Aston Business School

Correlation and Regression analysis

BN2255 – Business Analytics in Practice

Association between variables

- Question: Does TV advertising lead to increased sales?

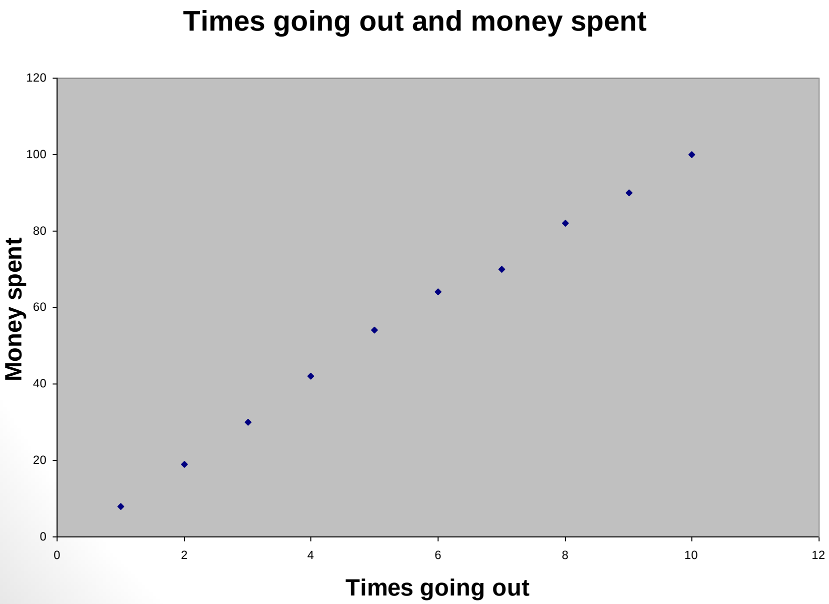
- Are the two variables related?
  - number of ads played and sales
- What kind of a relation?
  - are increasing number of ads played associated with increased sales?
- Is the relation strong or weak?
  - how important is the number of ads played in ‘explaining’ changes in sales?
- Can we predict the values of one variable by using the values of the other?
  - can we predict the amount of sales by the number of ads played?

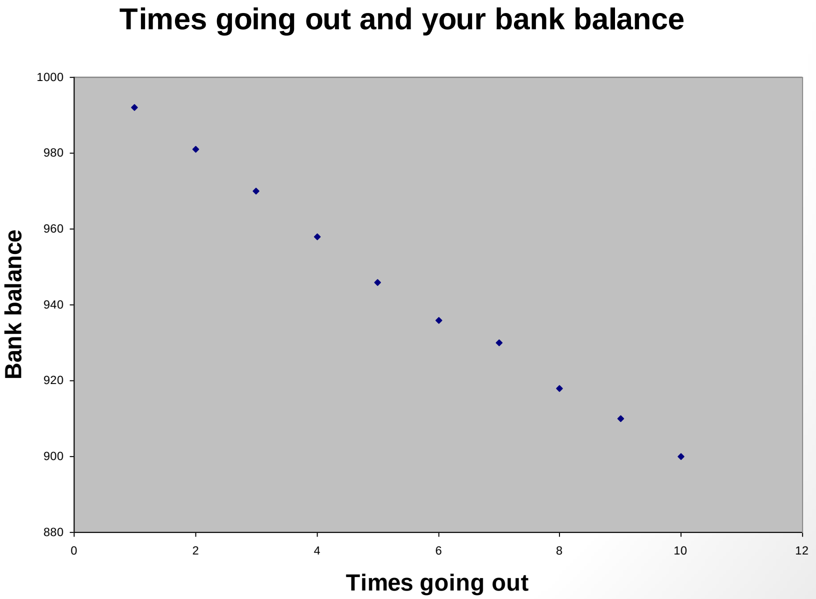
Are two variables related? Graphical approach

- A visual inspection can be helpful in identifying the presence of a relationship between two variables.
- A diagram that can give an idea about the relationship between two variables is the scatter diagram.
- One variable is plotted along the x axis and the other on the y axis. Each point is identified by a coordinate (x, y), e.g. (4, £5k).
- From the graph we will be able to tell whether a relationship exists or not and also the type of the relationship
  - eg. linear or non-linear, positive or negative

Scatter diagrams (1)

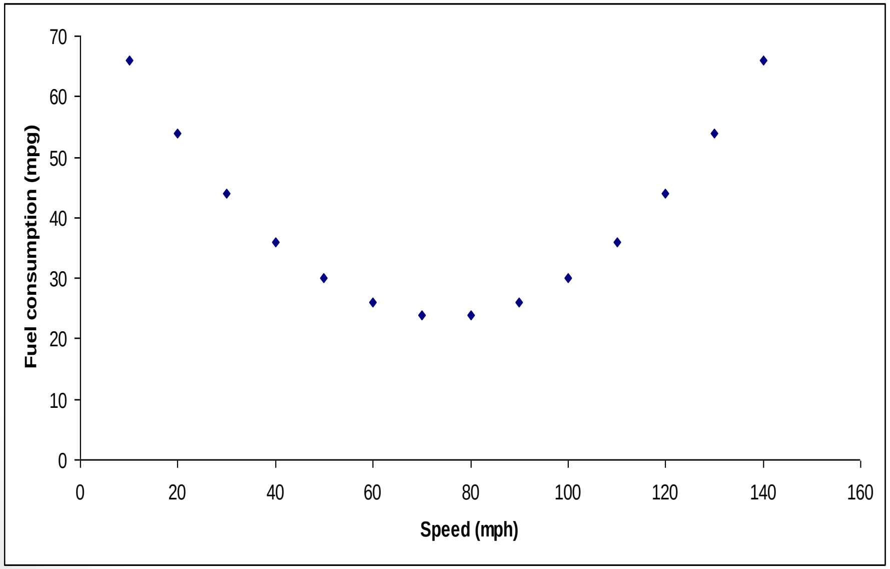
- A linear positive relationship can be observed when if one of the variables increases, the other one increases as well
  - eg. the number of times you go out every week, and the amount of money that you spend.
- A linear negative relationship can be observed when one of the variables increases, the other one decreases
  - eg. the number of times you go out and your bank balance

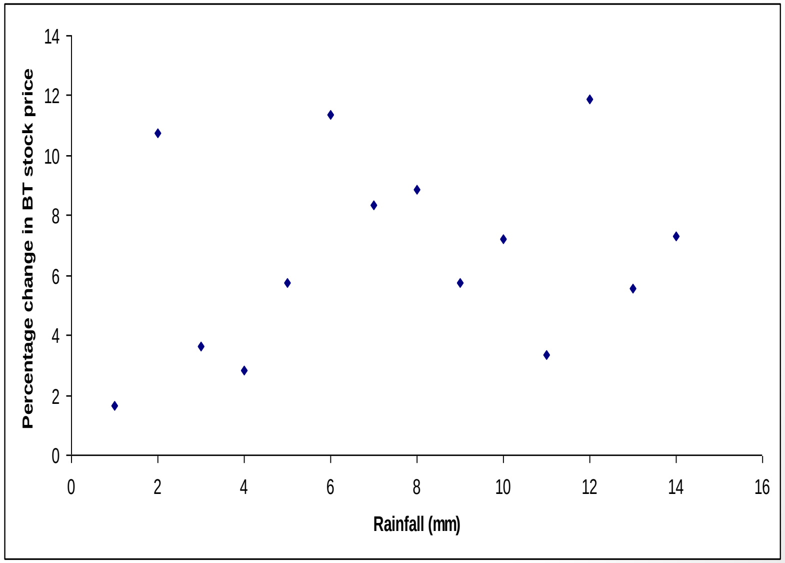




Scatter diagrams (2)

- sometimes the relationship between variables is not linear
  - eg. driving speed and fuel consumption
- scatter diagrams also easily demonstrate cases where there appears to be no relationship
  - eg. changes in stock market prices and rainfall





Strength of association - Pearson's Correlation Coefficient (r)

- A way to measure the strength and the direction (positive/negative) of the linear association between two variables is the Pearson's correlation coefficient (r)
  - takes values between +1 and -1

- Correlation is a measure of how much two variables change together; it is a measure of the association between the two variables.

If r = {
+1, then we have a perfect positive linear relationship
0, then there is no linear relationship
-1, then we have a perfect negative linear relationship

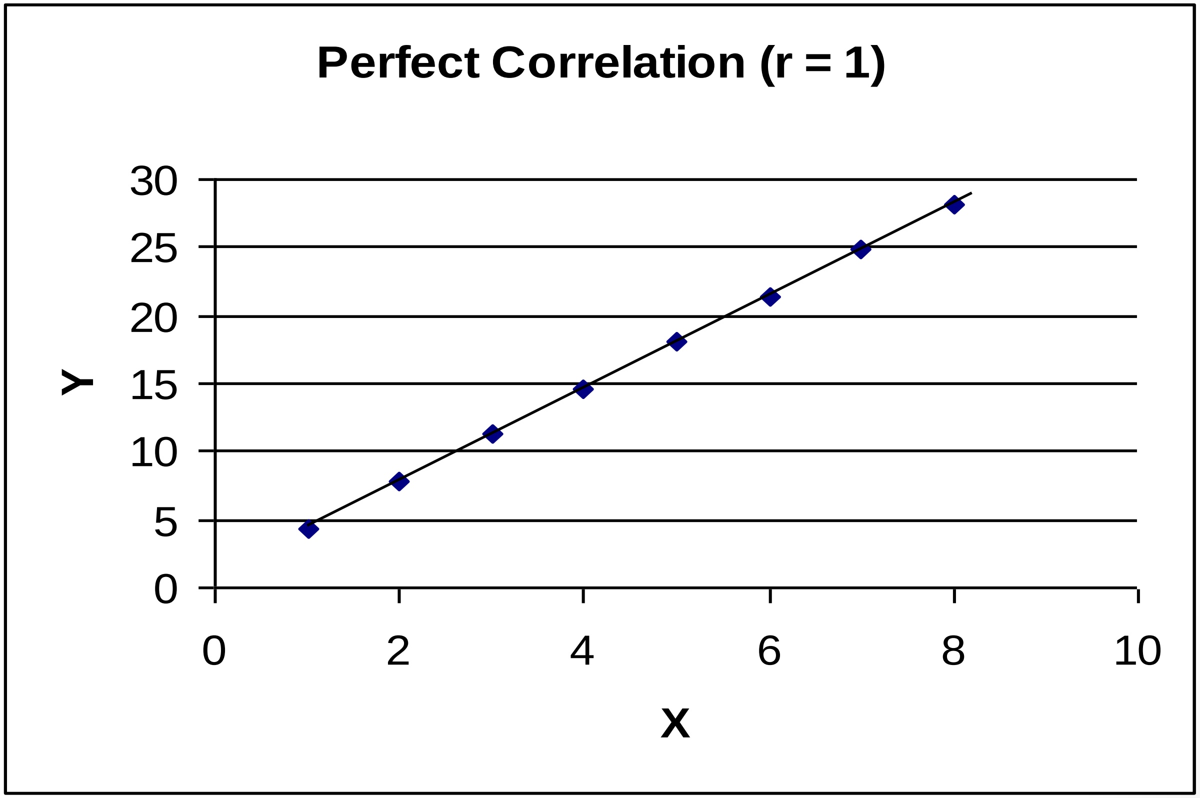
- Excel can readily calculates this using the CORREL() function

Correlation values

- If all the points in a data set fall on a straight line then we have perfect correlation
  - If the line is has a positive slope, then we have perfect positive correlation and $r = 1$
  - If the line is has a negative slope, then we have perfect negative correlation and $r = -1$
- If we get a value of **r** that is close to either 1 or -1 then we can say that a strong linear (either positive or negative) association exists between the two variables.
- A value of **r** close to zero (0) means that the linear association is very weak or that there is no association at all.

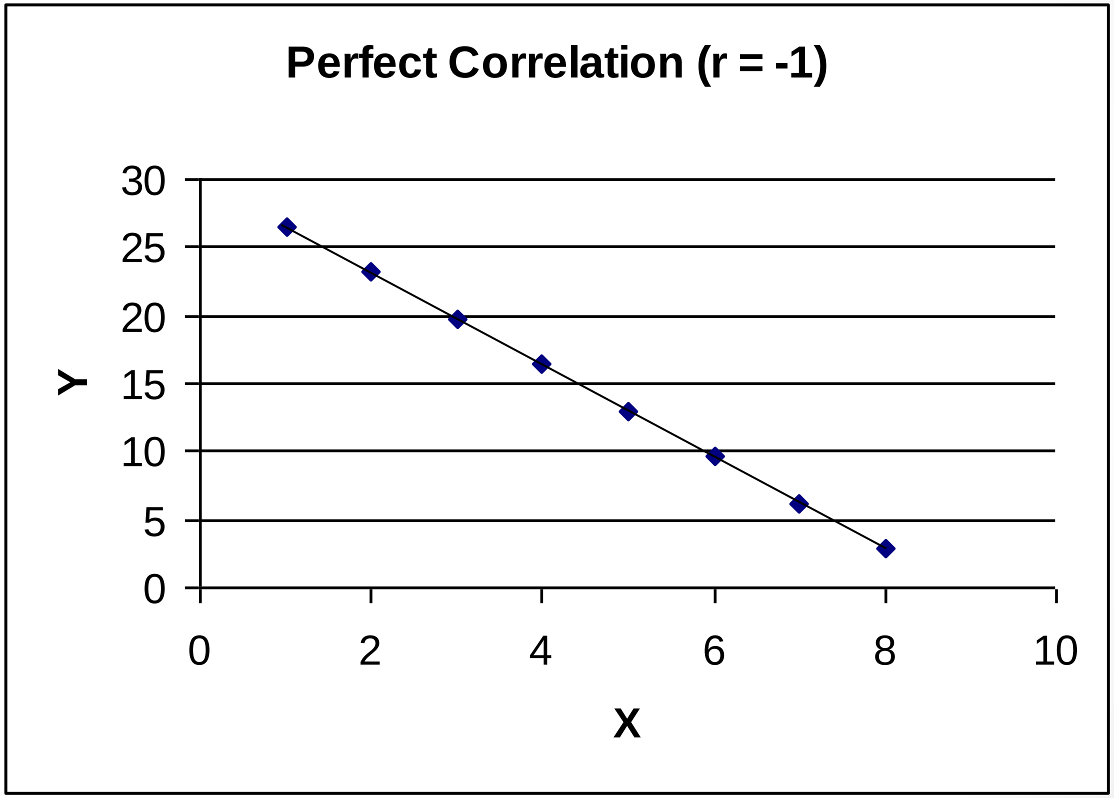
Perfect positive correlation

If **r = +1,** then an increase in one of the variables is exactly matched by a linear increase in the other variable.



Perfect negative correlation

If **r = -1,** then an increase in one of the variables is exactly matched by a linear decrease in the other variable.



Indicative characterisations of correlation strength

- The strength of correlation is subjective and depends on a number of factors
  - sample size, results of other statistical tests and measures but most importantly the variables we are considering
- Table below represents a ‘rule of thump’ for most business applications

Value of the Correlation Coefficient	Strength of Correlation
1	Perfect
0.7 - 0.9	Strong
0.4 - 0.6	Moderate
0.1 - 0.3	Weak
0	Zero

Correlation vs Causality

- Warning! ***Correlations does not always imply causality***
  - due to the presence of spurious correlations it is easy to mix causality with correlation

- Spurious Correlation: Two variables may have no direct causal connection but may be wrongly inferred that they have due to the presence of another variable that causes both to happen.

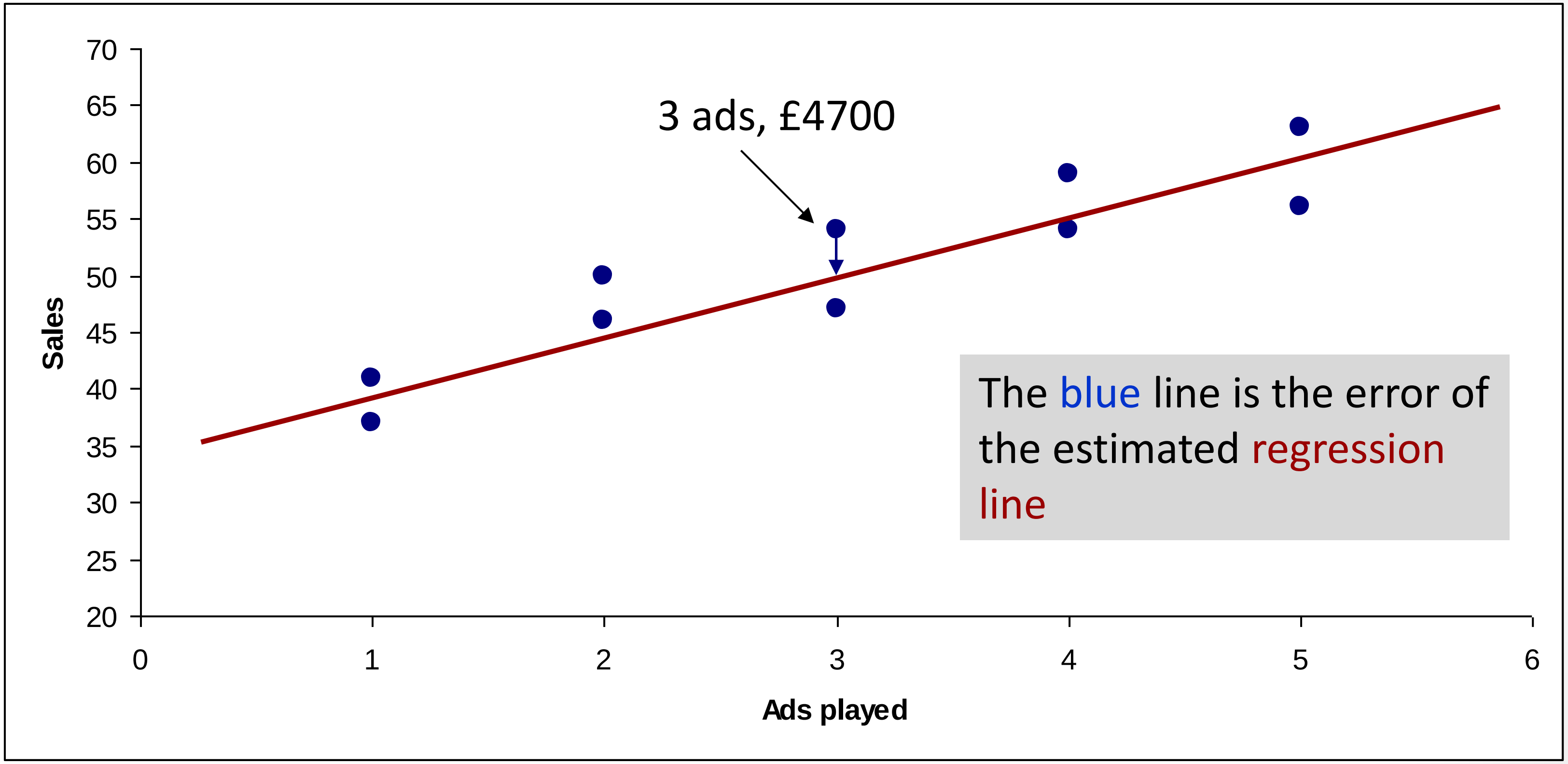
- Example: It has been found that the sales of ice cream is highly correlated with the rate of accidents in the swimming pool. Can we infer that an increase of swimming pool accidents leads to increased ice-cream sales?
  - no, because both are caused by rising temperature (summer months)!

Describing linear relationships- Regression analysis

- if two variables are strongly correlated, their relationship can be described using a simple linear expression
  - $y = a + bx + e$

- y = dependent variable
  - what we are trying to predict or explain (eg sales)
- x = independent variable
  - what we think can be used to predict or explain the dependent variable (eg number of ads played)
- a, b → unknown parameters (regression coefficients)
  - these define the linear relationship (regression line)
    - a is the constant
    - b is the slope
- e is called the error or the residual of the regression
  - If correlation is not perfect, the regression will not be able to full describe the relationship between y and x
  - e is simply the part of the dependent variable that is not explained by the independent variable

Drawing the regression line



3 ads, £4700

The blue line is the error of the estimated regression line

In order to make sure we find the most accurate regression line, we need to draw it in such a way so that it minimises the square of all these errors

Method of least squares

- The equation is going to be of the form:
  - y = **a** +**b**x

- Apply the method of least squares to find the slope **b** and intercept **a** of this line
  - Excel has a number of formulas to calculate **a** and **b** individually
    - eg, SLOPE(), TREND(), LINEST(), etc
  - However, the 'Regression Analysis' option from the 'Data Analysis' add-in offers a more complete picture
    - See this week's hands-on session on how to install the 'Data Analysis' add-in to carry our a regression analysis

Regression output

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	35.850	2.596	13.808	0.000
# Ads played	4.950	0.783	6.323	0.000

Constant, a

Slope, b

Significance of independent variable. Variable significant if P-value **<0.05**

- The equation is going to be of the form: y = a +bx + e
- So, the regression model is:
  - Sales = 35.85 + 4.95 (# ads played) + e
- Forecasting: What would the sales be if the number of ads played was 3?
  - Sales = 35.85 + 4.95 (3) = 50.7

P-values

- regression coefficients are estimates and as such are subject to statistical error
  - regression provides both expected values and standard errors
    - these can be used to calculate p-values

- p-values in a regression model give the probability that the coefficient in question is equal to zero
  - if p-value=0.05, there is a 5% probability that the coefficient is equal to zero
  - the lower the p-value, the more confident we are that the variable in question can be used to ‘explain’ the depended variable
  - if the p-value is lower than a certain probability level (usually 5%), the coefficient is **statistically significant** (at 5%)

Coefficient of determination (R^2)

- How good is our regression model in ‘explaining’ the behaviour of the depended variable?
  - How close is our regression line to our observations?
- the Coefficient of Determination (R^2) measures ***how well the regression line fits the data***
  - it represents **the total variation of the dependent variable that is explained by the regression line**

- A value of R^2 near 1 indicates a good fit, where nearly all the variation in y is explained
  - the Regression Analysis’ option in Excel automatically calculates R^2 and adjusted R^2 values for a given regression model

Multiple regression analysis

- Sometimes, there are multiple variables that help explain the depended variable (y)
  - Eg, store sales (y) might depend on the size of the store (x_1), the number of different product types (x_2) and the distance from the nearest competitor (x_3)
  - In this case, we want to estimate:
    - $y = a + b_1 x_1 + +b_2 x_2 + \ldots. +b_n x_n$, where n is the number of independent variables
- Excel can easily accommodate multiple regression analysis through the ‘Data Analysis’ add-in
- Results are interpreted in a similar manner as the single regression case
  - p-values give the probability that the coefficient of the selected independent variable is 0
  - Goodness-of-fit is measured using the ***adjusted*** $\boldsymbol{R^2}$ instead of the simpler R^2 estimate

Is this a good model?

- Significance of parameters
  - we usually assess the relevance of each independent variable based on its p-value
    - if the p-value is above a certain threshold, the effect of the independent variable to the dependent variable is likely to be minimal
    - as with significance levels, this threshold is usually set at 5%
- Coefficient of determination
  - for our model to be useful for forecasting, it needs to explain a large proportion of the variation of the dependent variable
  - as with the correlation coefficient, ‘large’ is subjective
    - as a rule of thump, we would like (adjusted) R^2 to be above 80%
    - anything below 60% is probably not suited for forecasting