Reference card

Identifying patterns

Line of best fit

A **straight line** drawn through a scatter plot. This line helps us to **visualise the relationship between two variables** by representing the **trend** in a dataset.

Independent variable

Plotted on the **x-axis**. The factor that is being **changed or manipulated**, i.e. the variable **causing the effect**.

Dependent variable

Plotted on the **y-axis**. The factor that is being **measured or observed** as a result of the changes made to the independent variable, i.e. the variable **we want to study**.

Equation of the line

A **straight line equation** that is used to **make predictions** and **gain insights**.

```
y = mx + c
```

y = dependent variable
m = slope: Indicates the rate of change in y per unit change in x.
x = independent variable
c = y-intercept: The point at which the line intersects with the y-axis.

Linear regression

The process of fitting a straight line to a set of data points to describe the relationship. The line of best fit is the visual representation of this straight line.

Evaluating the line

We can use various tools and metrics to **analyse** the line of best fit and **evaluate** its ability to represent the data.

Equation of the line

The line can be evaluated by **analysing the variables** in the equation.

slope (`m`)

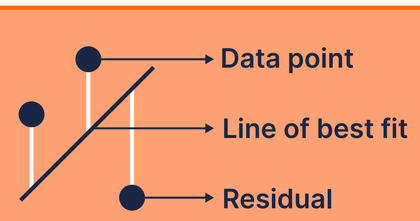
- **Sign**: Positive indicates a positive relationship; negative indicates a negative relationship.
- **Value:** A higher value indicates a steeper slope, i.e. a change in x is associated with a larger change in y.

y-intercept (`c`)

- The predicted value of y when x is equal to zero, i.e. it is the value of y when x does not have any effect or influence.

Residual

The distance between each data point and the line. The line of best fit minimises the sum of the squared residuals for all the data points.



R-squared

Measures the goodness of fit of a line. Tells us how much of the variation in the dependent variable is explained by the independent variable.

```
=RSQ(data_y, data_x)
```

$$R^2 = 1 - \frac{RSS}{TSS}$$

RSS = sum of squares of residuals
TSS = total sum of squares

Correlation

Measures the **strength and direction** of a linear **relationship** between two variables.

```
=CORREL(data_y, data_x)
```

Heatmap

Visually represents the relationship between multiple variables by using colour to indicate the value of a cell.

Linear relationship: The heatmap shows a gradual colour change.

Non-linear relationship: The heatmap shows a random distribution of colours.

EXPLORE AI ACADEMY