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \cdot \sqrt{n(\sum y^2) - (\sum y)^2}}$

We'll calculate the coefficient of correlation for data in our example:

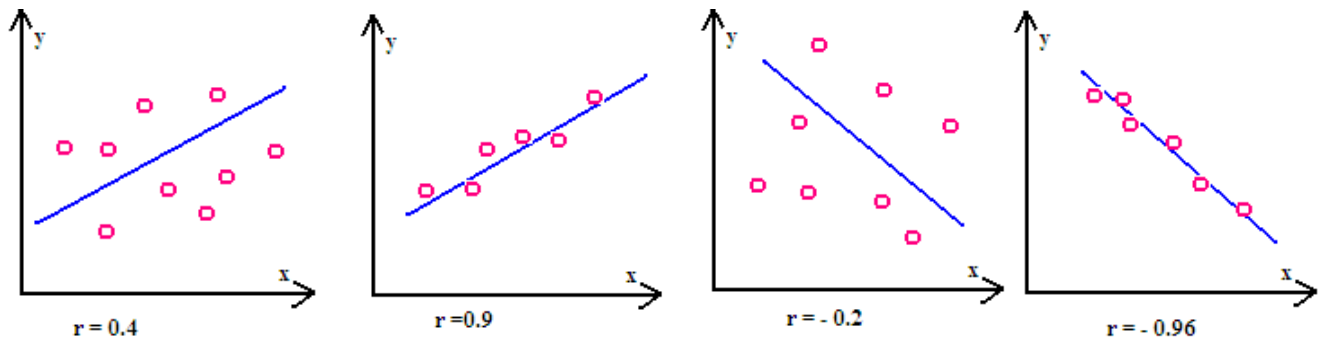
$$\sum y^2 = 80^2 + 100^2 + 110^2 + 120^2 + 140^2 = 62500$$

$$r = \frac{5 \cdot 4280 - 35 \cdot 550}{\sqrt{5 \cdot 343 - 35^2} \cdot \sqrt{5 \cdot 62500 - 550^2}} = \frac{2150}{\sqrt{490} \cdot \sqrt{10000}} \approx \frac{2150}{2213.594} \approx 0.971$$

The value of r that is close to 1 indicates that our formula will give us a reliable prediction for sales level, based on years of experience of the employee.

What does coefficient of correlation tell us?

The correlation coefficient is always a number between -1 and 1 . The picture bellow shows how its value numerically describes our data:



The equation may be used as a source for reliable prediction if the correlation coefficient is a number that is close to -1 or 1 . That means that your observed values are close to the least square line (*second and fourth picture above*). If not, the value of the correlation coefficient is closer to 0. Such a small value for the coefficient of correlation indicates that the observed data are widely spread around, so our formula is not reliable source of prediction (*like on first and third picture above*).

It is usually more convenient to numerically measure reliability of our formula using the square of correlation coefficient. Some books and computer software are using symbol R^2 for it (read as 'r square') In our example;

$$R^2 = 0.971^2 \approx 0.943$$

R^2 is always a number between 0 and 1. Values of R^2 that are close to 1 indicate reliable formula.

FINDING THE LEAST SQUARES LINE USING EXCEL

Problem:

A sales manager noticed that the annual sales of his employees increase with years of experience. We'll use Excel to graph his data and create a formula that will help him estimate annual sales based on years of experience.

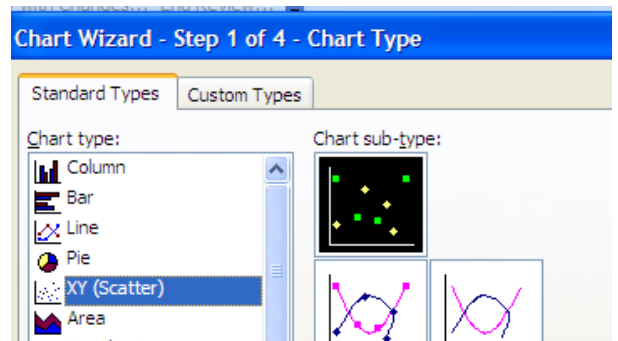
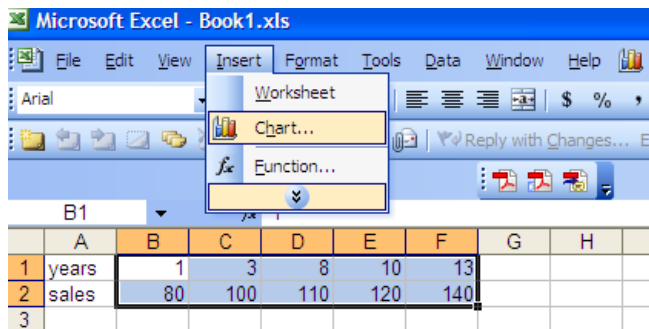
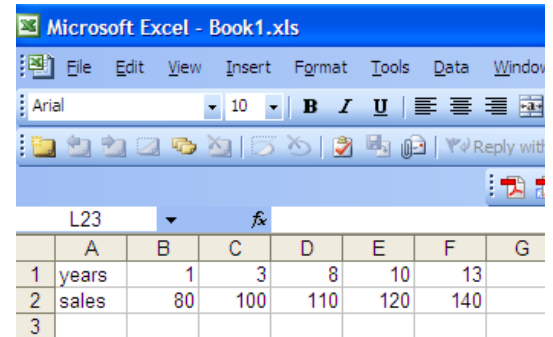
Years of experience	1	3	8	10	13
Annual sales in thousands	80	100	110	120	140

Work

Open a new blank sheet in Excel and type in data as **shown to the right**.

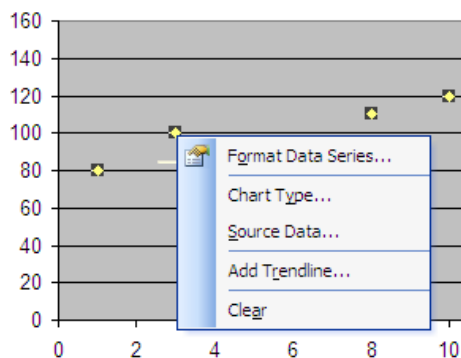
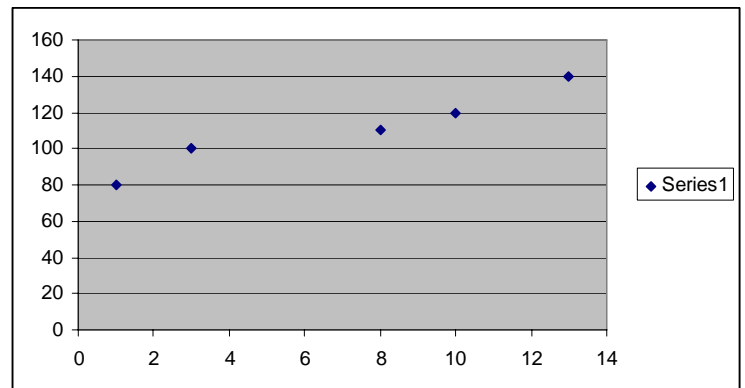
Step in cell B1. While keeping left finger down, pool cursor to cell F2. This procedure will highlight (color blue) coordinates of points that should be graphed. Once your data are highlighted, click on *insert* and select *chart* from the menu as **shown bellow left**.

New Chart Wizard pop-up screen will ask you what kind of chart do you want. Select *XY(Scatter)* that compares pairs of values as **shown bellow right**.



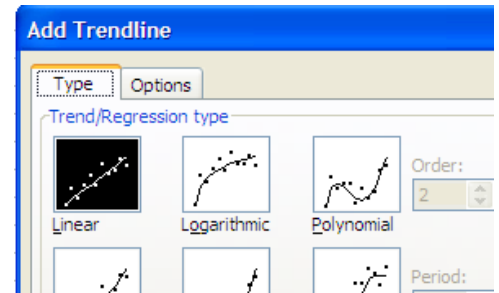
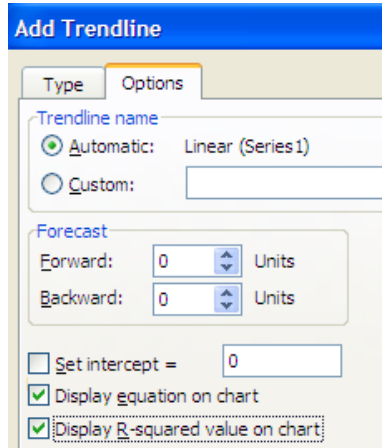
Click on *finish*. A simple scatter graph of your data will appear, as **shown to the right**.

The next step is to draw the least squares line and calculate its equation and R square. Carefully right-click on any data point on the graph. A small pop-up screen will come out as **pictured bellow**. Select *Add trendline*.

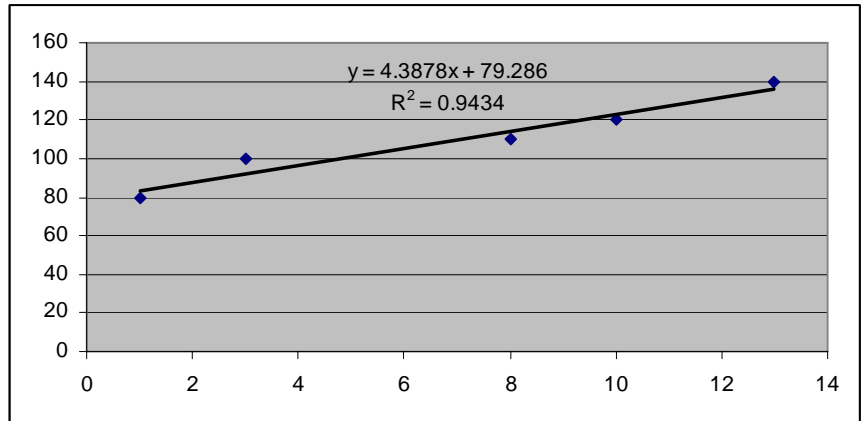


A new pop-up screen will come out. Select *Linear Trend* and click on tab *Options* as on **the picture to the right.**

In *Options* tab check *Display equation on chart* and check *Display R-Squared value* as shown on **picture bellow to the left.** Click *OK* to obtain the final picture.



The final picture shown bellow features the equation of the least squares line $y = 4.3878x + 79.286$ and the value of R square $R^2 = 0.9434$



Practice questions:

- 1) The table below gives monthly sales y (in thousands), corresponding to advertising expenditures x (in thousands).

Advertising expenditures x	0	1	2	3	4	5	6
Monthly sales y	3	9	11	15	17	20	27

Create the Least Squares Line. Examine if advertising expenditures appear to have strong effect on monthly sales by calculating and interpreting R^2 , and if so, predict the monthly sales if 11 thousands dollars is spent on advertising.

(Answer: $y = 3.57x + 3.86$ and $R^2 = 0.97$, so if we spend 11 thousand dollars on advertising the monthly sales will increase to 43 thousand dollars.)

- 2) A hospital conducted study to determine relation between age and blood pressure of their patients. The table bellow shows collected data. Find the equation of the least squares line and use it to calculate blood pressure of a 50 years old patient.

Age x	43	48	56	61	67	70
Pressure y	128	120	135	143	141	152

(Answer: $y = 0.964x + 81.048$, so 50 years old patient should have 129.)

- 3) A researcher wishes to see whether there is a relationship between number of hours of study and test scores on exam, so she collected data shown in the table bellow. Find the equation of the least squares line and use it to calculate how many hours should a student study to obtain 93 percent on the test.

Hours of study x	6	2	1	5	2	3
Score on test y	82	63	57	88	68	75

- 4) The table bellow compares rents for one-bedroom and two-bedroom apartments in 7 different cities. Find the equation of the least squares line and R square. A one-bedroom rent in a doorman building in Lower Manhattan averaged \$3000 in august 2007. Calculate a two bedroom rent using the obtained formula.

One-bedroom rent x	782	486	451	529	618	520	845
Two bedroom rent y	1223	902	739	954	1055	875	1455