WESTMINSTER

INTERNATIONAL UNIVERSITY IN TASHKENT

An Accredited Institution of the University of Westminster (UK)

LECTURE 8
Correlation and Regression

Temur Makhkamov
Indira Khadjieva
QM Module Leaders
tmakhkamov@wiut.uz
i.khadjieva@wiut.uz

Office hours: by appointment
Room IB 205
EXT: 546

Lecture outline

- Define and calculate correlation coefficient
- Find the regression line and use it for regression analysis
- Define and calculate coefficient of determination (R-squared)

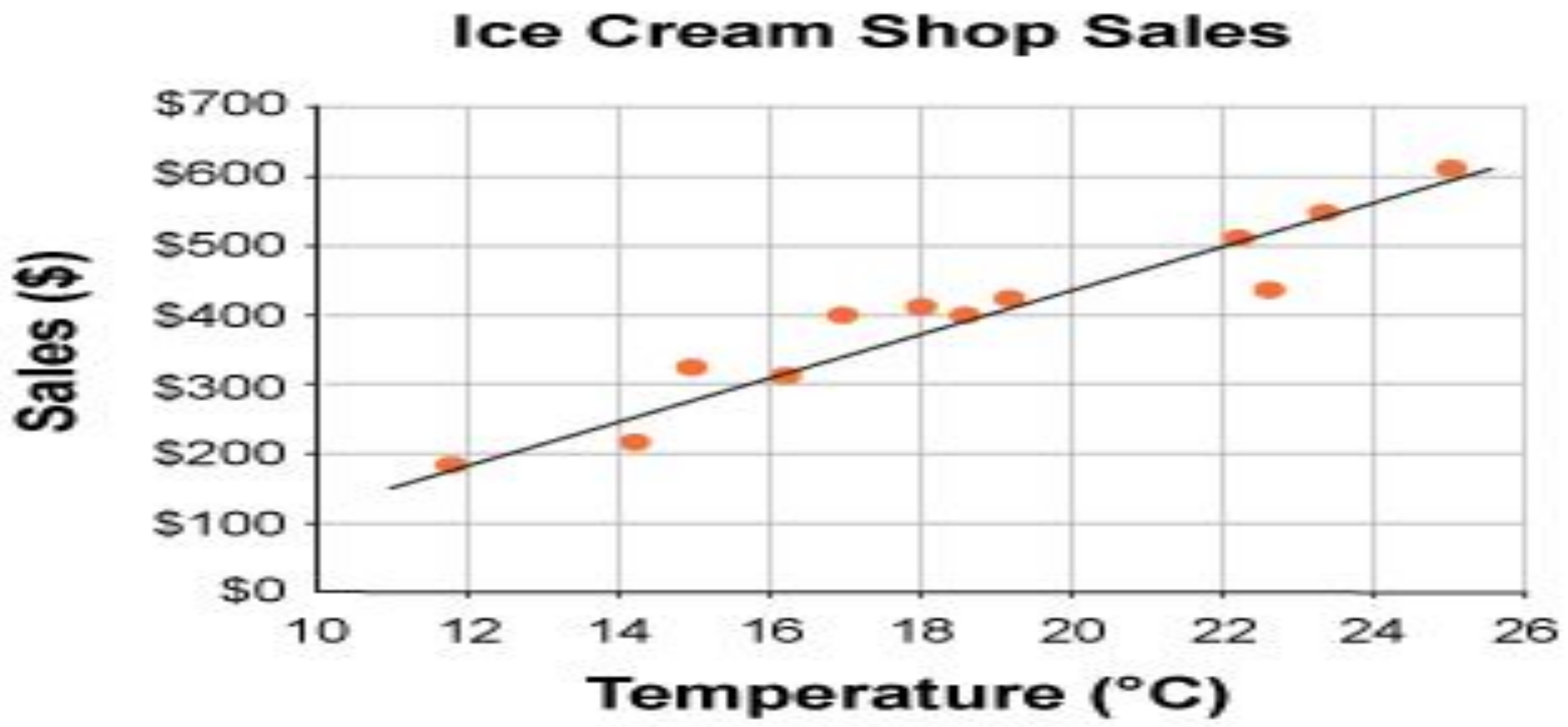
CORRELATION

- Correlation is a measure of the strength of a linear relationship between two quantitative variables

  SIMPLY, it's how two variables move in relation to one another.

- Measures the relationship, or association, between two variables by looking at how the variables change with respect to each other

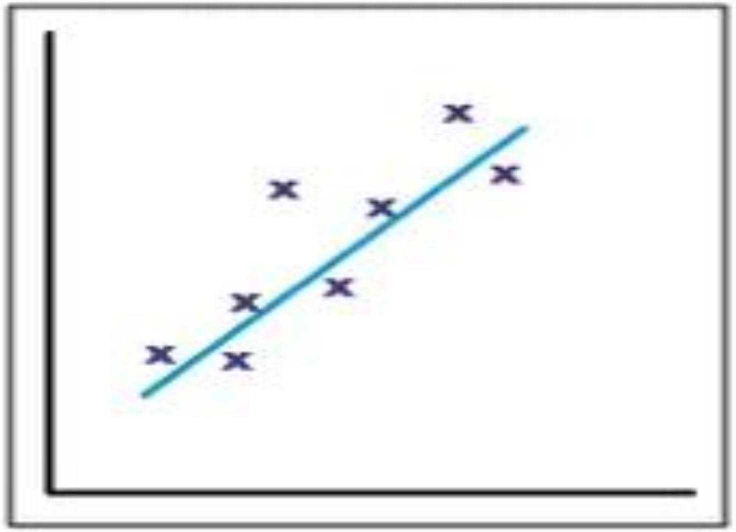
- The correlation coefficient is a value that indicates the strength of the relationship between variables. The coefficient can take any values from -1 to 1.


Ice Cream Shop Sales
Sales ($)
$700
$600
$500
$400
$300
$200
$100
$0
10
12
14
16
18
20
22
24
26
Temperature (°C)

Doing exersice & BMI (Body Mas Index)

TYPES OF CORRELATION

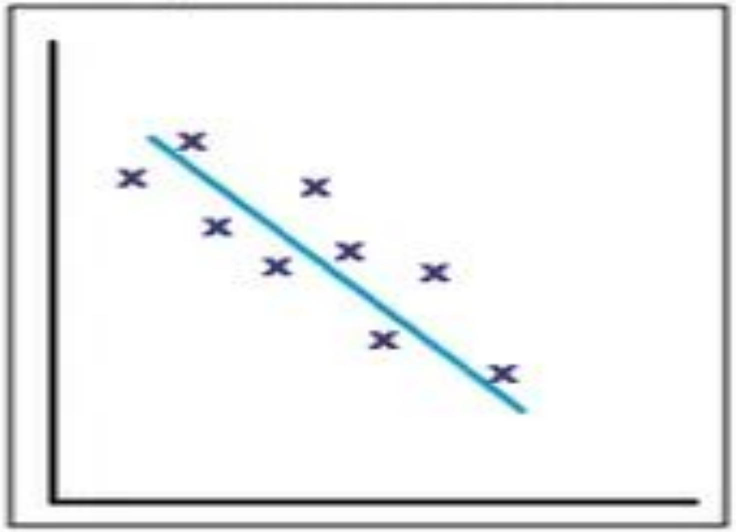
Positive correlation



The points lie close to a straight line, which has a positive gradient.

This shows that as one variable **increases** the other **increases**.

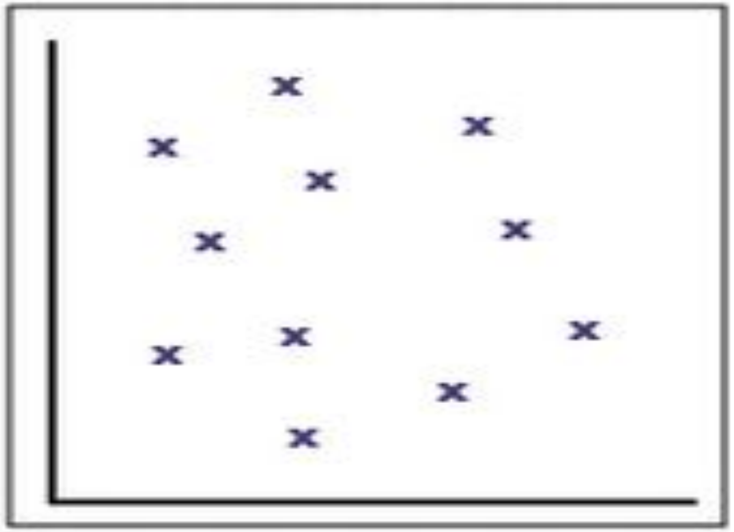
Negative correlation



The points lie close to a straight line, which has a negative gradient.

This shows that as one variable **increases**, the other **decreases**.

No correlation



There is no pattern to the points.

This shows that there is **no connection** between the two variables.

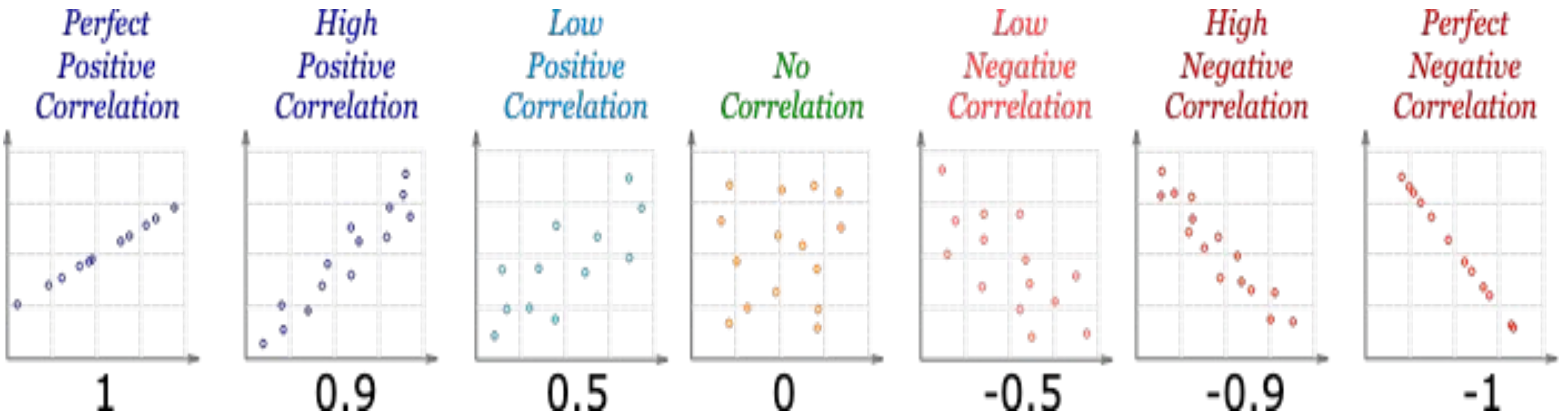
POSITIVE CORRELATION EXAMPLES

- ✔ As the number of trees cut down increases, the probability of erosion increases.
- ✔ As you eat more antioxidants, your immune system improves.
- ✔ The more time you spend running on a treadmill, the more calories you will burn.
- ✔ The longer your hair grows, the more shampoo you will need.
- ✔ The more money YOU save, the more financially secure YOU feel.
- ✔ As you drink more coffee, the number of hours you stay awake increases.
- ✔ As a child grows, so does his clothing size.
- ✔ The more you exercise your muscles, the stronger YOU get

Negative Correlation Examples

- ❖ A student who has many absences has a decrease in grades.
- ❖ If the sun shines more, a house with solar panels requires less use of other electricity.
- ❖ The older a man gets, the less hair that he has.
- ❖ The more one cleans the house, the less likely there are to be pest problems.
- ❖ The more one smokes cigarettes, the fewer years he will have to live.
- ❖ The more one runs, the less likely one is to have cardiovascular problems.
- ❖ The more vitamins one takes, the less likely one is to have a deficiency.
- ❖ The more iron an anemic person consumes, the less tired one may be.

CORRELATION COEFFICIENT



Measuring association between the variables

-

The covariance measures linear dependence between two variables.

Covariance:

$$Cov(x, y) = \sum \frac{(x - \bar{x})(y - \bar{y})}{n}$$

Cov > 0 indicates that two variables move in the same direction

Cov < 0 indicates that two variables move in opposite direction

CORRELATION COEFFICIENT

- The correlation coefficient that indicates the strength of the relationship between two variables can be found using the following formula:

To standardize the covariance we need to divide it by the product of two separate standard deviations.

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{\sum(x-\bar{x})(y-\bar{y})/n}{s_x s_y}$$

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$$

where:

- r_{xy} – the correlation coefficient of the linear relationship between the variables x and y
- x_i – the values of the x-variable in a sample
- $\bar{x}$ – the mean of the values of the x-variable
- y_i – the values of the y-variable in a sample
- $\bar{y}$ – the mean of the values of the y-variable

Finding Correlation

Jake is an investor. His portfolio primarily tracks the performance of the S&P 500 and he wants to add a stock of Apple Inc. Before adding Apple to his portfolio, he wants to assess the correlation between the stock and the S&P 500 to ensure that adding the stock won't increase the systematic risk of his portfolio.

	S&P 500	Apple
2017	2275	29,48
2018	2743	39,1
2019	2531	38,07
2020	2541	79,58
2021	3756	127,14

Finding Correlation

Using the formula below, Jake can determine the correlation between the prices of the S&P 500 Index and Apple Inc.

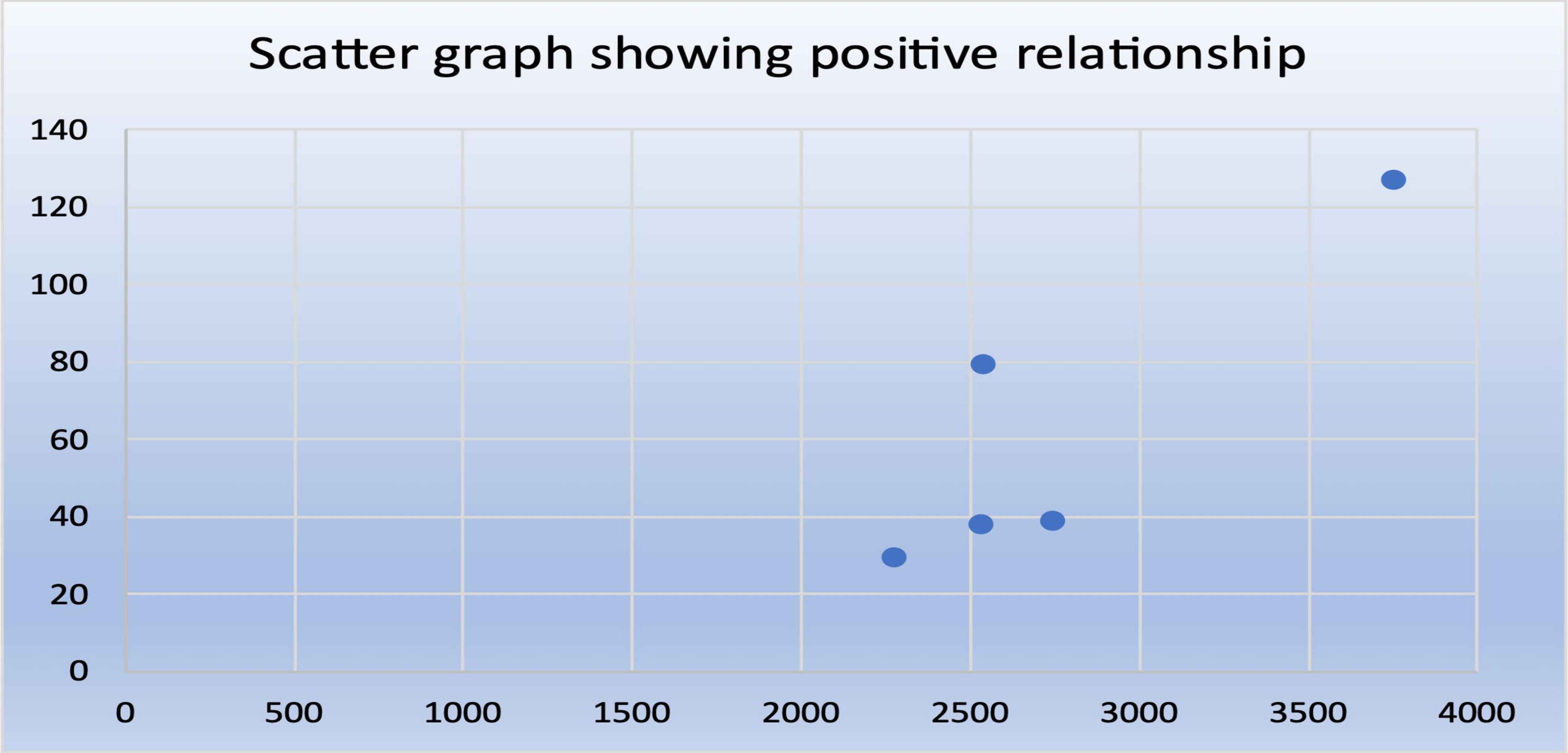
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{82639.886}{\sqrt{1327508.8 * 6704.6099}} = 0.876$$

The coefficient indicates that the prices of the S&P 500 and Apple Inc. have a high positive correlation. This means that their respective prices tend to move in the same direction. Therefore, adding Apple to his portfolio would, in fact, increase the level of systematic risk.

Calculation

	S&P 500 (x)	Apple (y)	x-xmean (a)	y-ymean (b)	a*b	(x-xmean)^2	(y-ymean)^2
2017	2275	29,48	-494,2	-33,194	16404,4748	244233,64	1101,84164
2018	2743	39,1	-26,2	-23,574	617,6388	686,44	555,733476
2019	2531	38,07	-238,2	-24,604	5860,6728	56739,24	605,356816
2020	2541	79,58	-228,2	16,906	-3857,9492	52075,24	285,812836
2021	3756	127,14	986,8	64,466	63615,0488	973774,24	4155,86516
Total	13846	313,37			82639,886	1327508,8	6704,60992

Mesuring association between variables



Strengths of Correlation

- Correlation allows the researcher to investigate naturally occurring variables that maybe unethical or impractical to test experimentally. For example, it would be unethical to conduct an experiment on whether smoking causes lung cancer.

- Correlation allows the researcher to clearly and easily see if there is a relationship between variables. This can then be displayed in a graphical form.

Limitations of Correlation

- Correlation is not and cannot be taken to imply causation. Even if there is a very strong association between two variables we cannot assume that one causes the other.

- Correlation does not allow us to go beyond the data that is given. For example, suppose it was found that there was an association between time spent on homework (1/2 hour to 3 hours) and Grade of student (30 to 40). It would not be legitimate to infer from this that spending 6 hours on homework would be likely to generate 80 marks.

Regression

If the relationship between variables exists (as we can see from correlation coefficient) we would be interested in predicting the behaviour of one variable, say y, from behaviour of the other, say x

Regression analysis is a well-known statistical learning technique useful to infer the relationship between a *dependent variable* ***Y*** and *independent variables.*

- ***predictor, explanatory*** or ***independent variable*** denoted ***x*** ;

- ***dependent variable, response, or outcome*** denoted by ***y*.**

Regression Analysis

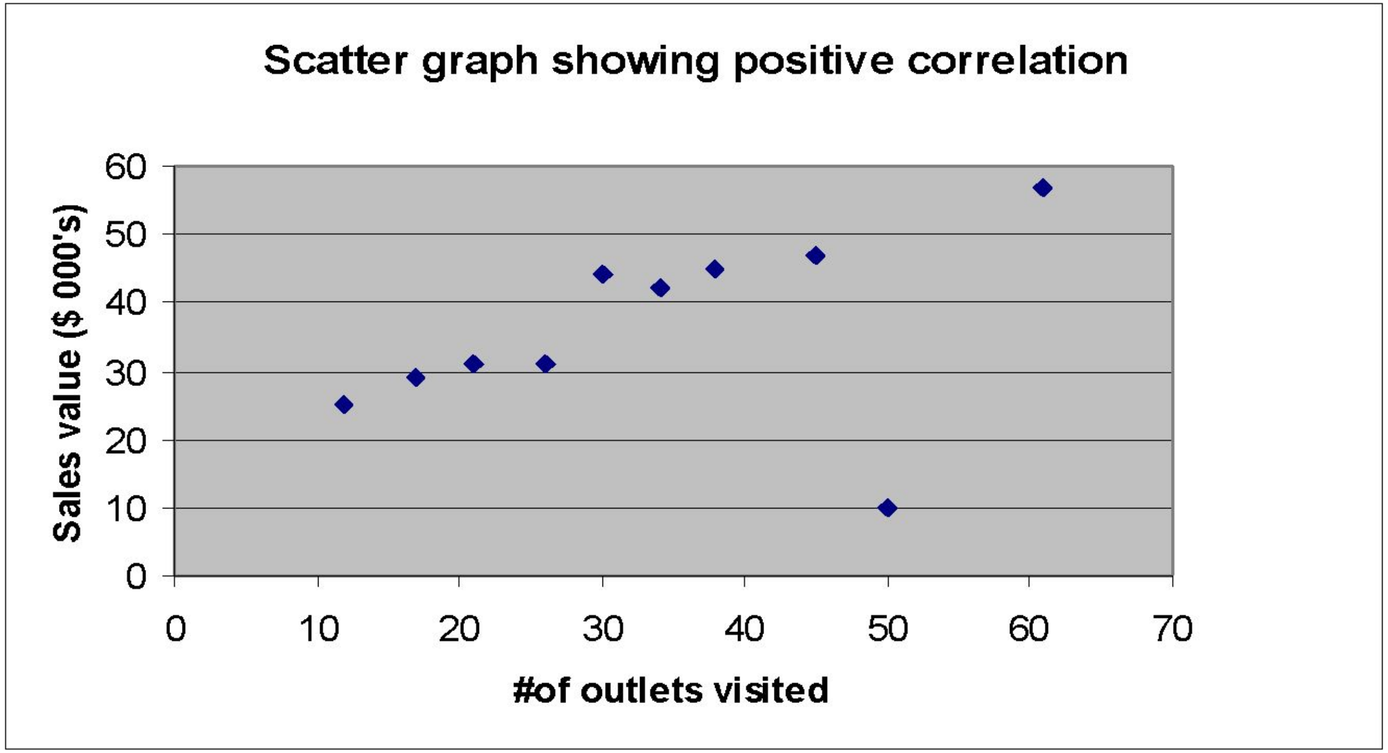
WESTMINSTER
INTERNATIONAL UNIVERSITY IN TASHKENT
An Accredited Institution of the University of Westminster (UK)

Rep. no.	Value of last quarter's sales ($000s)	Number of retail outlets visited regularly
1	10	50
2	25	12
3	29	17
4	31	21
5	31	26
6	42	34
7	44	30
8	45	38
9	47	45
10	57	61

Regression Analysis

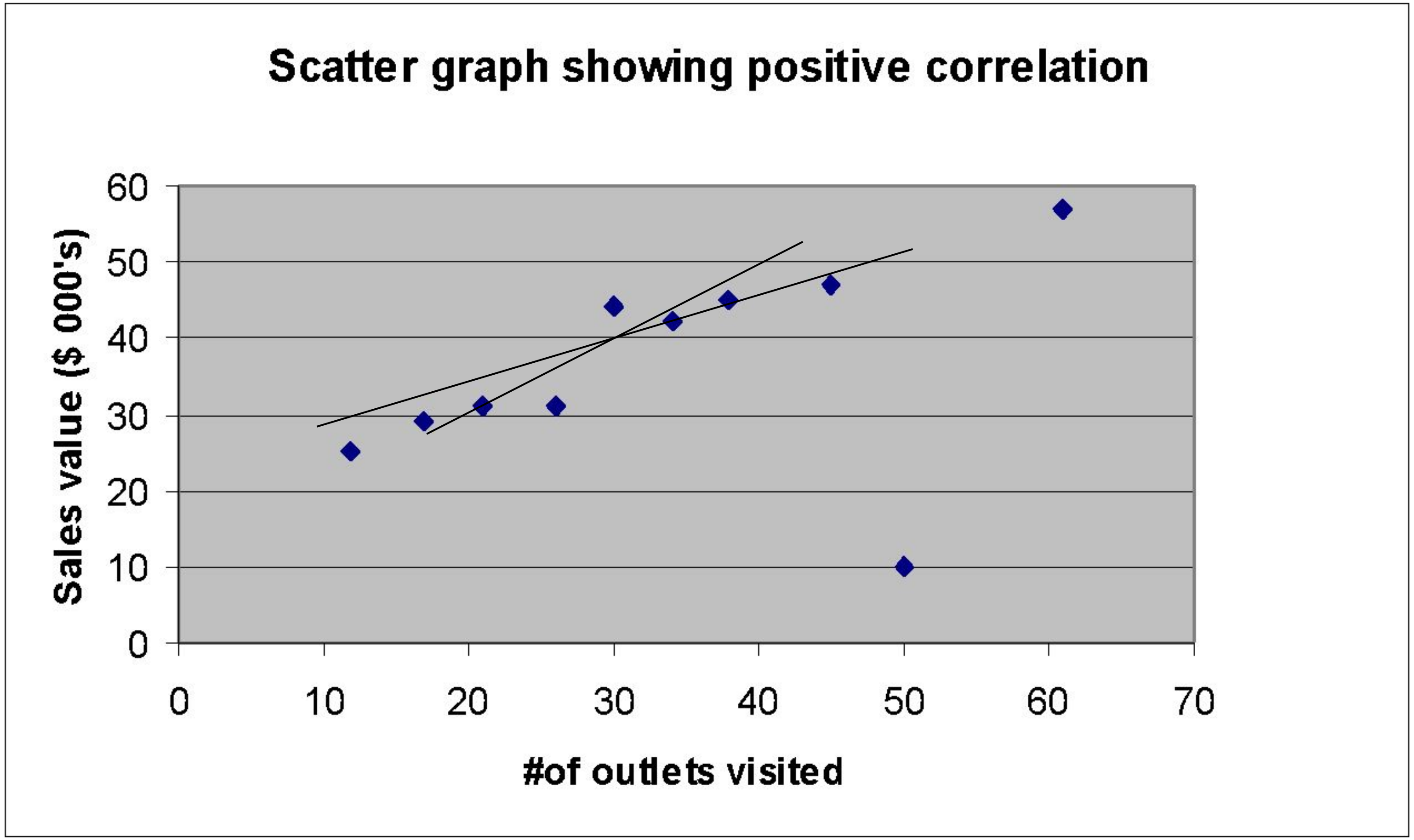
- Relationship between the sales and number of outlets visited could be well approximated by the line :
- **Sales=a+ b *number of outlets visited** (where a is a number of sales when no outlet is visited (x=0)

Or **y=a+bx**



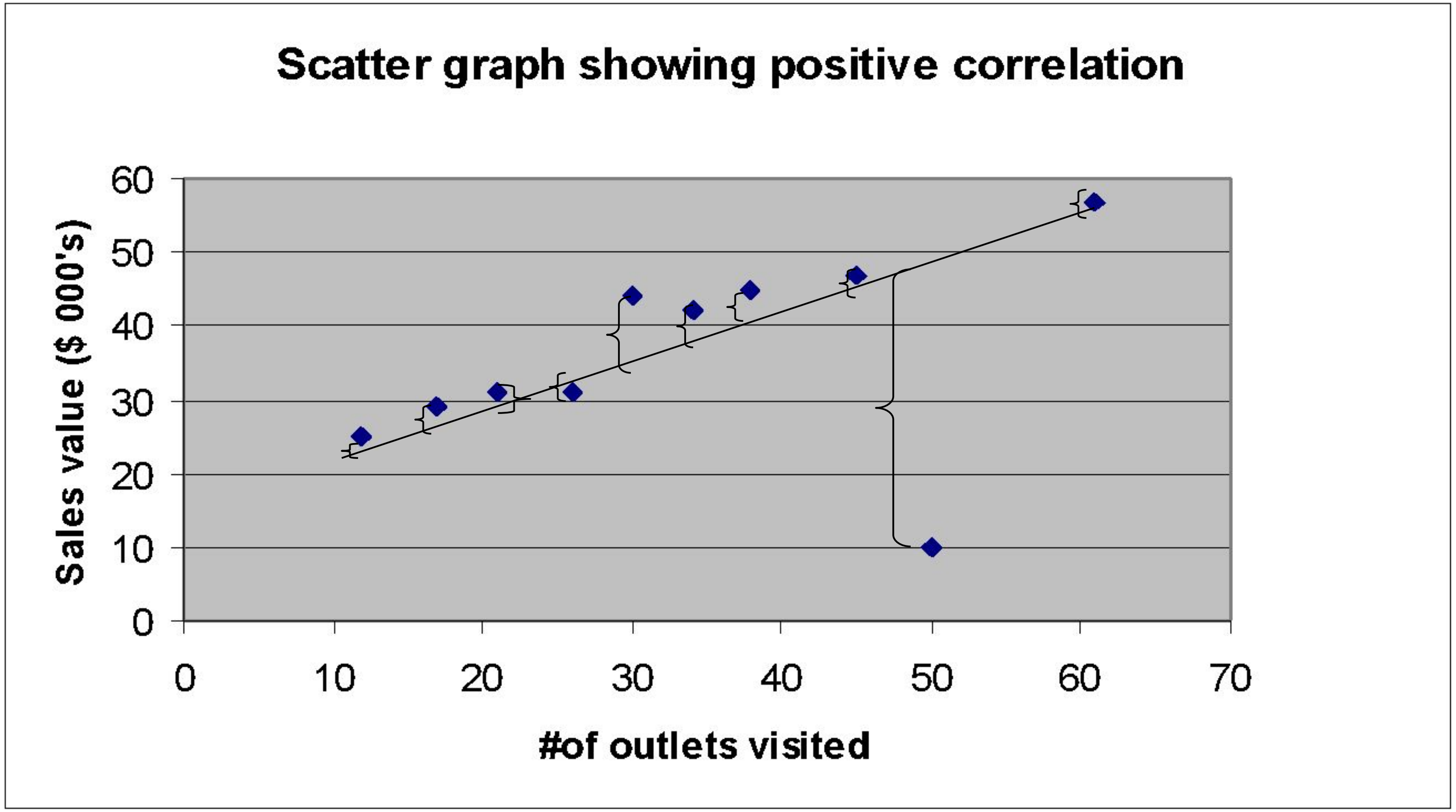
Regression Analysis

The problem is we could draw many possible lines. Which one to choose?



Regression Analysis

Well, try to find a line that minimizes the sum of squared distances between the data and the line (see the graph!) to ensure a better fit!



Regression Analysis

- For example, let's estimate the regression line for our data on sales minimizing the sum of **squared** differences between data and the line:

  Sales=a+ b *number of outlets visited

- Coefficient **b** of such line could be found using the following formula

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

- Coefficient **a** of such line could be found using the following formula $a = \bar{y} - b\bar{x}$

Regression Analysis

Rep. no.	Value of last quarter’s sales ($000s) (y)	Number of retail outlets visited regularly (x)	xy	x^2
1	10	50	500	2500
2	25	12	300	144
3	29	17	493	289
4	31	21	651	441
5	31	26	806	676
6	42	34	1428	1156
7	44	30	1320	900
8	45	38	1710	1444
9	47	45	2115	2025
10	57	61	3477	3721
Total	361	334	12800	13296

Regression Analysis

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10*12800-334*361}{10*13296-111556} = 0.3469$$

$$a = \bar{y} - b\bar{x} = 36.1 - 0.3469*33.4 = 24.512$$

Interpretation of Regression Analysis

Simple regression analysis

sales=24.5120+0.3469 x

Wow, we now could predict the sales by looking at number of outlet visited by sales representatives!

In our case, if we increase the number of outlets visited by sales representative by one the sales will increase by 0.3469 thousand dollars or 346.9 $

Regression Analysis (homework)

2nd method of finding coefficient of Regression Line

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

Mesuring quality of regression equation

- Coefficient of determination – R squared – is a statistical measure of how close the data are to the fitted regression line.

It takes values between 0 and 1, which is the same as 0% and 100 %, respectively.

$$R^2 = \rho^2(x,y) = 0.3965^2 = 0.1572$$

What does it imply?????

Essential readings

- Curwin J. and Slater R, Quantitative methods for Business Studies, 6th ed, Ch 15-17