Introduction to Statistical Modelling

STAT2507A
Chapter 3
Describing Bivariate Data

BIVARIATE DATA

- **Bivariate data** results when two variables are measured on a single experimental unit
  - Each variable can be described individually and **relationship** between the two variables can be explored
  - Bivariate data can be described with **graphs** and **numerical measures**

GRAPHS FOR QUALITATIVE VARIABLES

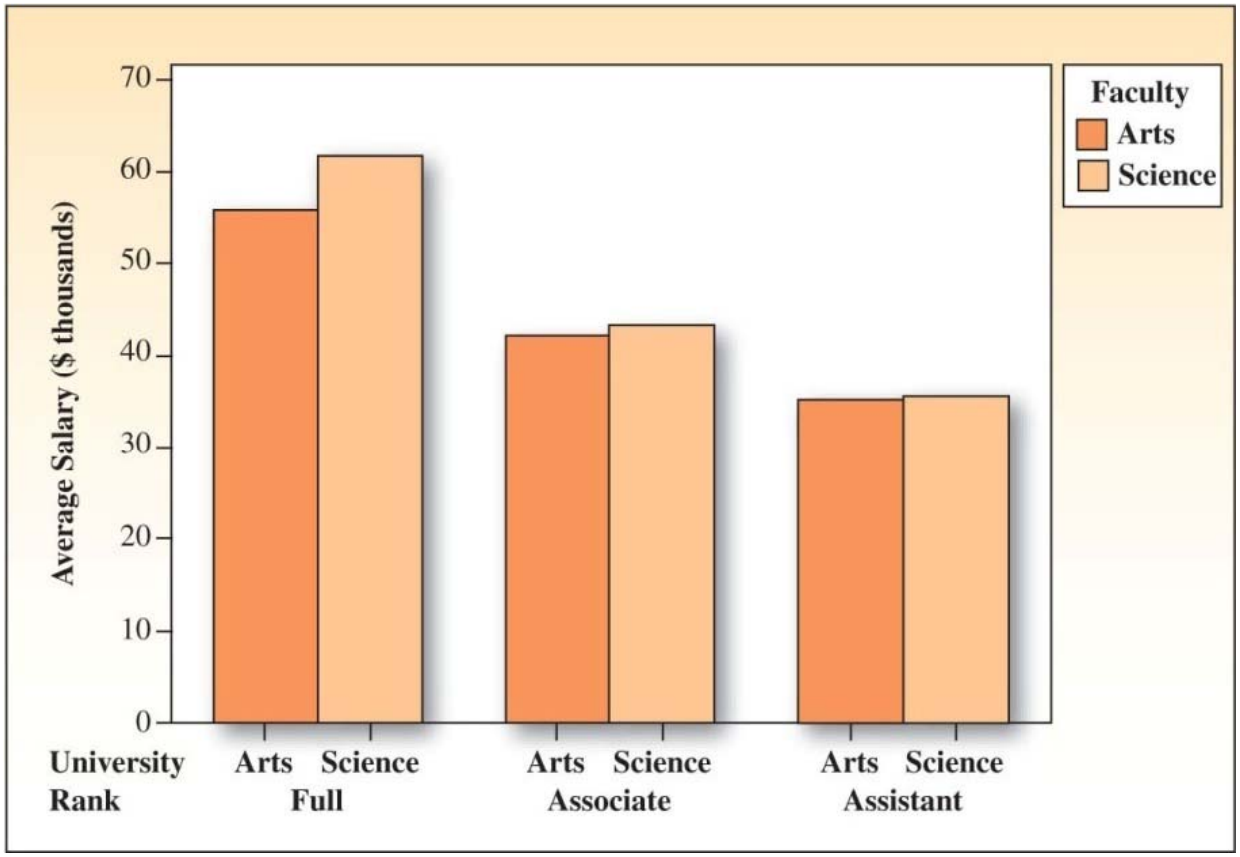
- When at least one of the variables is qualitative, **comparative pie charts** or (side-by-side or stacked) **bar charts** can be used

	Full Professor $ (count) (%)	Associate Professor $ (count) (%)	Assistant Professor $ (count) (%)	Total (count) (%)
Arts	55.8 (24) (16%)	42.2 (57) (38%)	35.2 (69) (46%)	(150) (100%)
Science	61.6 (60) (24%)	43.3 (78) (31.2%)	35.5 (112) (44.8%)	(250) (100%)
Total	(84)	(135)	(181)	(400)

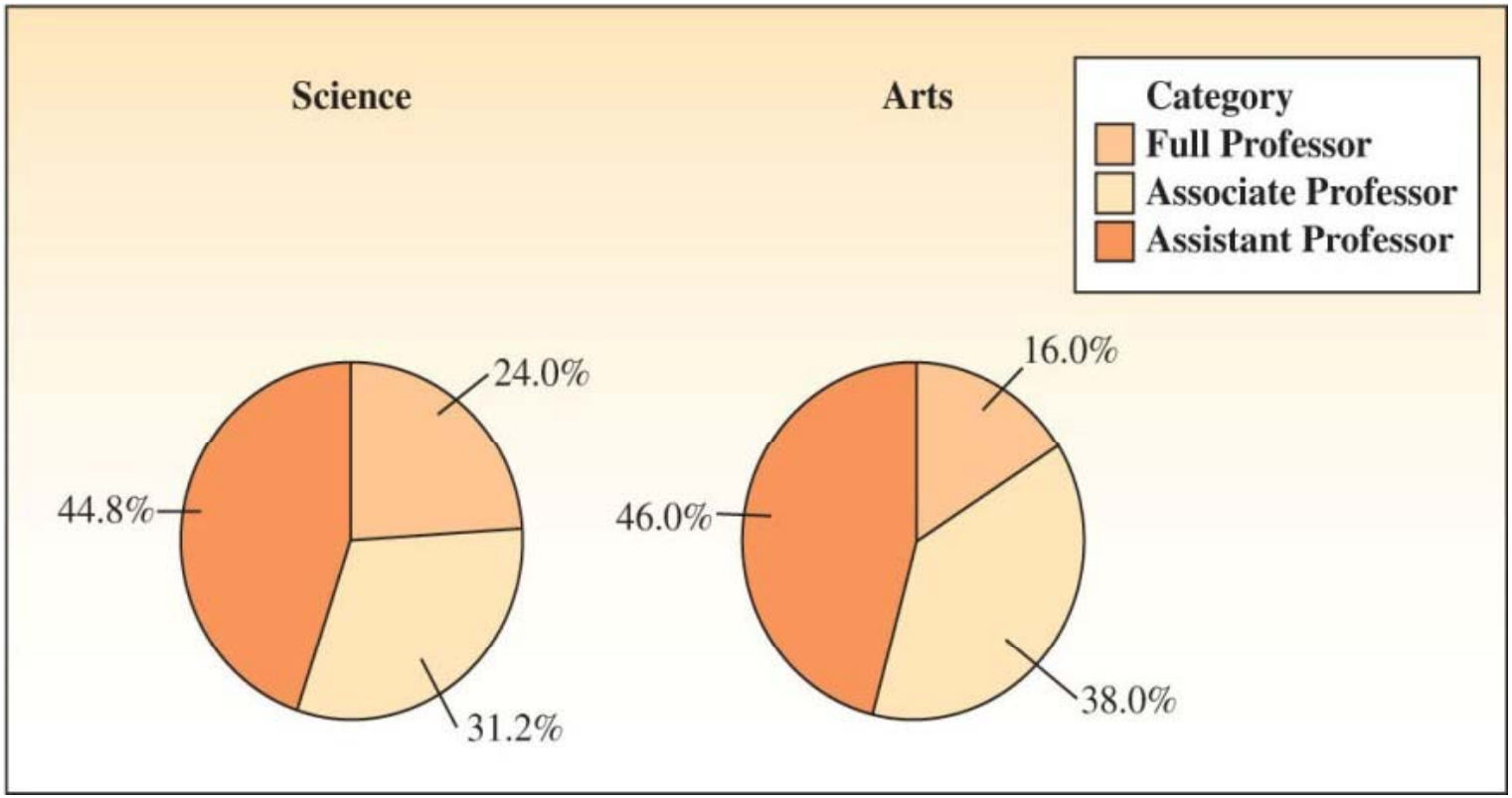
GRAPHS FOR QUALITATIVE VARIABLES

- Average salary of full, associate and assistant professors from Arts and Science to be compared. Side-by-side bar chart
- Are science professors paid more than professors in the faculty of Arts

SIDE-BY-SIDE BAR CHART

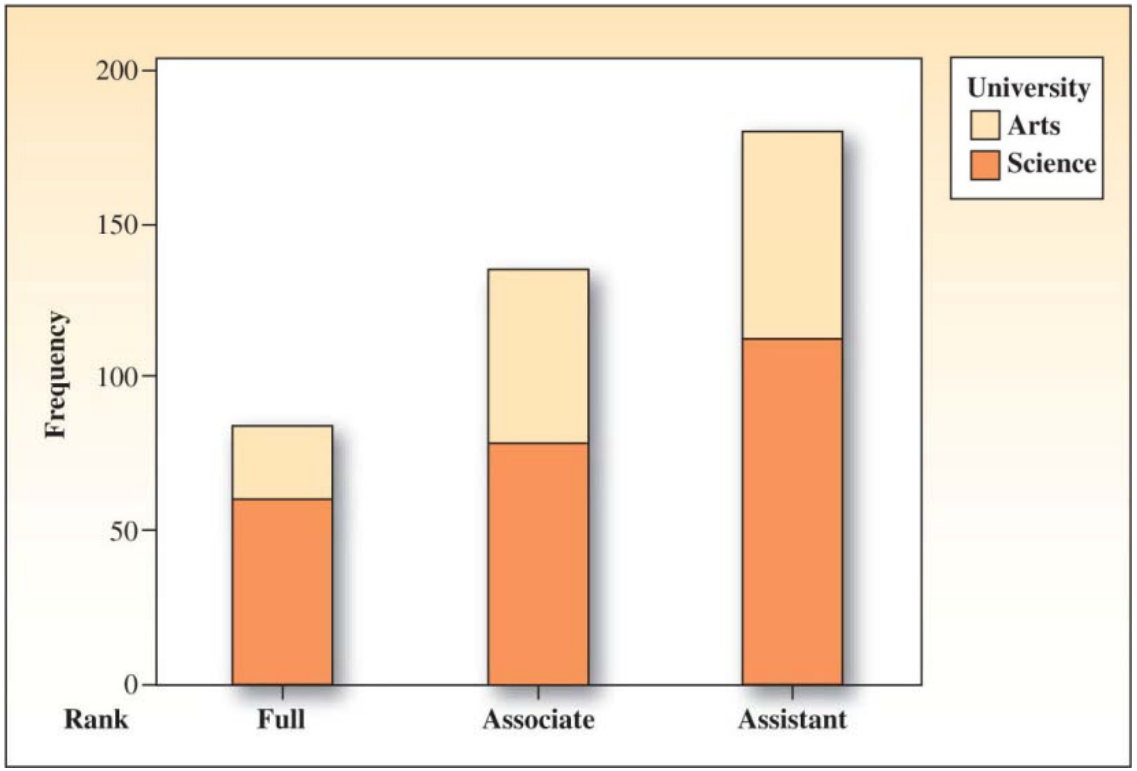


COMPARATIVE PIE CHART



GRAPHS FOR QUALITATIVE VARIABLES

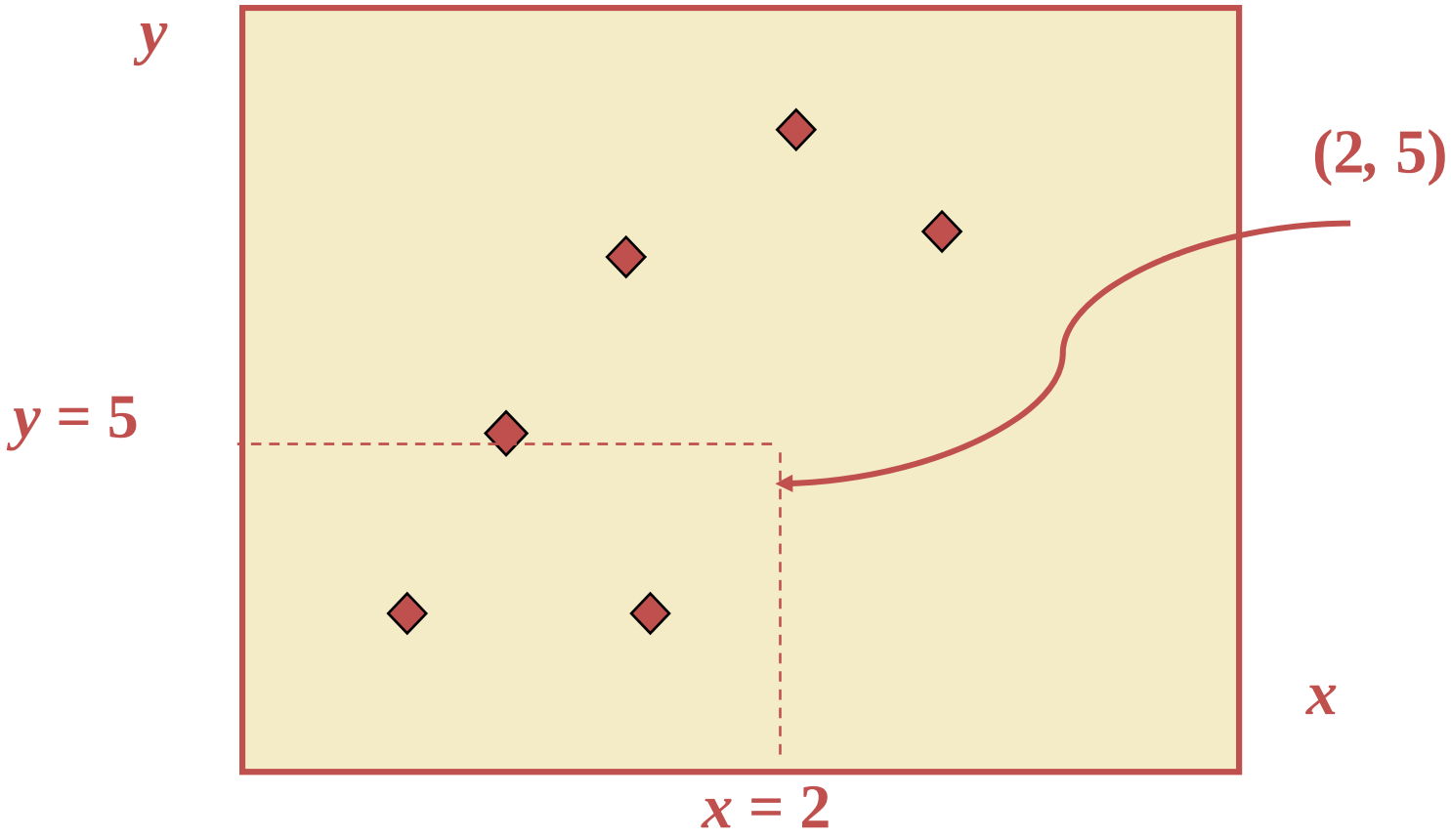
- Stacked bar chart: comparing the numbers of different ranking of professors from Arts and Science



TWO QUANTITATIVE VARIABLES

- When both of the variables are quantitative, one variable is called x and other y.
- A single measurement on a experimental unit is a pair of numbers (x, y) that can be plotted using a two-dimensional graph called a **scatter plot**.

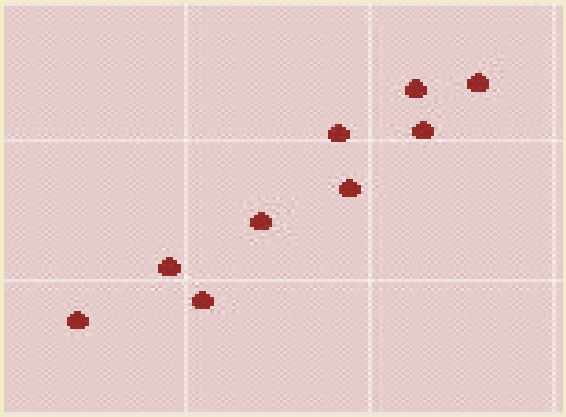
Two Quantitative Variables



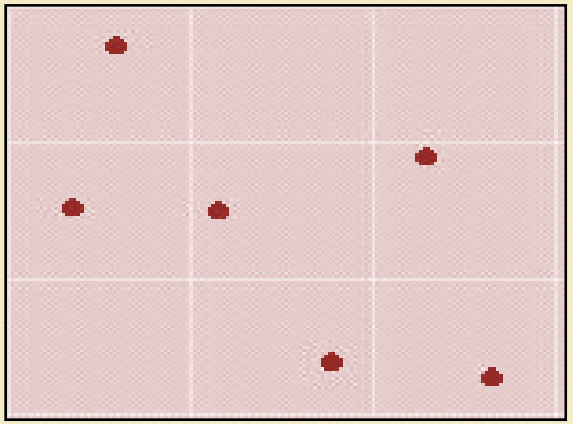
DESCRIBING THE RELATIONSHIP BETWEEN TWO VARIABLES

- What **pattern** or **form** do you see?
  - Straight line upward or downward
  - Curve or no pattern at all
- How **strong** is the pattern?
  - Strong, moderate, or weak
- Are there any **unusual observations**?
  - Clusters or outliers

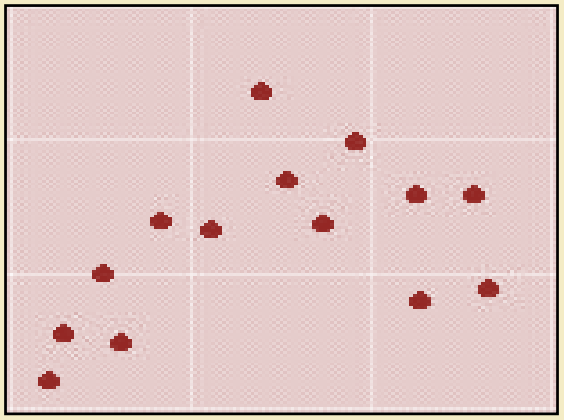
SCATTER PLOT EXAMPLES



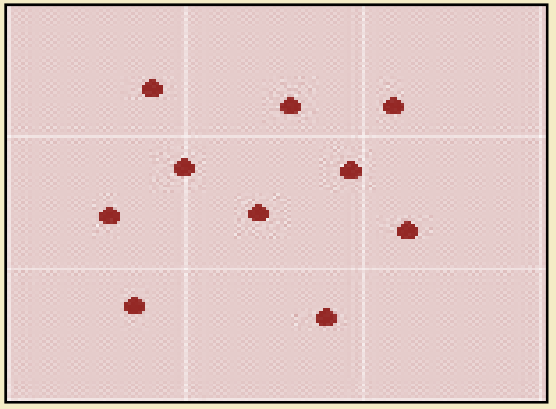
Positive linear - strong



Negative linear - weak



Curvilinear



No relationship

NUMERICAL MEASURES: QUANTITATIVE BIVARIATE DATA

- Assume that the two variables *x* and *y* exhibit a **linear pattern** or **form**
- Two numerical measures:
  - The **form** of the relationship:
    linear, curvilinear, clustered, randomly scattered.
  - The **strength** and **direction** of the relationship between x and y.

NUMERICAL MEASURES: QUANTITATIVE BIVARIATE DATA

- Assume that two variables x and y exhibit a linear pattern or form
  - **Correlation coefficient, r:** used to measure the strength and direction of the relationship between x and y

$$r = \frac{S_{xy}}{S_x S_y}$$

  - Where S_{xy} – covariance between x and y

FORMULAE

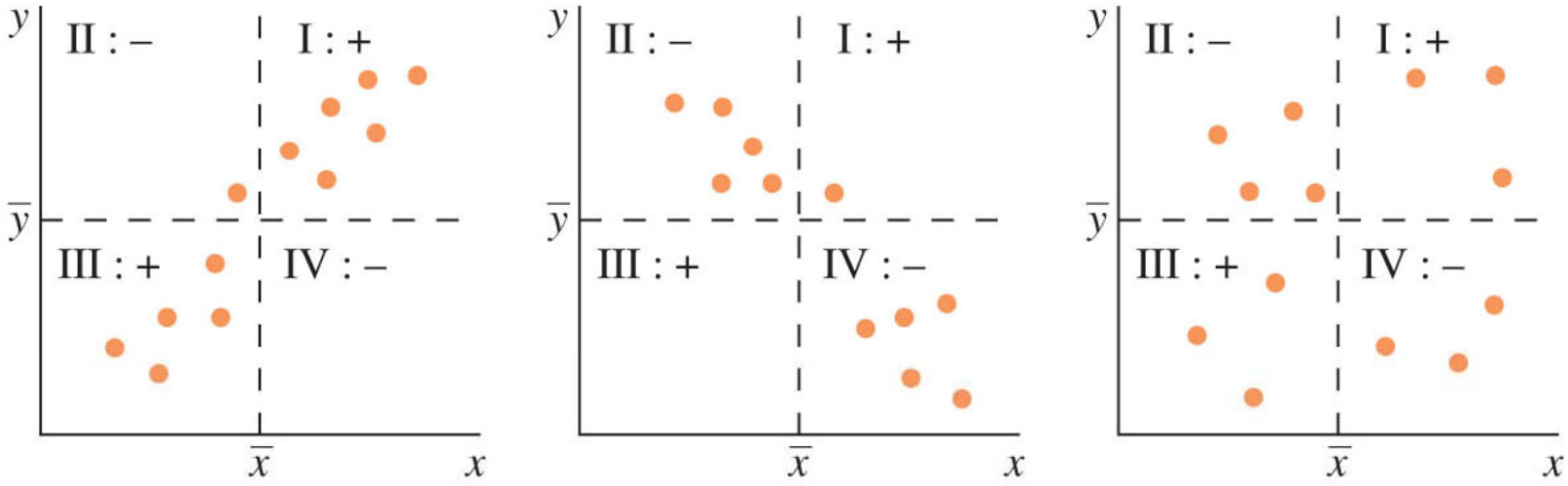
$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{n-1}$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

INTERPRETING R

$-1 \leq r \leq +1$	Sign of r indicates direction of the linear relationship
$r \approx 0$	Weak relationship. Random scatter of points
$r \approx 1$ or -1	Strong relationship; positive or negative
$r = 1$ or -1	All points fall exactly on a straight line

INTERPRETING R



(a) Positive pattern (b) Negative pattern (c) No pattern

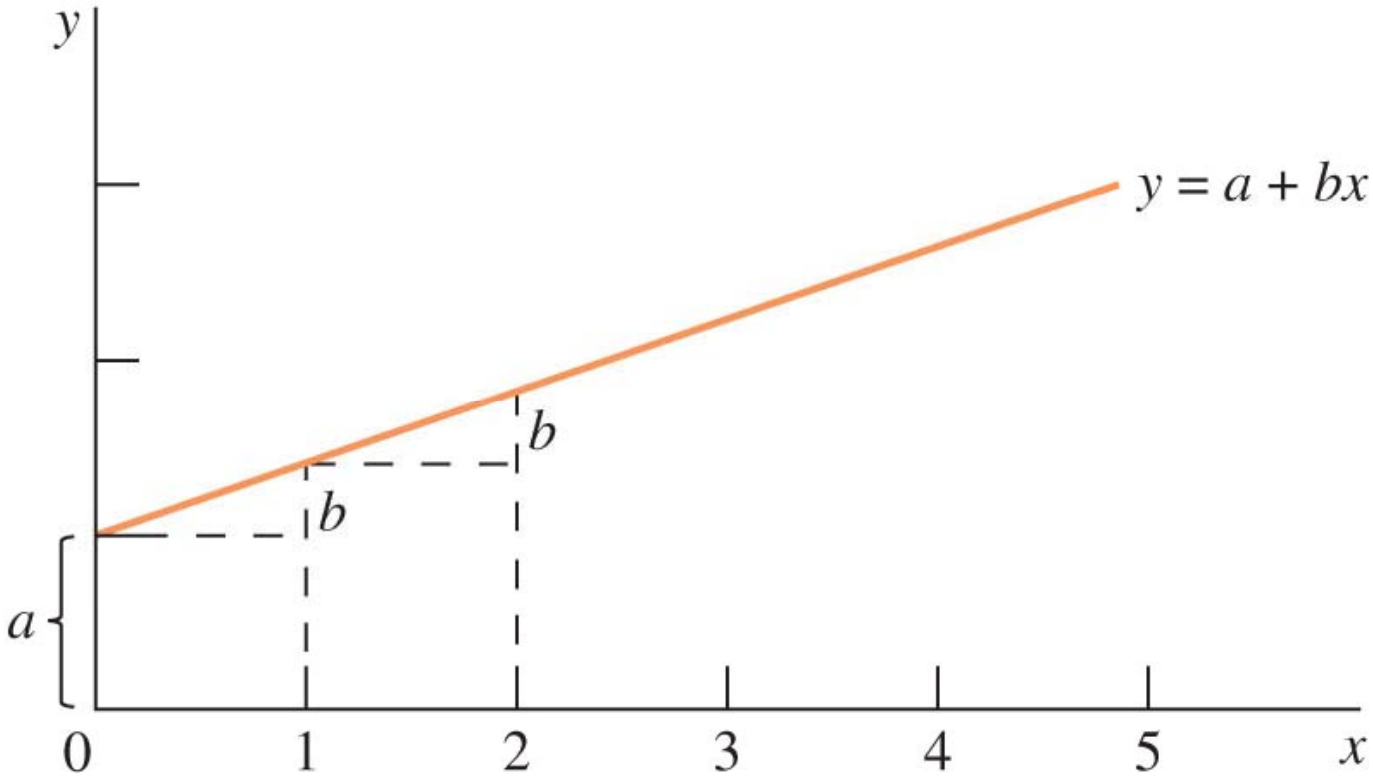
THE REGRESSION LINE

- Sometimes the two variables x and y are related in a particular way. i.e. The value of y depends on the value of x. Then
  - Y – dependent variable
  - X – independent variable

THE REGRESSION LINE

- Linear relationship between x and y can be described by fitting a line as best through the points
- Linear regression line: $y = a + bx$
  - a = y –intercept of the line
  - b = slope of the line

THE REGRESSION LINE



The Regression Line: Formula

- To find the slope and y-intercept of the best fitting line, use

- Slope: $$b = r\frac{S_y}{S_x}$$

- Y-intercept: $$a = \bar{y} - b\bar{x}$$

The Regression Line: Formula and Example

- Example: Living area x and selling price y of five homes

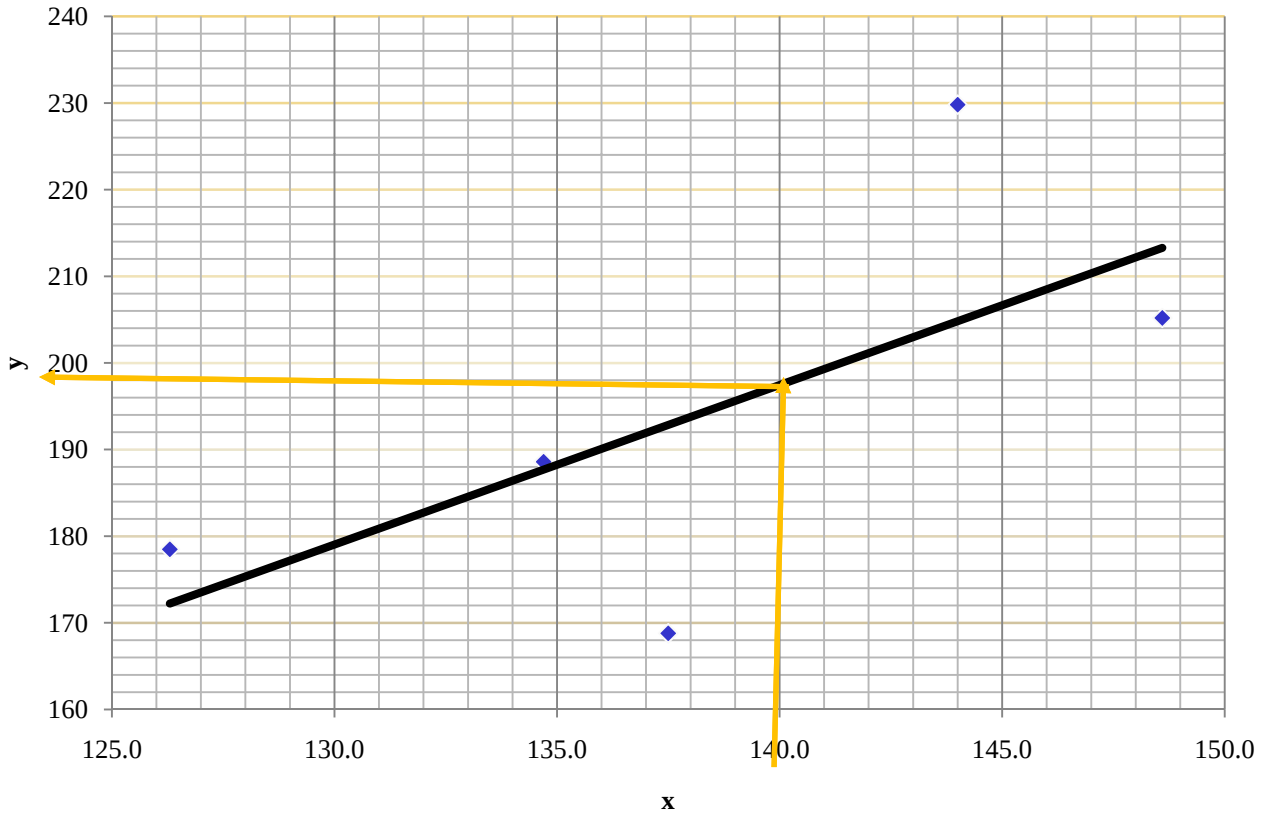
Residence	1	2	3	4	5	Sum
X (Area in m^2)	126.3	134.7	137.5	144.0	148.6	691.1
Y (Price in $ 1000s)	178.5	188.6	168.8	229.8	205.2	970.9
XY	22544.55	25404.42	23210.00	33091.20	30492.72	134742.89

EXAMPLE

- Calculate r, b, a
- Obtain fitted Regression line
- Predict the selling price for another residence with 140 m^2 of living area

PLOT

- The least-squares regression line



SUMMARY

- Describing two qualitative variables
  - Side-by-side pie charts
  - Comparative bar charts
    - Side-by-side
    - Stacked

SUMMARY

- Describing two Quantitative variables
  - Scatter plots
    - Linear or non-linear
    - Strength of relationship
    - Unusual observation: outliers
- Covariance and Correlation coefficient

SUMMARY

- The best-fitting regression line
  - Calculating slope and intercept
  - Graphing the line
  - Using the line for prediction