Least Squares Regression

Using data to make predictions

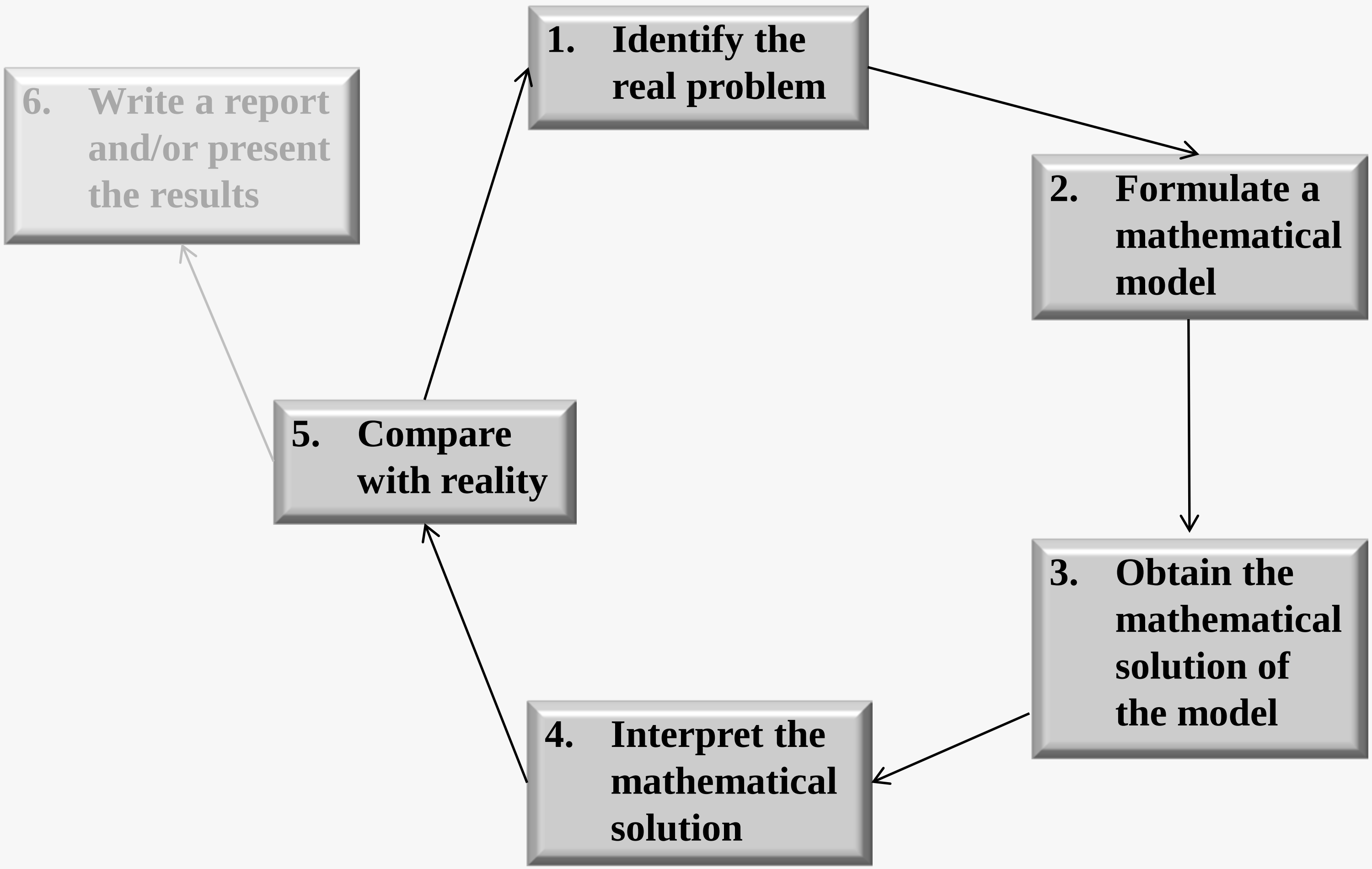
- May need to solve a problem by answering questions like
  - "what is the relationship between quantity of carbon monoxide and temperature in the Earth's atmosphere"
  - "How many broadband users will there be in the UK in five years from now"
- We need to create an equation that describes the dependent (response) variable(s) with the independent variable
- Used mainly for estimating and predicting

What is the relationship?

year	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Price (pence)	51.8	54.6	54.7	59.9	65.6	69.8	84.3	78.9	74.0	74.4	81.7

- Data for average unleaded fuel prices in Scotland.
- The mechanistic model between the petrol price and a given year is not clear (unknown)
- How do we build an empirical model for this data?

Modelling flowchart



General principles of data analysis

PLOT your data

To understand the data, always start with a series of graphs

INTERPRET what you see

Look for overall pattern and deviations from that pattern

Numerical **SUMMARY?**

Choose an appropriate measure to describe the pattern and deviation

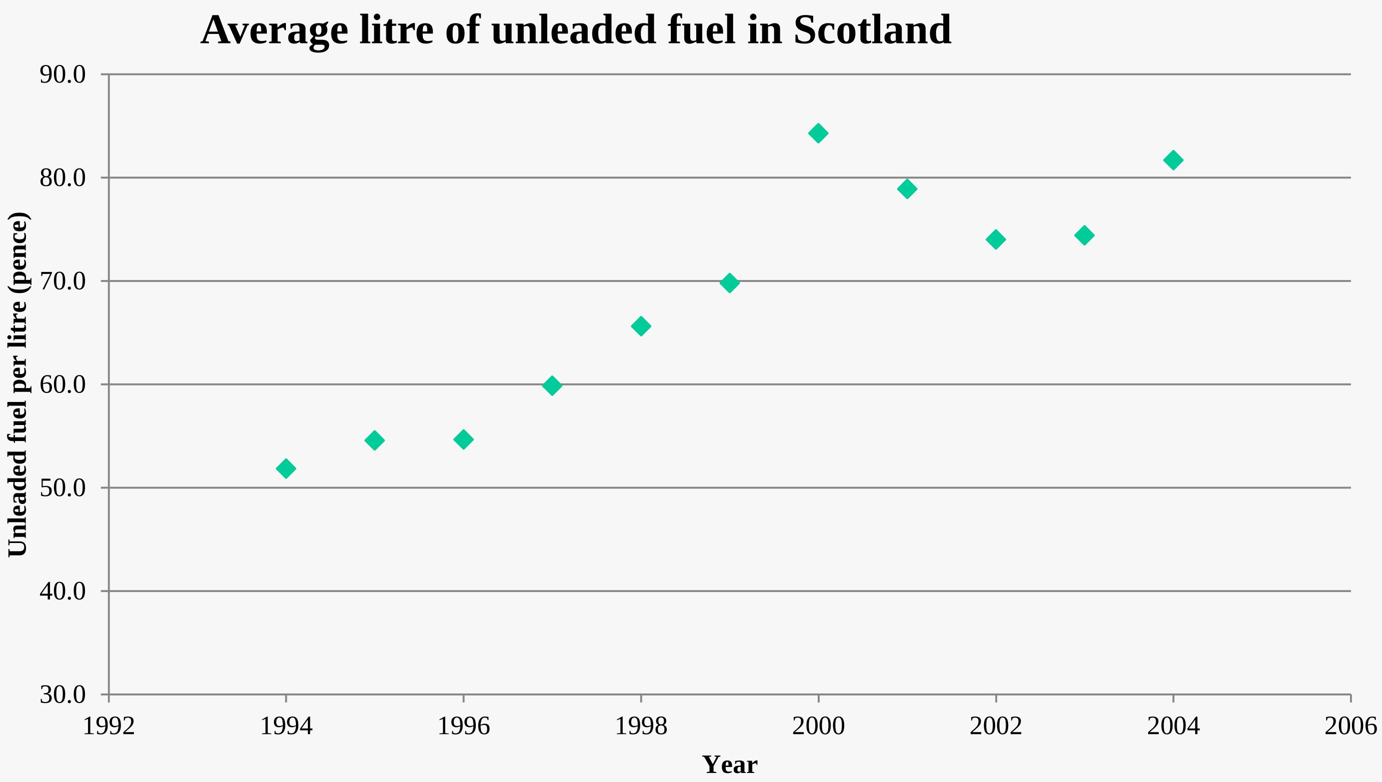
Mathematical **MODEL?**

If the pattern is regular, summarize the data in a compact mathematical model

What do we see...

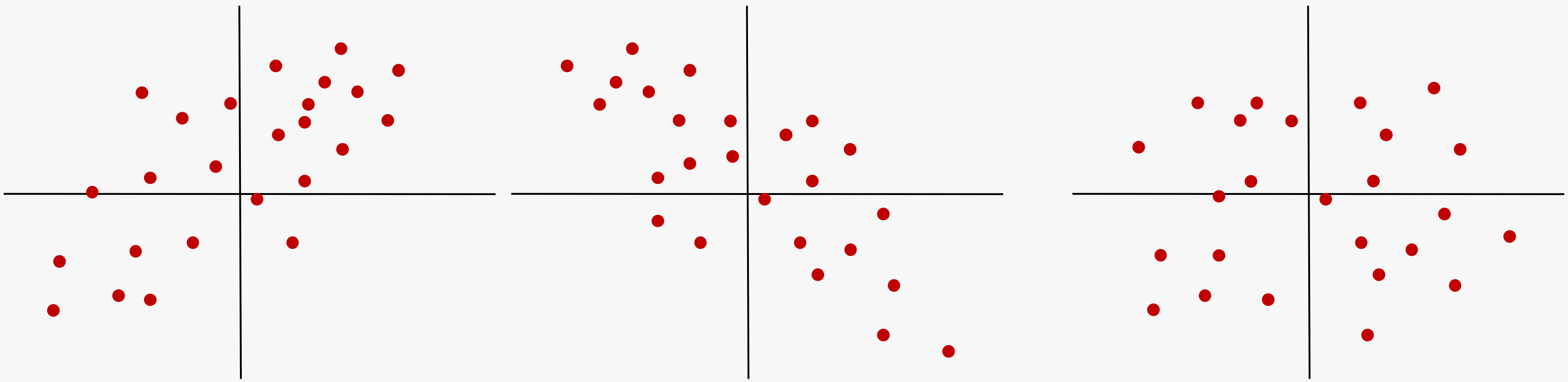
- The first thing to do when building this "***model***" is to plot the data.
- To do this we select one of the variants and label it ***x*** and label the other variant ***y***
- Set ***x*** to be the **year** and ***y*** to be the **price of unleaded fuel** (in pence)
- This graph is called a scatter plot/diagram and can be very informative

Scatter diagram of the data



- Now what do you see ????

Examples of scatter diagrams



General principles of data analysis

PLOT your data

To understand the data, always start with a series of graphs

INTERPRET what you see

Look for overall pattern and deviations from that pattern

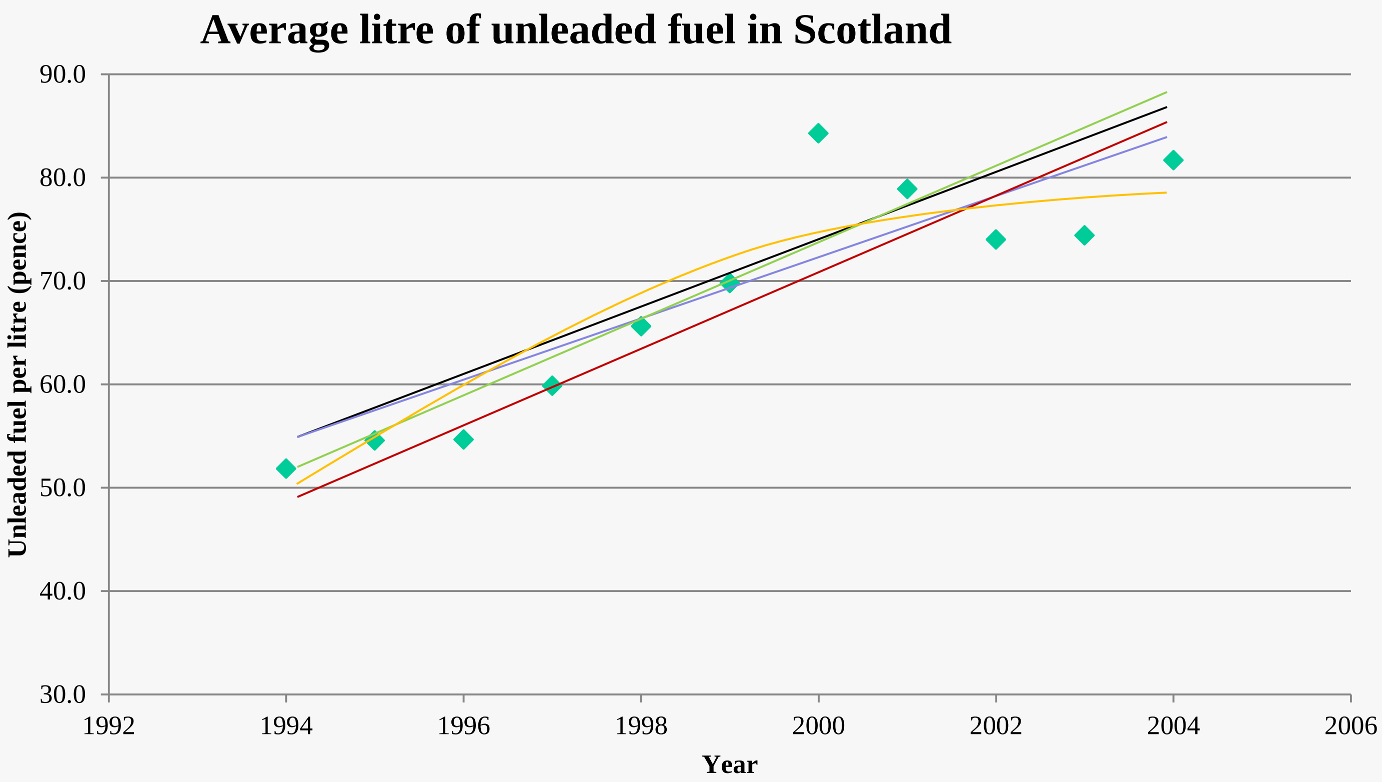
Numerical **SUMMARY?**

Choose an appropriate measure to describe the pattern and deviation

Mathematical **MODEL?**

If the pattern is regular, summarize the data in a compact mathematical model

Suggested models

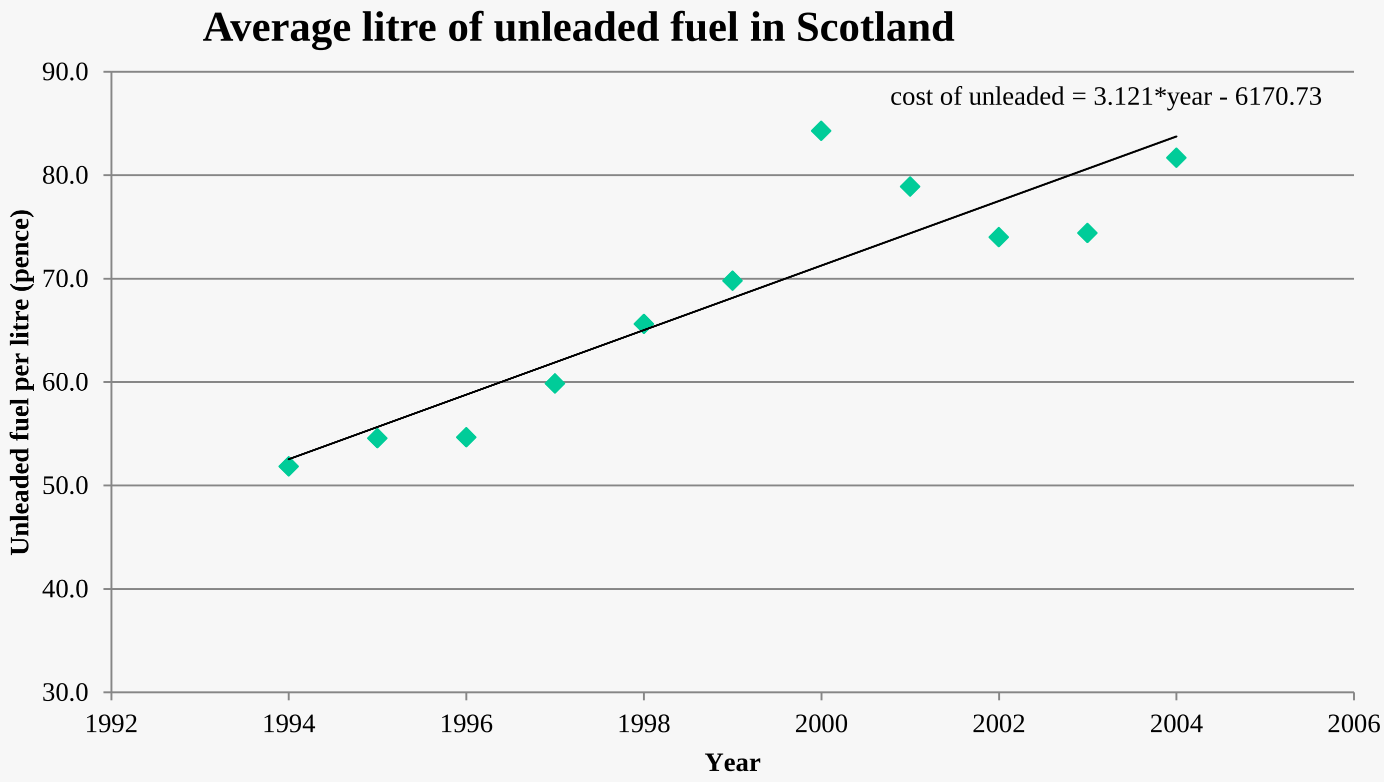


- One outlier in year 2000 otherwise a possible straight line fit
- What's the **best** line through these points ?

General principles of data analysis

PLOT your data	To understand the data, always start with a series of graphs
INTERPRET what you see	Look for overall pattern and deviations from that pattern
Numerical **SUMMARY?**	Choose an appropriate measure to describe the pattern and deviation
Mathematical **MODEL?**	If the pattern is regular, summarize the data in a compact mathematical model

A simple model – straight line



- Here’s the “**best**” so how do we get this?

General principles of data analysis

PLOT your data

To understand the data, always start with a series of graphs

↓

INTERPRET what you see

Look for overall pattern and deviations from that pattern

↓

Numerical **SUMMARY?**

Choose an appropriate measure to describe the pattern and deviation

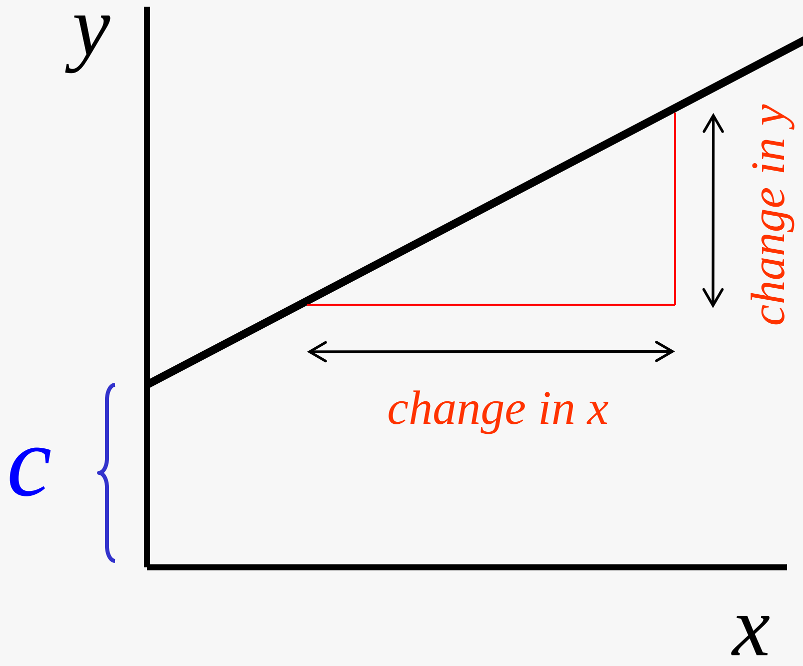
↓

Mathematical **MODEL?**

If the pattern is regular, summarize the data in a compact mathematical model

Back to school first...

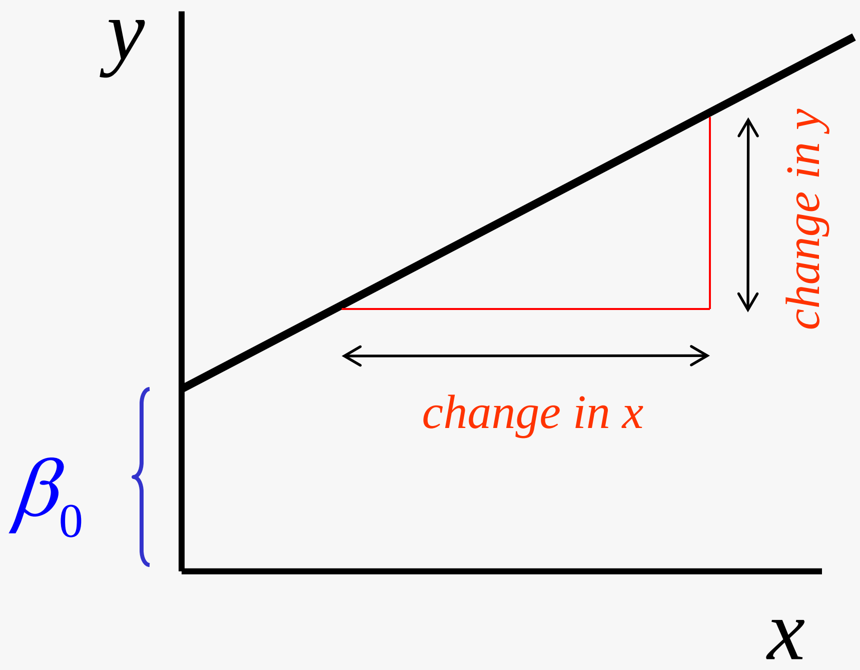
- Simple equation of a straight line $y = mx + c$
- m is the slope and c is the intercept



$$m = \frac{change\ in\ y}{change\ in\ x}$$

Update our notation

- Simple equation of a straight line $y = \beta_1 x + \beta_0$
- β_1 is the slope and β_0 is the intercept

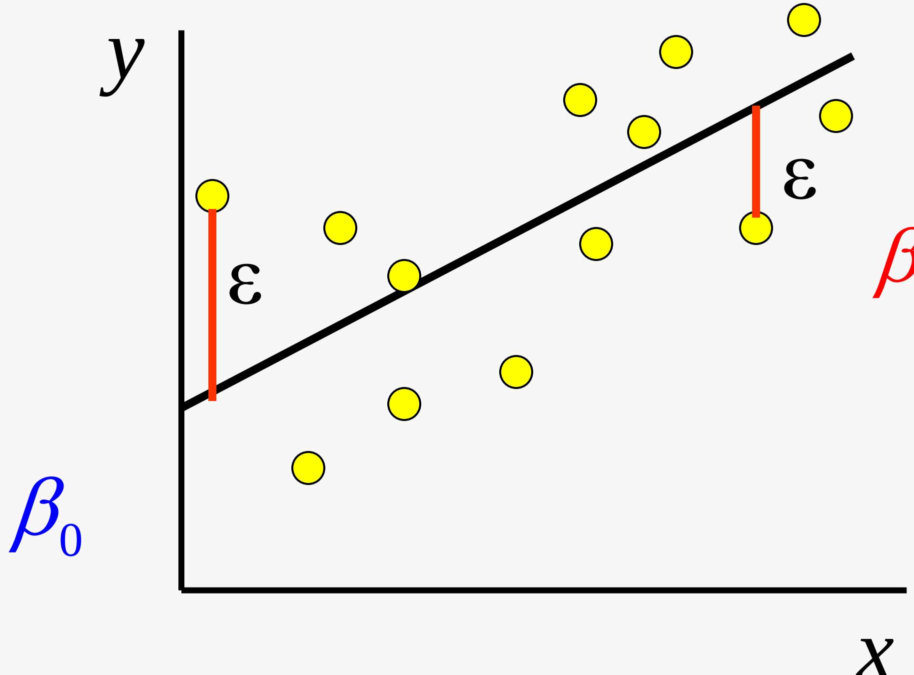


$$\beta_1 = \frac{\text{change in } y}{\text{change in } x}$$

Errors or residuals

- For each *x* value there is a *y* value which is different from the actual *y* data so

$$\text{actual } \boldsymbol{y} = \boldsymbol{\beta_1 x} + \boldsymbol{\beta_0} + \varepsilon$$



$$\beta_1 = \frac{\text{change in } y}{\text{change in } x}$$

- We try to minimise these errors squared (ε^2) to get the best fit
- Ideal case all points coincide with line so there are no errors

Simple linear regression

- The line we are trying to fit is given by
- $$y = \beta_1 x + \beta_0$$
- This is called the ***simple linear regression*** line because
  1. The model is ***simple*** because there is only one independent variable (x) in the model
  2. The model is ***linear*** because it is linear in the regression coefficients β_1 and β_0
- We need to calculate the regression coefficients β_1 and β_0

$$\beta_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \qquad \beta_0 = \overline{y} - \beta_1 \overline{x}$$

- although the formulae may look complicated they are not!!!

Calculating the regression coefficients

- First calculate the *mean of all the x values* ($\overline{x}$) and the *mean of all the y values* ($\overline{y}$) given by the formulae

$$\overline{x} = \frac{sum\ of\ all\ x\ values}{number\ of\ x\ values} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \overline{y} = \frac{sum\ of\ all\ y\ values}{number\ of\ y\ values} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- Next calculate the *sum of squares of xy (S_{xy})* and the *sum of squares of xx (S_{xx})*

$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y})$$

$$S_{xx} = \sum (x_i - \overline{x})^2$$

- Finally we get $\beta_1 = \dfrac{S_{xy}}{S_{xx}}$ and $\beta_0 = \overline{y} - \beta_1 \overline{x}$

For our unleaded fuel data

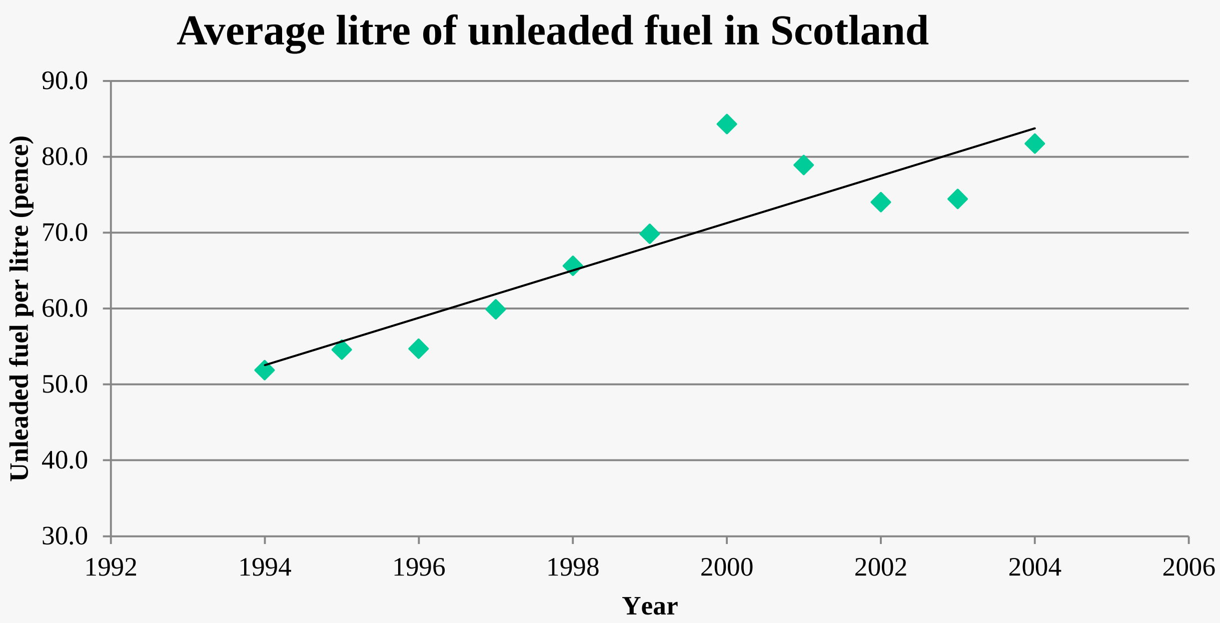
												sum of values	mean of values
year (x)	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	21989	1999
Price in pence (y)	51.8	54.6	54.7	59.9	65.6	69.8	84.3	78.9	74.0	74.4	81.7	749.65	68.15
y-mean of y	-16.3	-13.6	-13.5	-8.3	-2.5	1.6	16.2	10.8	5.8	6.3	13.6		
x-mean of x	-5	-4	-3	-2	-1	0	1	2	3	4	5		
(x-mean of x)(y-mean of y)	81.55	54.36	40.44	16.58	2.53	0	16.15	21.5	17.55	25	67.75	343.41	
(x-mean of x)²	25	16	9	4	1	0	1	4	9	16	25	110	

- So n=11, $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = 343.31$

$$S_{xx} = \sum (x_i - \bar{x})^2 = 110$$

$$\beta_1 = \frac{S_{xy}}{S_{xx}} = \frac{343.31}{110} = 3.121 \quad \beta_0 = \bar{y} - \beta_1 \bar{x} = 68.15 - 3.121 * 1999 = -6170.73$$

Interpretations from possible model

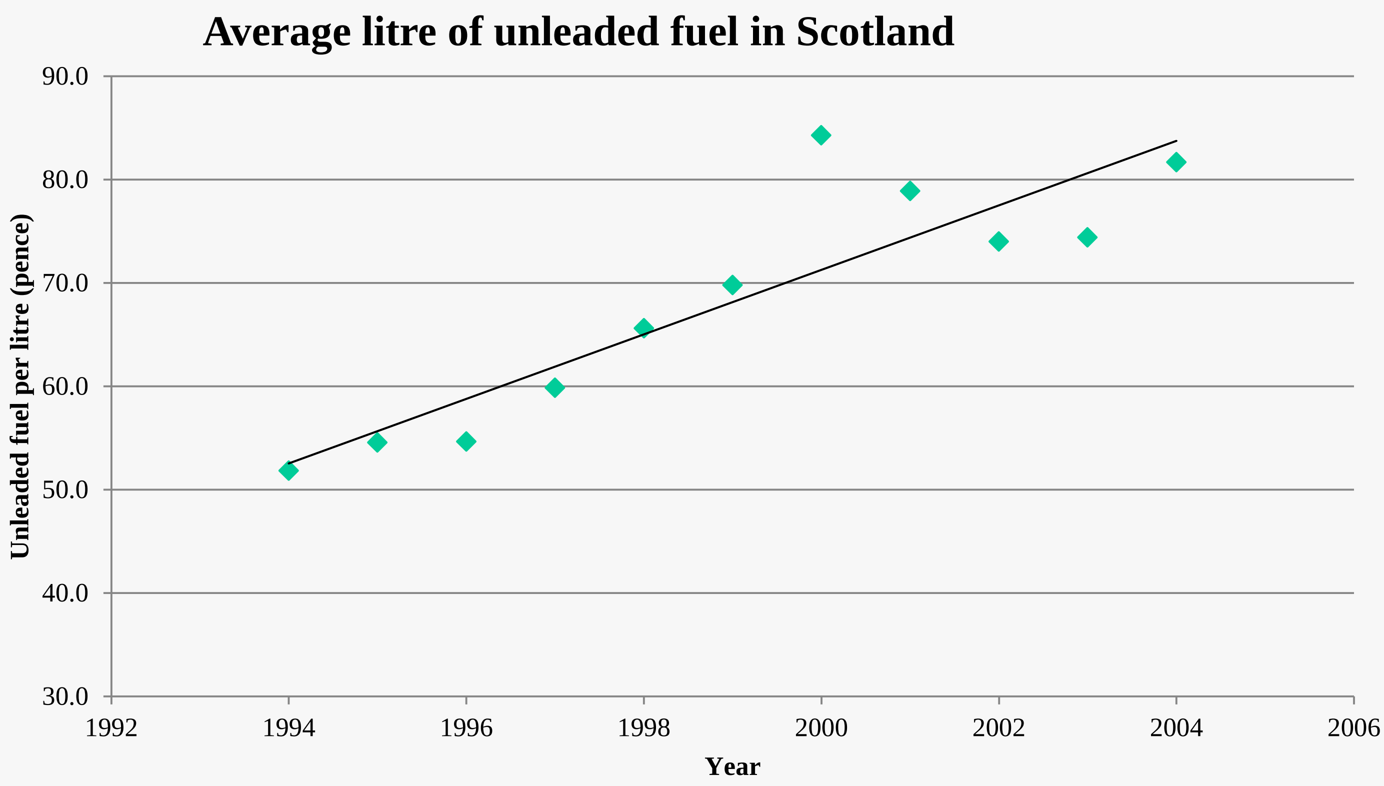


- cost of unleaded = 3.121*year - 6170.73
- model suggests
  - A positive relationship between the fuel price and time
  - cost of unleaded fuel has risen over 3p a year from 1994-2004

General principles of data analysis

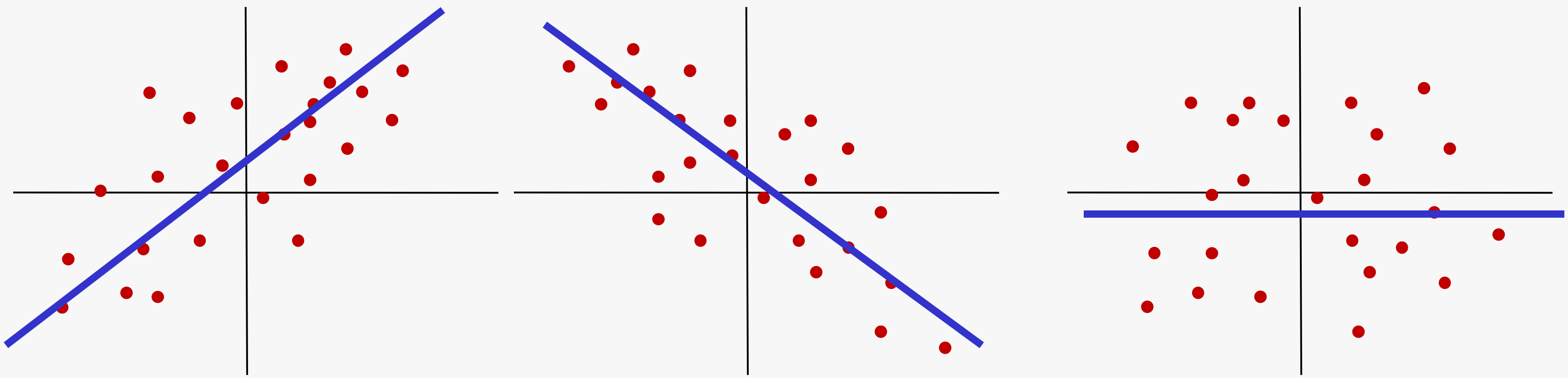
PLOT your data	To understand the data, always start with a series of graphs
INTERPRET what you see	Look for overall pattern and deviations from that pattern
Numerical **SUMMARY?**	Choose an appropriate measure to describe the pattern and deviation
Mathematical **MODEL?**	If the pattern is regular, summarize the data in a compact mathematical model

How good is it?



cost of unleaded = 3.121*year - 6170.73

Correlation



Positive correlation Negative correlation No correlation

- The ***correlation*** between the *x* and *y* data gives an idea of how well the model fits the data

Correlation coefficient (R^2)

- Need a measure of how strong or weak the correlation is between the two variables.

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

- Compute coefficient between 0 and 1
- 0 means no correlation
- 1 means perfect fit to data
- For our fuel data

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{343.41^2}{110 * 1343.5} = 0.798 \approx 0.8$$

- So this is a good fit and there is a strong **LINEAR** correlation between the fuel price and time range of 1994-2004.

How good as a predictor?

- We can use the model to predict intermediate values, we call this ***interpolating***

- This is generally safe but as we can see we have an outlier so if the model is used to predict the value of fuel in 2000 what will it be?

$$\text{cost of unleaded}_{2000} = 3.121*2000 - 6170.73 = 71.3\text{p}$$

the actual value was 84.3p this represents a worst case error of 15% !!!! (NOTE: 100* (84.3-71.3)/84.3)

- What was the average fuel price between 1994 and 2004?

this is given by $\bar{y} = 68.2\text{p}$

How good as a predictor?

- We can use the model to predict future values, but must exercise caution because we are ***extrapolating***
- What was the fuel price in 1992?

$$\text{cost of unleaded}_{1992} = 3.121*1992 - 6170.73 = 46.3\text{p}$$

sounds reasonable?

- What was the fuel price in 2010?

$$\text{cost of unleaded}_{2010} = 3.121*2010 - 6170.73 = 102.5\text{p}$$

sounds reasonable?

- What was the fuel price in 1800?

$$\text{cost of unleaded}_{1800} = 3.121*1800 - 6170.73 = -552.9\text{p}$$

sounds reasonable?

Final thoughts

- Developing a simple model to represent a set of data can be an effective way to
  - Use the model to answer questions about the data within the range (interpolation)
  - Use the model to answer questions about the data beyond the range (extrapolation)
- An analysis of how good the "fit" of the straight line is can be made by determining the correlation coefficient
- Always question the sensibility of the results