

# Session 7.5: Correlation and Linear Regression

Dr Annabel Bradbury, TRL Limited

- 1 Relationships
- 2 Correlation coefficients
- 3 Linear regression
- 4 Confidence intervals

- 1 Relationships
- 2 Correlation coefficients
- 3 Linear regression
- 4 Confidence intervals

## 1: Up and down - the randomness of numbers

- “There is more chance about than many of us think”  
(Blastland and Dilnot)
- *Numbers go up and down for no more reason than chance* - without any intervention!
- “One big wave does not mean a rising tide”
- Mistaking chance ups or downs can have dire consequences:
  - a failure to understand what really works
  - spending money on things that don't work
  - ignoring what does work

## Chance

- *When 2 things happen together they are often related - but not always!*
- Apparently unusual events happening simultaneously do not necessarily share the same cause
- Chance distributions are often clustered; but gaps also occur
- Most phenomena have multiple causes
- Clusters can be big!
- When we see patterns or clusters in numbers everyone wants an explanation; the most overlooked explanation is that there is no explanation but that it was just a matter of chance
- How to minimise the effect of chance? Study a large number of a population

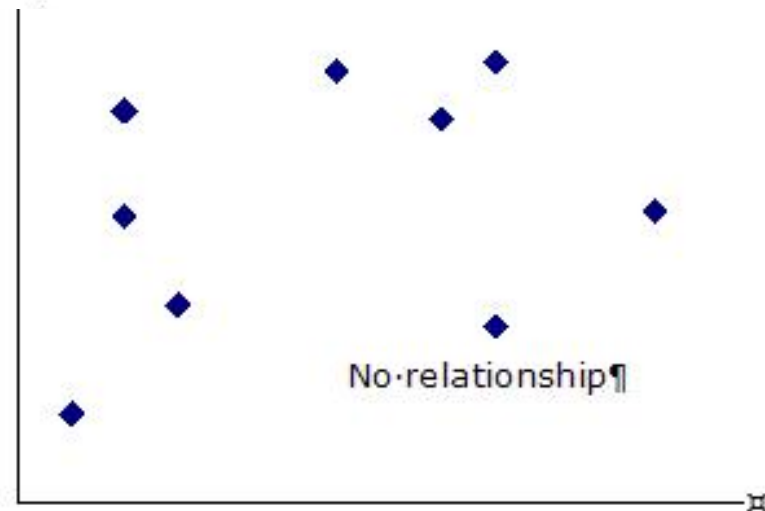
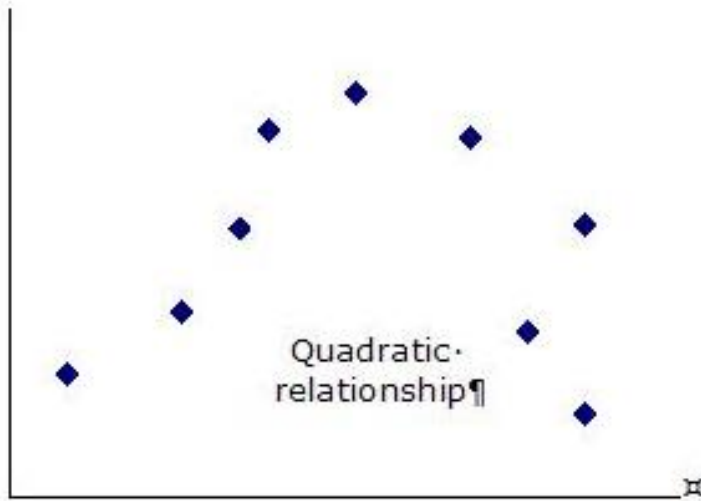
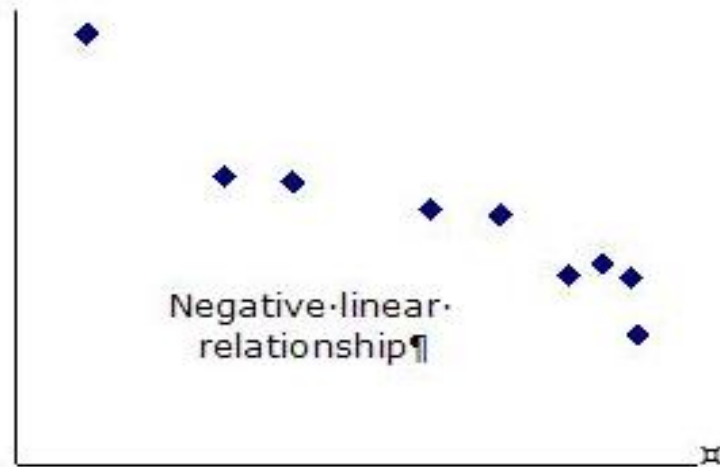
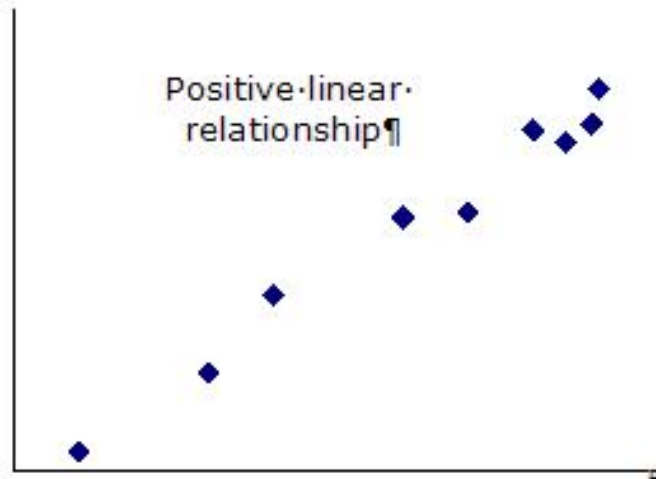
### Coins

- Coins are unbiased
- Heads and tails should occur about evenly
- But there will be sequences

### Test

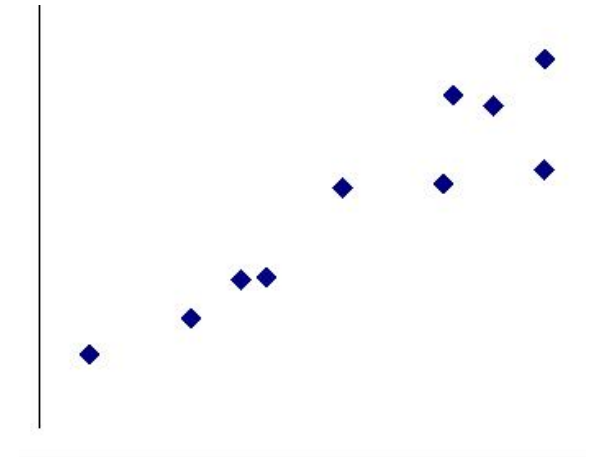
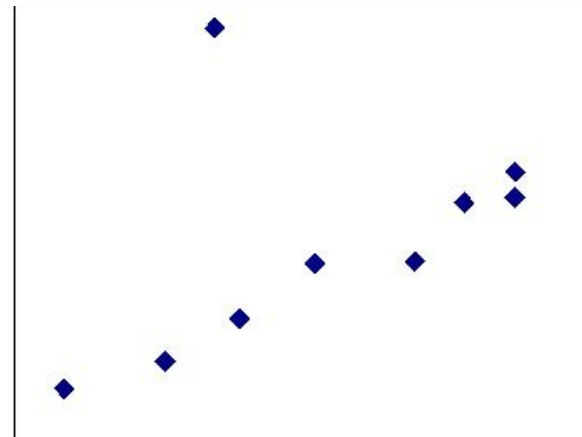
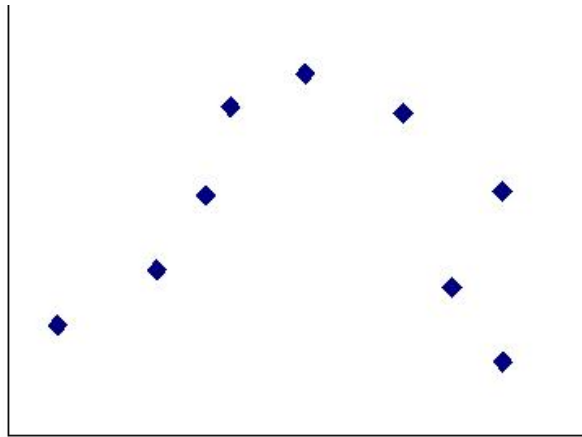
- Toss a coin 10 times
- Record heads and tails
- Sum heads and tail – see if they are equal
- See which is longest sequence

## Is there a relationship between 2 continuous variables?



## It is always helpful to draw a scatter plot first!

- Non linear relationship (transform x values?)
- Outliers (investigate why?)
- Equal variances
- Check for normality



- **Dependence** refers to any statistical relationship between two random variables or two sets of data
- **Correlation** refers to any of a broad class of statistical relationships involving dependence
- Dependent or correlated variables are not independent

- 1 Relationships
- 2 Correlation coefficients
- 3 Linear regression
- 4 Confidence intervals

- (Pearson's or Spearman's rank) Correlation Coefficient:
  - A measure of the strength and direction of the linear relationship between two variables:

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

- Coefficient of Determination= $R^2$ 
  - A measure of how well future outcomes are likely to be predicted by the model

## Example: Calculate correlation coefficient

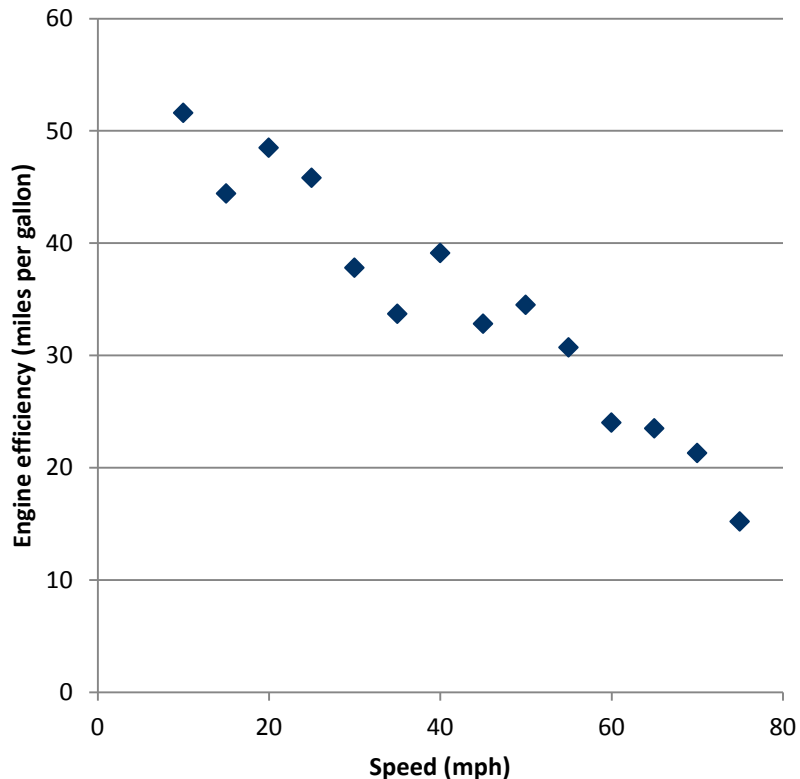
### Correlation coefficient

- Direction and strength of linear relationship
  - +/- tells you whether relationship is positive or negative
  - Value tells you how strong a relationship it is
- Produce a scatter plot using Excel and compute the correlation (dataset mpg)

**Data → data analysis → correlation**

## Example: Scatter plot and correlation coefficient

Engine efficiency v speed data



Excel correlation output

	<i>speed</i>	<i>mpg</i>
<i>speed</i>	1.0	-0.966
<i>mpg</i>	-0.966	1.0

## Interpreting your correlation output

### Correlation coefficient

- **Weak relationship**

R between (0.10 and 0.29) or (-0.10 and -0.29) is small

- **Reasonable relationship**

R between (0.30 and 0.49) or (-0.30 and -0.49) is medium

- **Strong relationship**

R between (0.50 and 1.00) or (-0.50 and -1.00) is large

### Coefficient of determination

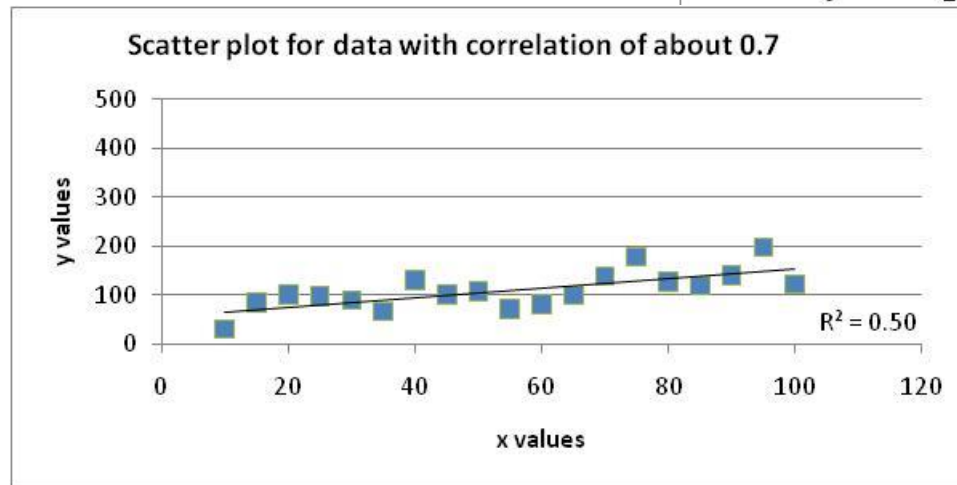
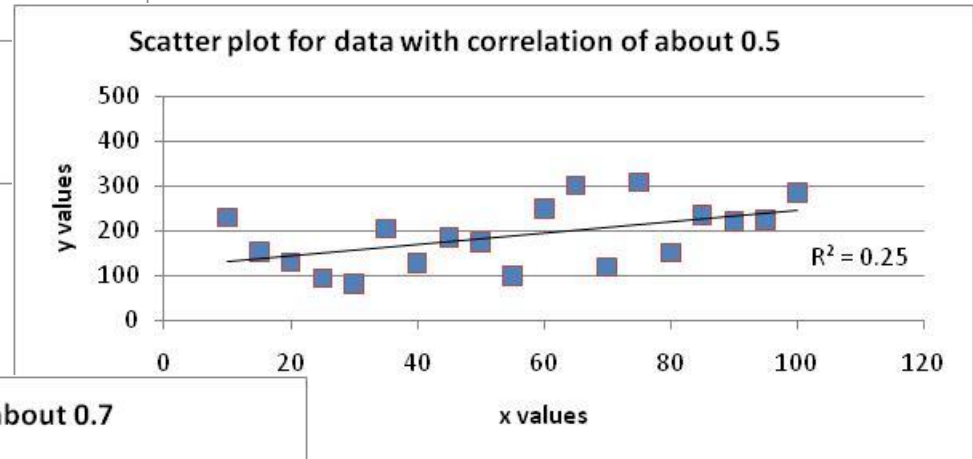
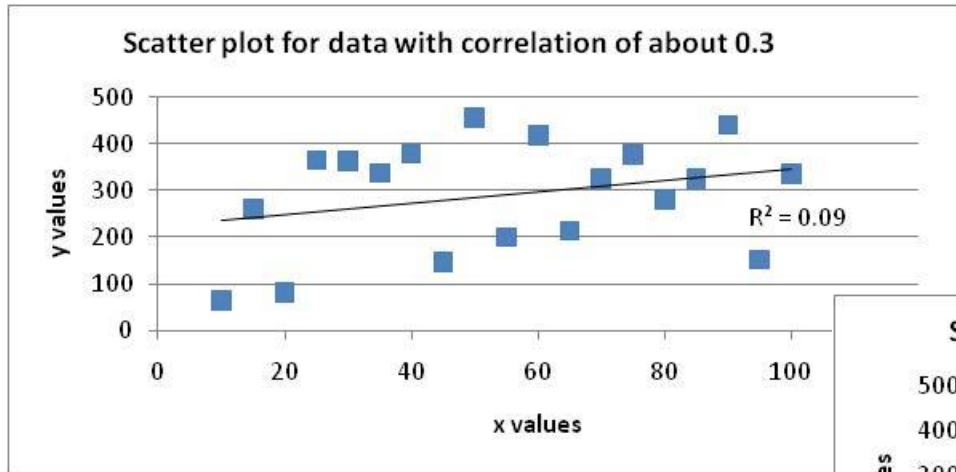
- How much variation the two variables share
- Square r and multiply by 100%

- In the previous example of engine efficiency and speed :

$$R^2 = (0.966)^2 = 93.3\%$$

Only 6.7% of the “noise” (error) in the engine efficiency data is unexplained by the speed being travelled

## Interpreting your correlation output



- 1 Relationships
- 2 Correlation coefficients
- 3 Linear regression
- 4 Confidence intervals

### The theory

- Once a relationship has been discovered we need to estimate the relationship

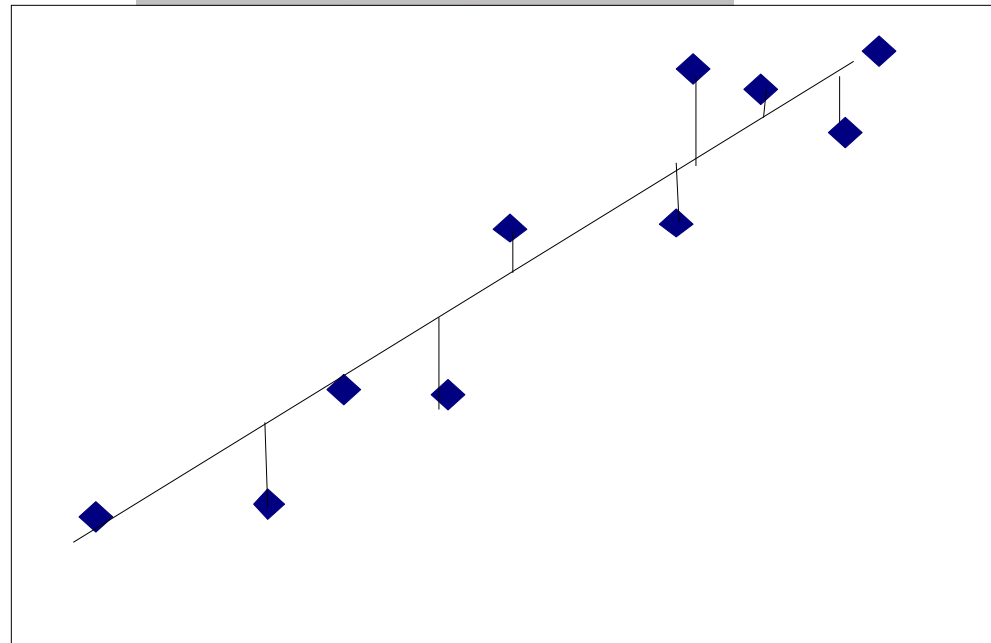
### Variables

- Dependent (or response) variable (y):
  - Values controlled and selected by experimenter
  - Not subject to experimental variation
- Explanatory or regressor variable or (x):
  - Observed change in response to explanatory variable
  - Has associated variability

## The method

- Trying to minimise the vertical distance between observed and fitted data
- Errors are the vertical lines  $\varepsilon_i = (y_i - \text{fitted } y_i)$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$S = \sum \varepsilon_i^2$$

## Calculating

- Slope of the regression line is

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

- Precision of the estimate is given by the standard error

$$s.e.(\hat{\beta}_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- Where  $s$  is the standard error of the estimate

## Testing

- Hypotheses:
  - $H_0: \beta_1 = 0$  (no significant slope)
  - $H_1: \beta_1 \neq 0$
- Student's t-test

$$t = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)}$$

- Confidence interval

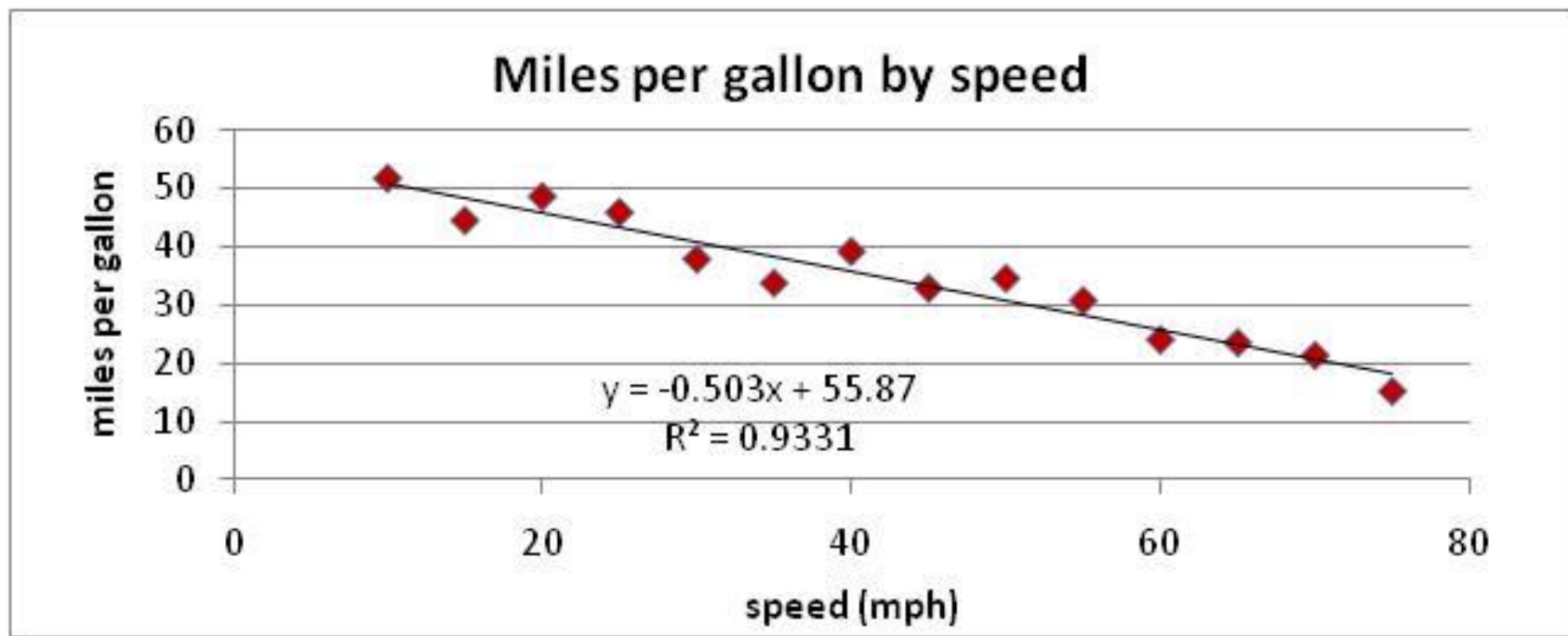
$$\hat{\beta}_1 \pm t_{n-2} * se(\hat{\beta}_1)$$

# Linear regression in Excel

## – using scatterplot

### Model summary

- Equation:  $\text{Mpg} = \text{'constant'} + \text{'slope'} * \text{speed} + \text{'error'}$
- Coefficient of determination (or R-squared)



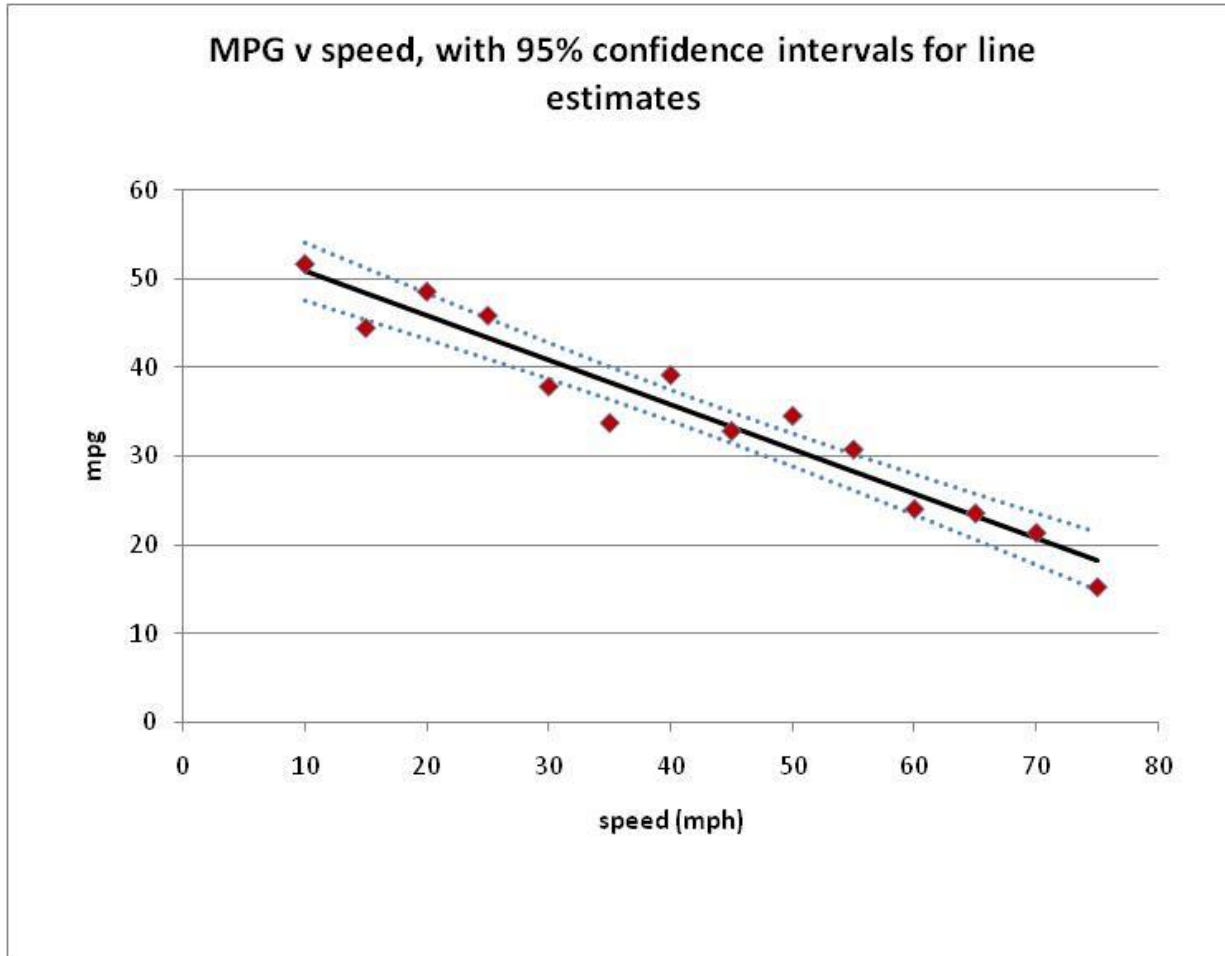
## Interpreting regression output - model summary

- Model summary shows:
  - Correlation coefficient (R ) – multiple correlation between the observed values and those fitted from the linear regression.
  - Coefficient of determination (R square).
  - Adjusted R square is an R squared value which takes into account the number of parameters in the regression (the coefficient of determination to use).
  - Standard error allows significance tests to be made on R.

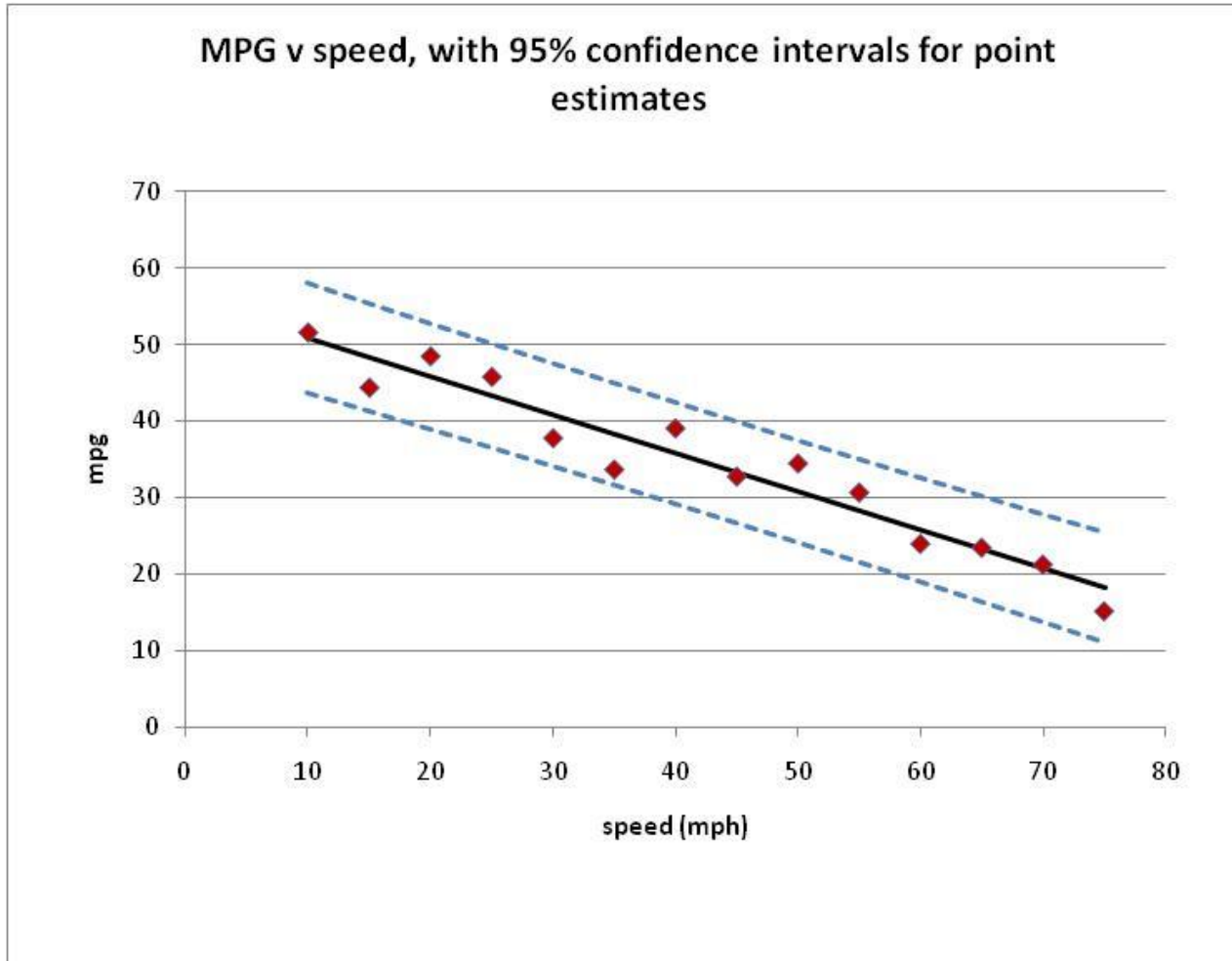
- 1 Relationships
- 2 Correlation coefficients
- 3 Linear Regression
- 4 Confidence intervals

- A **confidence interval (CI)** is an interval estimate of a population parameter used to indicate the reliability of an estimate. It is an observed interval (i.e. it is calculated from the observations), in principle different from sample to sample, that frequently includes the parameter of interest, if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the **confidence level**.
- **Confidence intervals** consist of a range of values (interval) that act as good estimates of the unknown population parameter.
- It is represented by a percentage; when we say, "we are 99% confident that the true value of the parameter is within our confidence intervals", we mean that 99% of the observed confidence intervals will hold the true value of the parameter.

## 95% confidence interval about the mean ('dotted')



# individual point estimates ('dashed')



## Key lessons

- *Correlation does not prove causation*
- Plausibility is part of the problem – we should demand more rigorous proof of causality
- Watch out for third factors that are influencing results
- *Need to eliminate all other possible causes of effect*



**Do You  
Have Any  
Questions?**