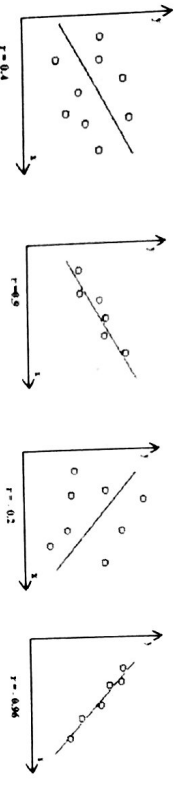


THE LEAST SQUARES LINE ('Line of Best-Fit Line')

What does coefficient of correlation tell us?

The correlation coefficient is always a number between -1 and 1. The picture below 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's sometimes more convenient to numerically measure reliability of our formula using the square of correlation coefficient. The symbol r^2 (read as 'r square') is called the **coefficient of determination**.
 - The coefficient of determination is always a number between 0 and 1.
 - Values of r^2 that are close to 1 indicate a reliable formula.
 - Values of r^2 that are close to 0 indicate an unreliable formula.
 - Values of r^2 that are closer to 0.5 indicate a moderately unreliable formula, where the equation's prediction dependability increases as it approaches 1 and decreases as it approaches 0.

Practice questions:

- 1) Use the table to answer the following questions.

x	3	5	7	8
y	8.7	7.9	6.2	5.8

a. What is the equation of the regression line?
 $y = -0.61x + 10.68$
 $r^2 = -0.99$

b. Find the coefficient of determination (r^2) and interpret its meaning (reliability).
 $r^2 = 0.98$
 yes 98% of variation in y can be explained

- 2) The table to the right shows the relationships between MPG (miles per gallon) and HP (horsepower) for some sports cars produced in year 2015.

	x=HP	y=MPG
Mitsubishi Lancer	148	29
Mazda Maza	158	23
Subaru WRX	268	24
Porsche Boxter	315	23
Chevy Camaro	323	20
Ferrari Spider	570	14

a. What is the equation of the regression line?
 $y = -0.03x + 30.89$
 $r = -0.91$
 Find the coefficient of determination (r^2) and interpret its meaning (reliability).
 $r^2 = 0.83$
 83% of variation in mpg can be explained by HP.

yes

- d. Estimate the MPG for a Chevrolet Corvette that has 460 HP engine.
 3) The Millers are buying a house in Atlanta so they'd like to estimate the yearly house taxes in Inman Park. Using data from a real estate web site they obtained 2014 taxes for 5 houses in that area.

x=house price in thousands	425	340	140	250	70
y=tax paid in year 2014	3570	2370	1000	1700	810

a. What is the equation of the regression line?
 $y = 7.58x + 34.07$
 $r = 0.97$
 b. Find the coefficient of determination (r^2) and interpret its meaning (reliability).
 $r^2 = 0.94$
 94% of change in tax paid can be explained by house price.
 c. Do you consider the equation to be reliable to predict further data (extrapolate)?
 yes
 d. Estimate yearly taxes for a \$200,000 house in Inman Park.
 $y = 7.58(200) + 34.07 = 15194.07$

- 4) 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

a. What is the equation of the regression line?
 $y = 3.57x + 3.86$
 $r = 0.99$
 b. Find the coefficient of determination (r^2) and interpret its meaning (reliability).
 $r^2 = 0.97$
 97% of change in monthly sales can be explained by advertising expenditure.
 c. Do you consider the equation to be reliable to predict further data (extrapolate)?
 yes

d. Predict the monthly sales if 11 thousand dollars are spent on advertising.
 $y = 3.57(11) + 3.86 = 43.13$

- 5) A hospital conducted study to determine relation between age and blood pressure of their patients. The table below shows collected data.

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

a. What is the equation of the regression line?
 $y = 0.96x + 81.05$
 $r = 0.90$
 b. Find the coefficient of determination (r^2) and interpret its meaning (reliability).
 $r^2 = 0.80$
 80% of change in pressure can be explained by age.
 c. Do you consider the equation to be reliable to predict further data (extrapolate)?
 yes
 d. Calculate the blood pressure of a 50 year old patient.
 $y = 0.96(50) + 81.05 = 111.05$

6) 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 below.

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

- a. What is the equation of the regression line?
 $y = 5.57x + 54.54$ $r = 0.92$
- b. Find the coefficient of determination (r^2) and interpret its meaning (reliability).
 $r^2 = 0.85$ 85% of change in test score can be explained by hours of study.
- c. Do you consider the equation to be reliable to predict further data (extrapolate)?
 yes

7) The table below compares rents for one-bedroom and two-bedroom apartments in 7 different cities.

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

- a. What is the equation of the regression line?
 $y = 1.54x + 98.53$ $r = 0.97$
- b. Find the coefficient of determination (r^2) and interpret its meaning (reliability).
 $r^2 = 0.95$ 95% of change in two bedroom rent can be explained by one bedroom rent.
- c. Do you consider the equation to be reliable to predict further data (extrapolate)?
 yes

- d. A one-bedroom rent in a downtown building in Lower Manhattan averaged \$3000 in August 2007. Calculate a two bedroom rent using the obtained formula.
 $y = 1.54(13000) + 98.53 = 84718.53$

8) Thanks to the progress of science the cancer survival rate is improving over time. The table shows 10-year survival rate for ovary cancer in England, diagnosed in year 1971, 1981, 1991 and 2001. According to the table 27% of the patients diagnosed in year 2001 survived over 10 years.

diagnosis year	71	81	91	101
percent lived over 10 years	18	22	25	27

- a. What is the equation of the regression line?
 $y = 0.3x - 2.8$ $r = 0.99$
- b. Find the coefficient of determination (r^2) and interpret its meaning (reliability).
 $r^2 = 0.98$ 98% of change in survival rate can be explained by diagnosis year.
- c. Do you consider the equation to be reliable to predict further data (extrapolate)?
 yes
- d. Estimate the 10-year ovary cancer survival rate for patients who are diagnosed in year 2015.
 $y = 0.3(115) - 2.8 = 31.7$ years