Statistics II

Railway Engineering Mathematics

Sheffield Hallam University

Lecture 24

Learning Outcomes

- Working with grouped data.
- Visualising grouped data using frequency polygons in EXCEL.
- Fitting simple curves and trendlines using EXCEL.

Grouped Data

When there are many different measurements with few/no repetition or just a large number of data, then it is only possible to make any real sense of the data if they are grouped together in intervals.

Quite often we are given data which is already grouped.

In this case, the formulae for calculating values such as averages and dispersion are slightly different, as we have to use the mid-point of the class as an estimate for all the data in that grouping.

Example 1: Grouped, continuous data

The heights (in cm) of 25 people of the same age were measured. The following table shows the data:

180.84	164.87	167.77	167.78	174.39
176.14	176.87	159.57	164.73	174.51
168.47	180.64	170.04	162.71	174.02
171.91	169.31	171.68	152.49	177.58
172.03	169.68	161.87	165.48	181.90

Summarise the data into a frequency table and find the mean.

Example 1: Grouped, continuous data

Group	Freq. f	Midpoint x_i	fx_i	C. f
$150 < x \leq 155$	1	152.5	152.5	1
$155 < x \leq 160$	1	157.5	157.5	2
$160 < x \leq 165$	4	162.5	650	6
$165 < x \leq 170$	6	167.5	1005	12
$170 < x \leq 175$	7	172.5	1207.5	19
$175 < x \leq 180$	3	177.5	532.5	22
$180 < x \leq 185$	3	182.5	547.5	25
	$\sum f = 25$		$\sum fx_i = 4257.5$	

$$\bar{x} = \frac{\sum fx_i}{\sum f} = \frac{4252.5}{25} = 170.1$$

The median is the 13^{th} value, thus 172.5 from the CF or the 170-175 group. The modal group is 170-175.

Graphing Data

Graphs can be used to quickly determine key characteristics of the data under analysis. Some possible graphs are:

- Bar charts and pie charts (discrete or qualitative data).
- Histograms/Frequency distributions (continuous, quantitative data).
- Frequency polygons.
- Cumulative frequency curves.

Example 2: Grouped, continuous data

The resistance (in kΩ) of a set of resistors are given in the following table:

Resistance R (kΩ)	Frequency f
$1.1 < R \leq 1.3$	9
$1.3 < R \leq 1.5$	22
$1.5 < R \leq 1.7$	15
$1.7 < R \leq 1.9$	11
$1.9 < R \leq 2.1$	8
$2.1 < R \leq 2.3$	5

Example 2: Grouped, continuous data

Determine the mean, median, mode and standard deviation (population). Plot a frequency polygon.

From the spreadsheet, we can determine that the mean is 1.61 kΩ, the median group is $1.5 < R \leq 1.7$ (the median value is the 35.5th value) and the modal group is $1.3 < R \leq 1.5$.

The standard deviation (population) formula for grouped data is as follows:

$$\sigma = \sqrt{\frac{\sum f\,(x - \overline{x})^2}{n}} \quad \text{grouped data}$$

From the spreadsheet, we can determine that the standard deviation is 0.29 kΩ.

Example 2: Grouped, continuous data

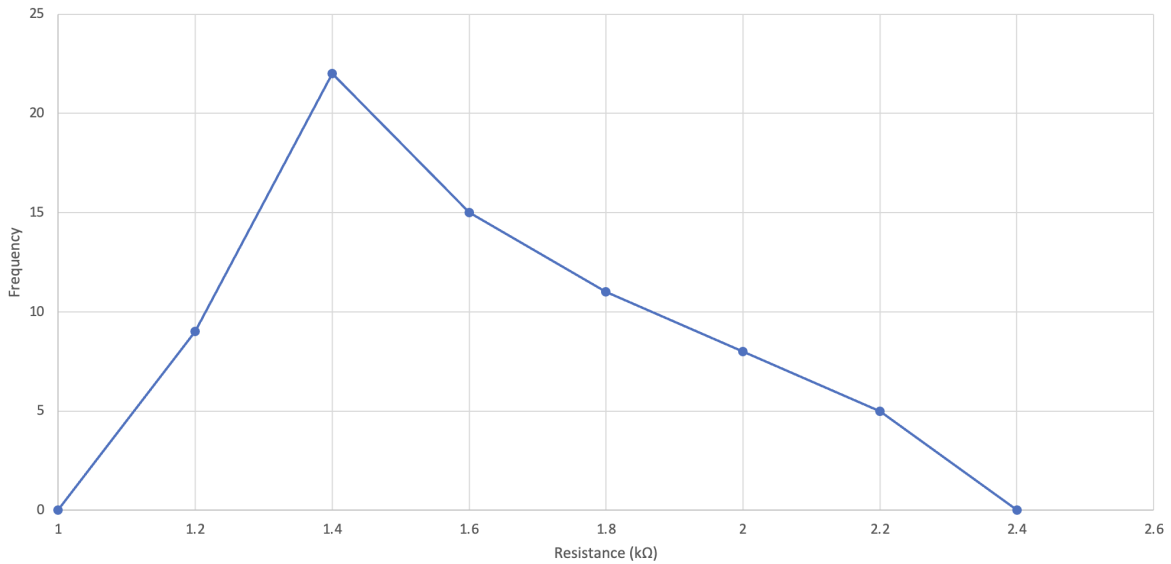
To plot the frequency polygon, we must first add a row at the start and the end of the data:

x_lower	x_upper	x_mid	f	f * x_mid
0.9	1.1	1	0	0
1.1	1.3	1.2	9	10.8
1.3	1.5	1.4	22	30.8
1.5	1.7	1.6	15	24
1.7	1.9	1.8	11	19.8
1.9	2.1	2	8	16
2.1	2.3	2.2	5	11
2.3	2.5	2.4	0	0
			70	112.4

This is so the curve will form a closed shape with the horizontal axis.

Example 2: Grouped, continuous data

The frequency is then plotted against the resistance (mid-point):



Here we have a visual of the spread of data.

Curve fitting

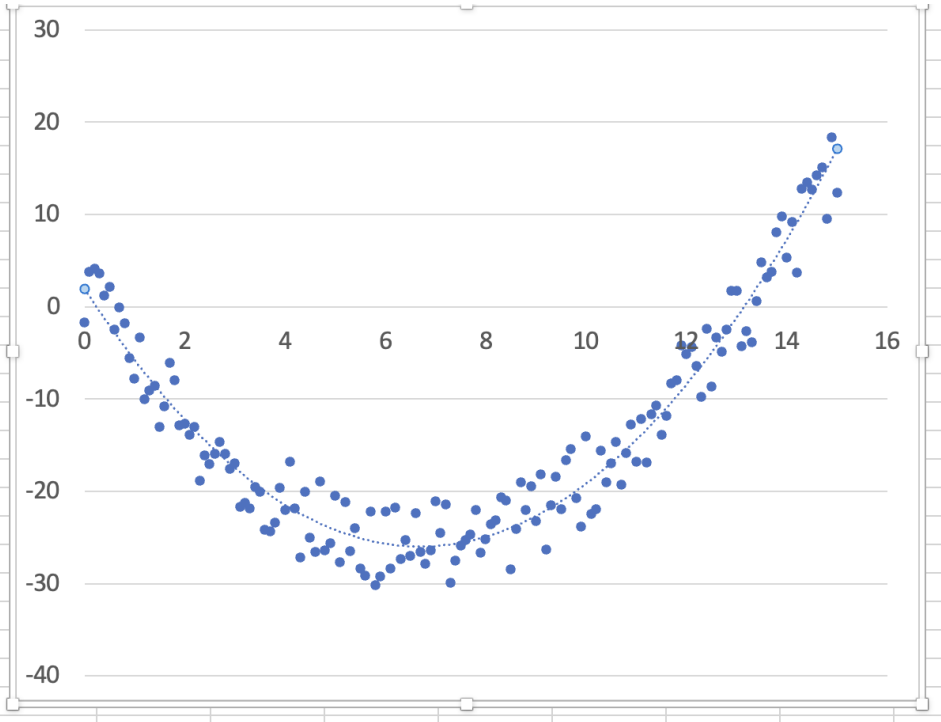
Often we can obtain a set of experimental data, and hypothesise that the relationship between the independent variable (that we control) and the dependent variable (that we measure) is described by some function. If we could determine the exact relationship, we could make further predictions by extrapolating the fitted curve.

Once we have decided on a general form of the relationship between our variables (e.g. linear, quadratic, exponential, power law), curve fitting is the process of **finding the set of parameter values** that best fits the set of experimental data.

Example 3

Suppose we have the set of data shown below:

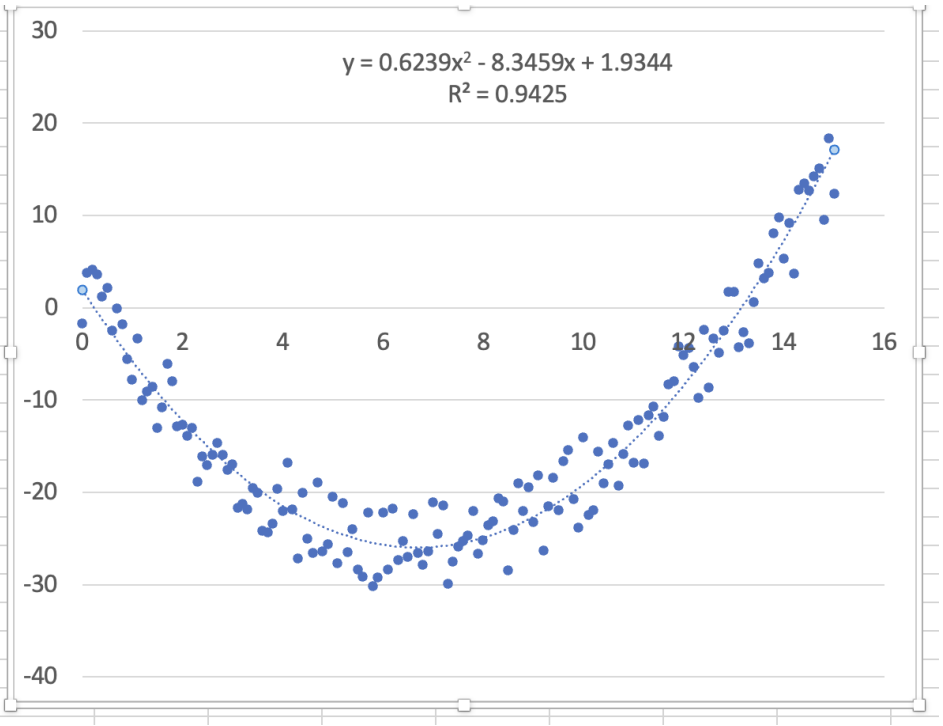
x	y
0	-1.75
0.1	3.756
0.2	4.124
0.3	3.554
0.4	1.196
0.5	2.1
0.6	-2.484
0.7	-0.106
0.8	-1.866
0.9	-5.614
1	-7.85
1.1	-3.374
1.2	-10.086
1.3	-9.136
1.4	-8.574
1.5	-13.05
1.6	-10.814
1.7	-6.116
1.8	-8.006
1.9	-12.884
2	-12.75
2.1	-13.954
2.2	-13.096
2.3	-18.876

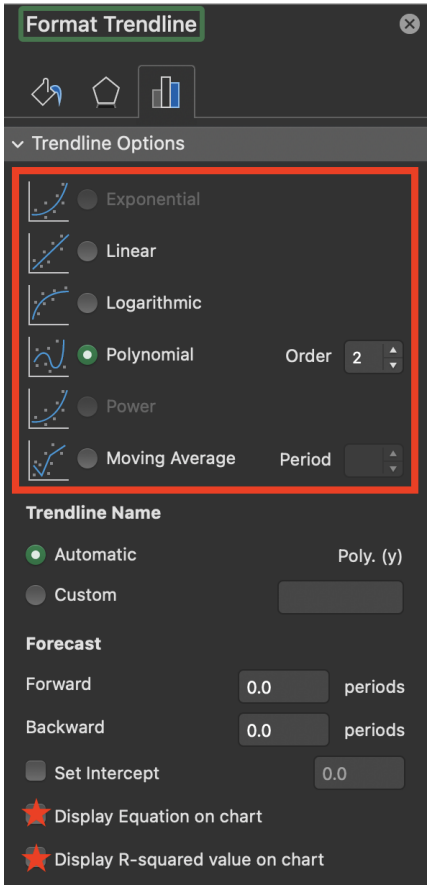


Perhaps a quadratic function $y = ax^2 + bx + c$ would fit. But what values of a, b and c would result in the *best* fit?

Example 3

x	y
0	-1.75
0.1	3.756
0.2	4.124
0.3	3.554
0.4	1.196
0.5	2.1
0.6	-2.484
0.7	-0.106
0.8	-1.866
0.9	-5.614
1	-7.85
1.1	-3.374
1.2	-10.086
1.3	-9.136
1.4	-8.574
1.5	-13.05
1.6	-10.814
1.7	-6.116
1.8	-8.006
1.9	-12.884
2	-12.75
2.1	-13.954
2.2	-13.096
2.3	-18.876

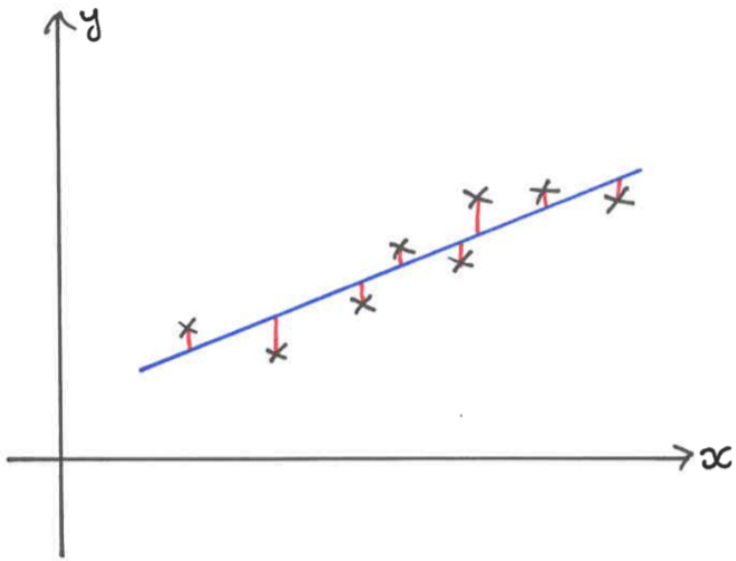




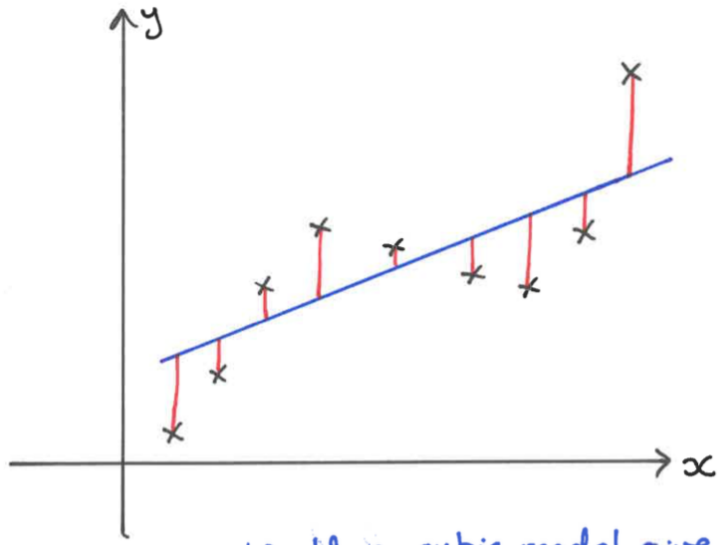
With the trendline tool we can specify fitting a 2nd-order polynomial (a quadratic function). The best such function is $y = 0.6239x^2 - 8.3459x + 1.9344$

Which model to choose?

In harder cases, we could choose several different models and fit the best parameter choices in each case.



$y = f(x) = mx + c$



Would a cubic model give a better fit?

How would we know which model described the data best by giving the closest fit?

R^2

To quantify the "goodness of fit" for each model, we can calculate the R^2 value, also called the coefficient of determination.

Calculating R^2

For a set of N data points (x_i, y_i), to which a model is fitted given by $y = f(x)$, we calculate R^2 using:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where: $SS_{res} = \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2$ and $SS_{tot} = \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2$

Interpreting R^2

Interpreting R^2

If R^2 is equal to 1, it means that the curve fits the data perfectly.

A smaller value (nearer to zero) indicates a poorer fit.

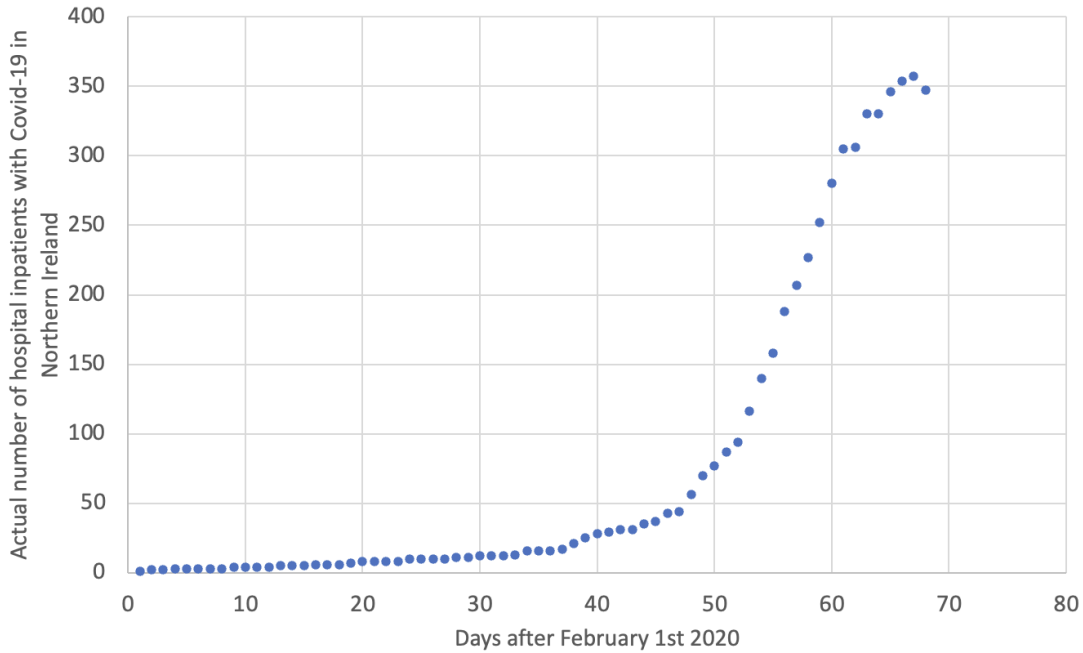
Given a data set, we can create a scatter plot, and undertake a curve-fitting procedure for each reasonable model to find the *best version of that model*. Then, compare the resulting R^2-values and determine which was the best overall best.

EXCEL's ability for curve fitting has limitations. Only certain simple functions can be fitted, and some cannot be fitted if there are zeros or negative values in the data.

Application: modelling the early spread of COVID-19

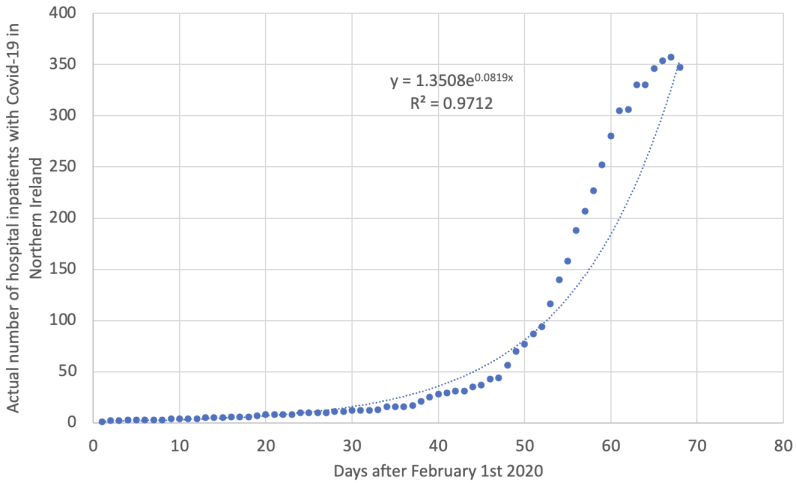
If we could use curve fitting to accurately fit a model to the data documenting the spread of coronavirus in the UK as it emerges, we may be able to estimate demand for healthcare and where and when to allocate resources.

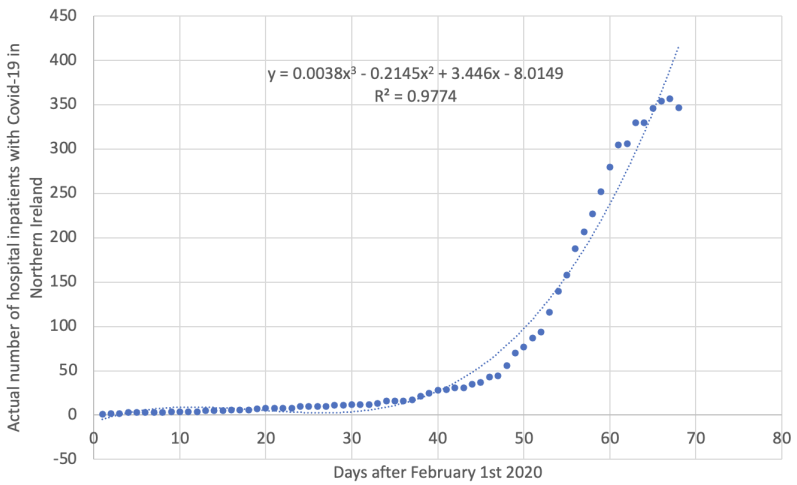
Consider this data, which shows the number of people in hospital in Northern Ireland with Covid-19 between February 1st and April 7th 2020.



Application: modelling the early spread of COVID-19

From the options available, the most reasonable (without using a very high-order polynomial) seem to be an exponential or a cubic function. They both fit the data very well ($R^2 \approx 0.97$).





Should we use these models to forecast hospitalisations in: (a) one week; (b) four months; (c) two years after the final datapoint?