Linear regression for numerical bivariate data (Part I)

BEA140 Quantitative Methods - Module 2

Example of numerical bivariate data

As an example, we will use the following sample of numerical bivariate data consisting of 8 days of data for the temperature and ice cream sales:

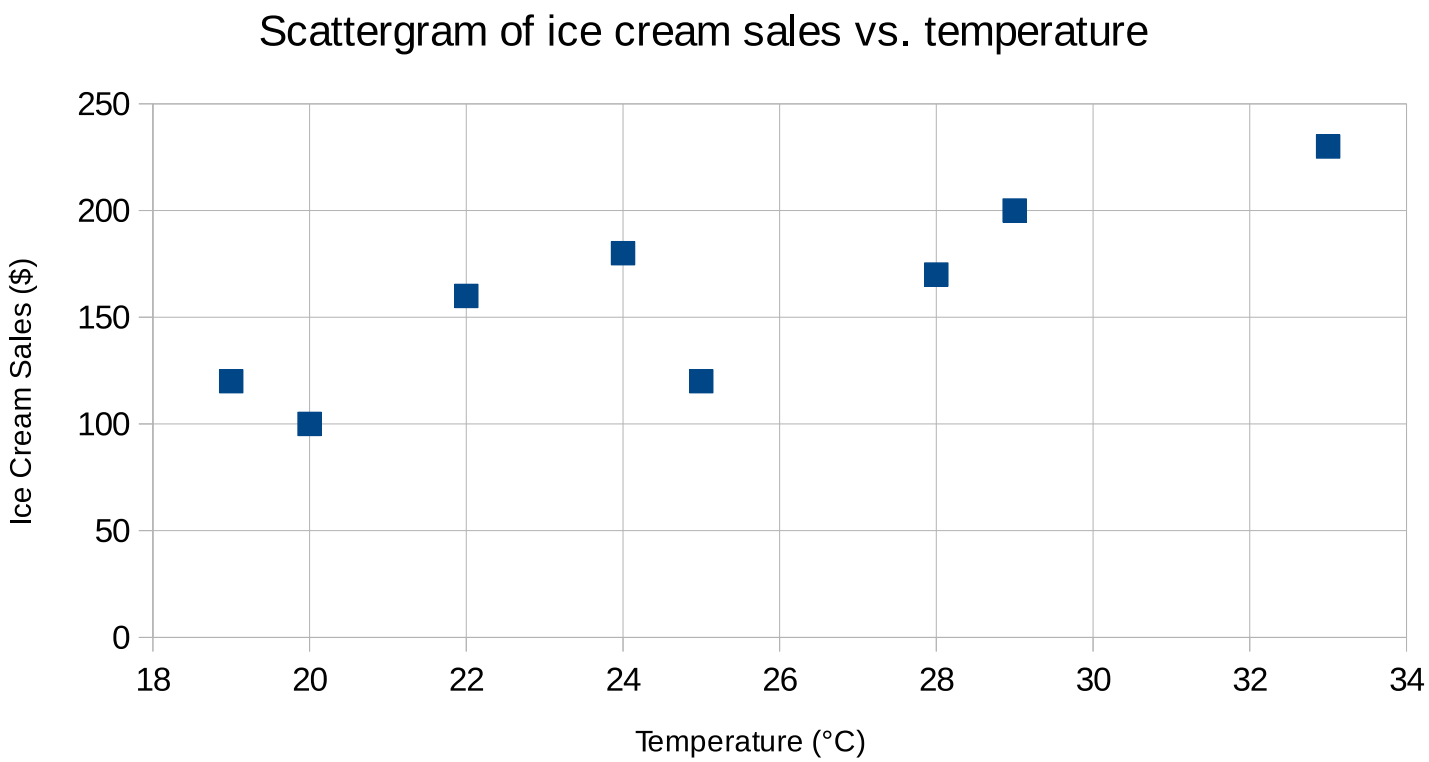
temp (°C)	ice cream sales ($)
29	200
22	160
28	170
19	120
25	120
24	180
20	100
33	230

Scattergrams

A **scattergram** for bivariate data is a 2d plot of the data with:

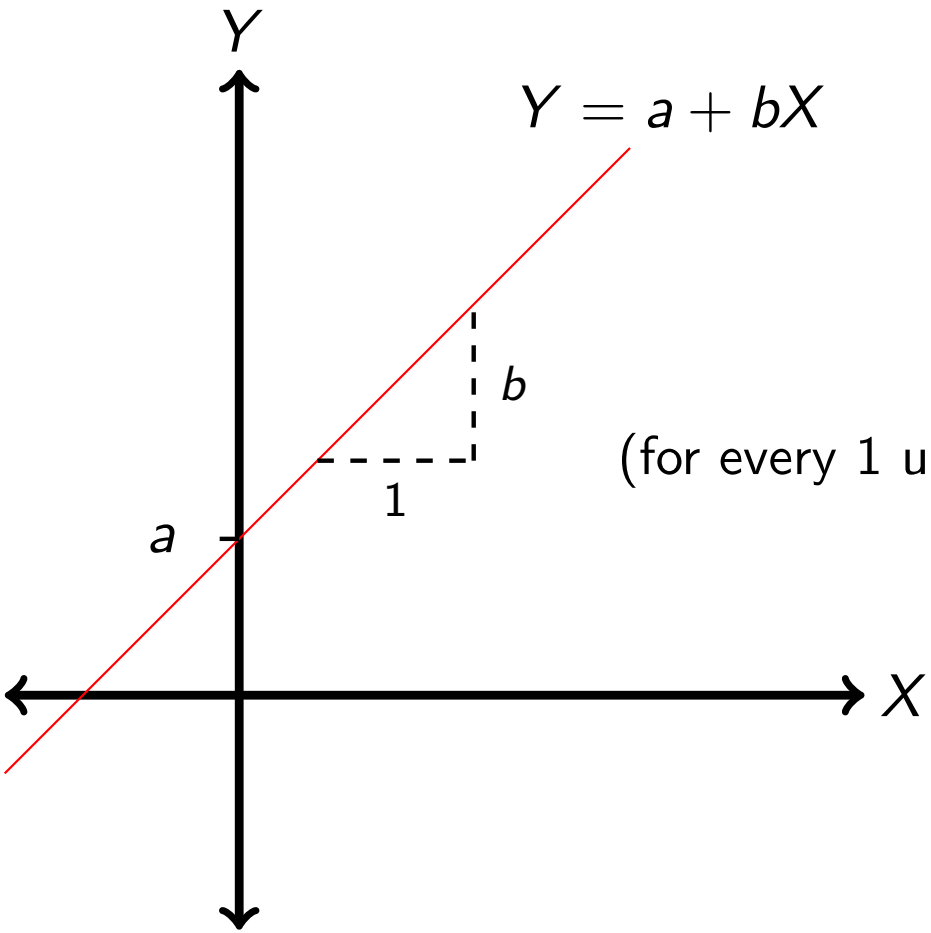
(i) one variable along the horizontal axis (which is often referred to as the independent variable); and

(ii) the other variable along the vertical axis (which is often referred to as the dependent variable).



Linear functions

At some point during your education you may have come across linear/straight lines of the form $Y = a + bX$.



$a = Y-$intercept
(where $X = 0$)

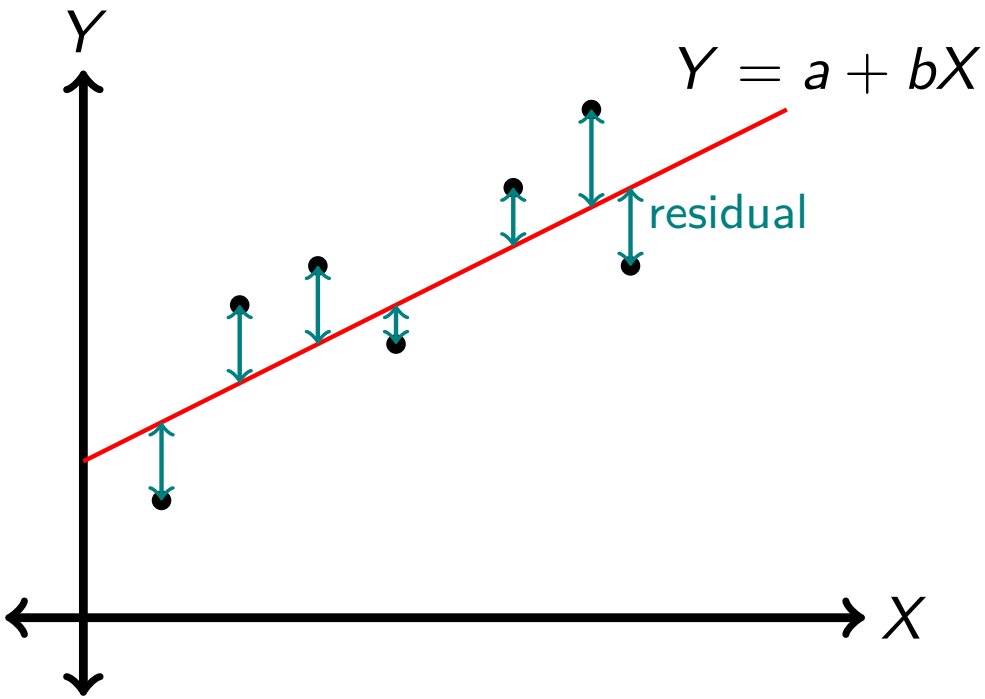
$b =$ slope
(for every 1 unit increase in X, Y increases by b)

Linear regression - Residuals

Given:

(i) some data points; and

(ii) a line $Y = a + bX$.



A **residual** is defined as the vertical distance from *an actual/observed value* and *the predicted/calculated value* (point on the line of best fit).

Linear regression - *Best Fit* Property

One way to think of how well a line fits some data points is to sum the squares of residuals, if the sum of one line is smaller than a second line then we say that the first line *fits the data better*.

It turns out that for a given data set, a unique line that *best fits* the data can be identified (i.e. a line for which the sum of squared residuals is lower than every other line). This line is referred to as **the ordinary least squares linear regression**.

Linear regression - Formulas

Formulas for the slope and intercept of the least squares linear regression line are:

$$\text{slope} = b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}; \text{ and}$$

$$\text{intercept} = a = \frac{\Sigma Y - b\Sigma X}{n}.$$

Note: It can be proved mathematically using calculus that the above formulas for a and b minimise the sum of squares of residuals (i.e. *best fit* the data), though the mathematically less-inclined need not fear as that is beyond the scope of this unit.

Linear regression - Example

Going back to our temperature and ice cream sale data:

temp (X)	ice cream sales (Y)	X^2	Y^2	XY
29	200	841	40000	5800
22	160	484	25600	3520
28	170	784	28900	4760
19	120	361	14400	2280
25	120	625	14400	3000
24	180	576	32400	4320
20	100	400	10000	2000
33	230	1089	52900	7590
200	1280	5160	218600	33270

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = 7.9375 \text{ and } a = \frac{\Sigma Y - b\Sigma X}{n} = -38.4375.$$

Hence the linear regression line for the data set is
$Y = -38.4375 + 7.9375X$.

Linear regression - Correlation does not imply causation

It is **extremely** important to note that causation can **not** be concluded from regression results, regression results establish a correlation (not causation).

Example: Countries with a higher per capita rate of heroin usage also tend to have a higher life expectancy.

A regression on such data establishes that there is a positive correlation between the two variables, however it does not establish that one variable influences the other, it could be that causation happens in neither direction but is instead influenced by other factors.

For instance, average wealth is much more likely to be the underlying cause for higher heroin consumption and life expectancy, though a regression is unable to prove causation here as well!

Linear regression - Properties

It can be mathematically proven that the following properties hold for the regression line of a data set:

(i) the regression line predicts the mean of Y at the mean of X, i.e. it passes through the point $(\overline{X}, \overline{Y})$;

(ii) the sum of residuals/errors is zero, i.e. sum of vertical distances between the data points and the regression line; and

(iii) the sum of residuals/errors squared is minimised (this is the best fit property from earlier on).

Linear regression - Interpolation & Extrapolation

Predicting the value of Y for a given value of X:

(i) within the range of data for X is referred to as **interpolating within the range of data**; and

(ii) outside the range of data for X is referred to as **extrapolating beyond the range of data**.

For example if predicting the sales of ice creams using our regression line calculated earlier:

- predictions made for temperatures between 19 and 33 degrees would be interpolating; and
- predictions made for temperatures less than 19 or greater than 33 would be extrapolating.

Note: When making predictions using regression results, interpolating within the observed range of X is much more valid than extrapolating outside the observed range of X (primarily because there is no data/evidence that the linear relationship continues outside the observed range of X).

. . . that's it for now, thanks for watching!

Don't forget that you can ask questions via:

(i) face-to-face lectures;

(ii) workshops or tutorials;

(iii) consultation hours; or

(iv) email.