Curve fitting and Correlation

Dr Gavin M Abernethy

Contents

Today we will cover...

- Curve fitting using SOLVER in Excel.
- The notion of "goodness of fit".
- R^2 correlation coefficients.
- An application to modelling the spread of coronavirus and estimating the R_0 reproduction value of the virus in the UK.

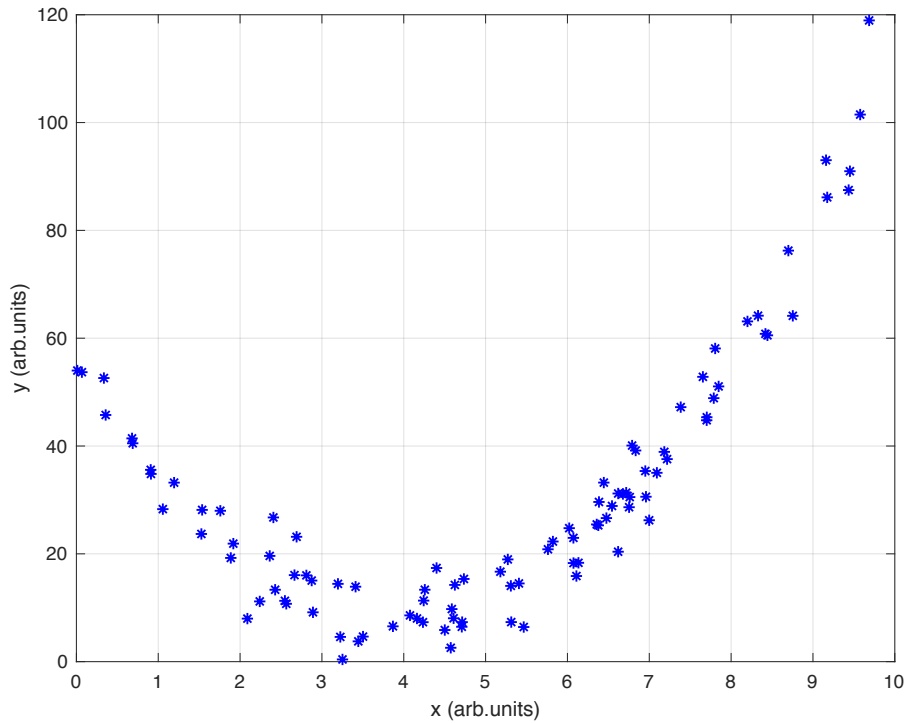
Curve fitting

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: data set

Suppose we have the set of data shown below:



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?

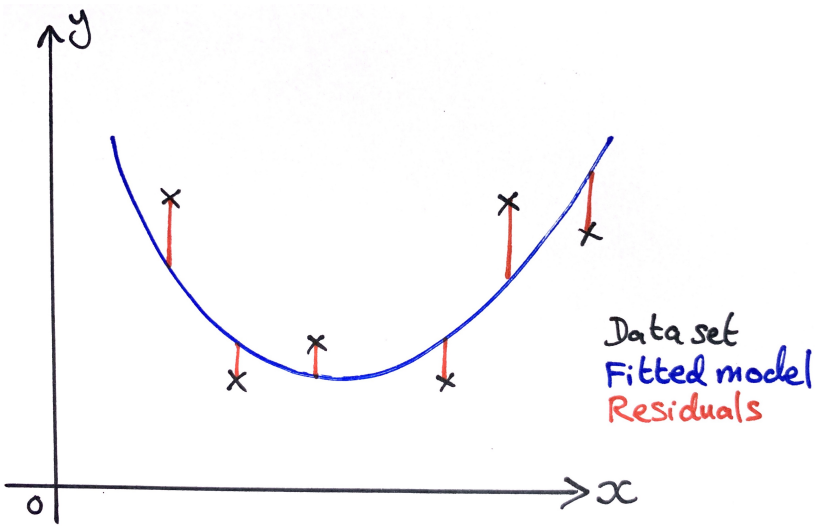
Curve fitting

The procedure for curve fitting, broadly speaking, is:

1. Select a function (called the "model") that you think might describe the data.
2. Choose some starting parameters (e.g. values for a, b and c in the previous quadratic function).
3. Use software (such as EXCEL or MATLAB) to adjust the parameters (a, b and c) until we find "good" values that yield the best fit.

Residuals

But how do we measure whether a model has a "good" fit?



The "residuals" are the differences between each actually-observed y-value and those predicted by our model with the current choice of parameter values.

The software tries many parameter choices, seeking the model (blue curve) that result in the smallest sum of the squared residuals (red distances b/w the model and the black data points).

Procedure: Curve Fitting in Excel

We can fit a curve in Excel using a built-in feature called SOLVER. Depending on your version of Excel, this may need to be enabled first:

- On PC, go to File>Options>Add-ins, and select Solver.
- For OSX, go to Tools>Excel Add-ins and again select Solver.

In either case, it should now appear at the far right hand side of the Data tab.

Procedure: Curve Fitting in Excel

Part I

1. Begin with columns containing the observed (actual) set of data, one column each with x-values and one with y-values.
2. Input some estimate parameter values in cells away from the columns of data.
3. Create a column of y-values predicted by the model, that reference the parameter value cells.
4. Create a column of the squares of the differences between the actual y-values and those estimated by the model. These are the "square residuals".
5. Sum all of these in a single cell.

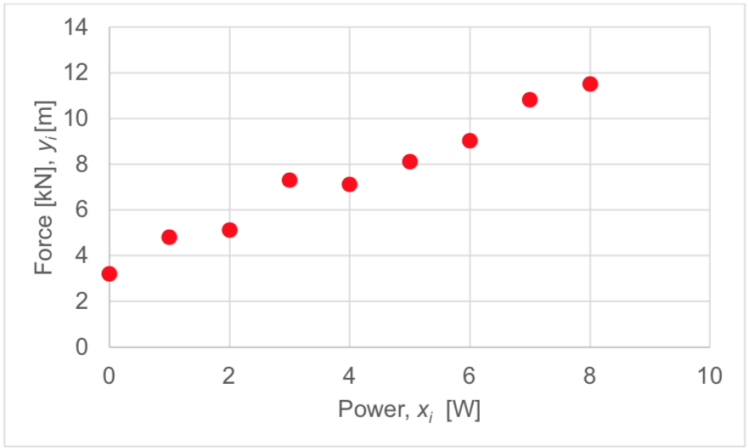
Procedure: Curve Fitting in Excel

Part II

6. Open SOLVER in the Data tab.
7. We want to **minimise** the sum of the square residuals.
8. By changing the variable cells containing the values of the parameters.
9. Ensure that "Make unconstrained variables non-negative" is NOT checked.
10. Click "Solve"!

Example (I/IV)

We will now demonstrate this procedure. The graph shows the force exerted, y_i, by a mechanical press as a function of its operating power, x_i.



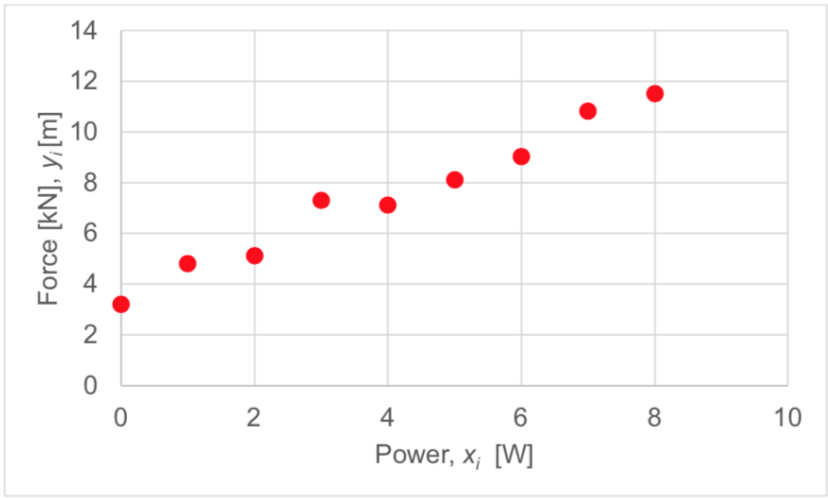
This data looks like a straight line, so we will try to fit the general linear function:

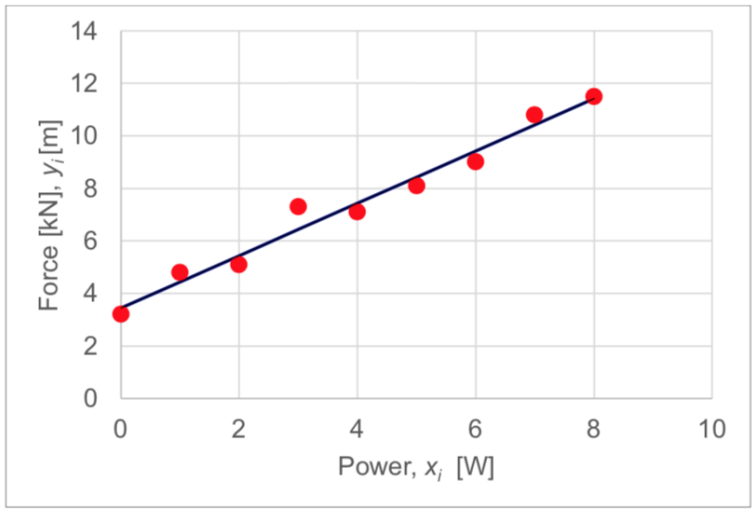
$$f(x) = mx + c$$

where m and c are our two unknown parameters that decide the gradient and intercept of the straight line.

Example (II/IV)

We will use the Excel workbook
`MMADLectureEx_CurveFittingAndRSquared_PART1.xlsx`

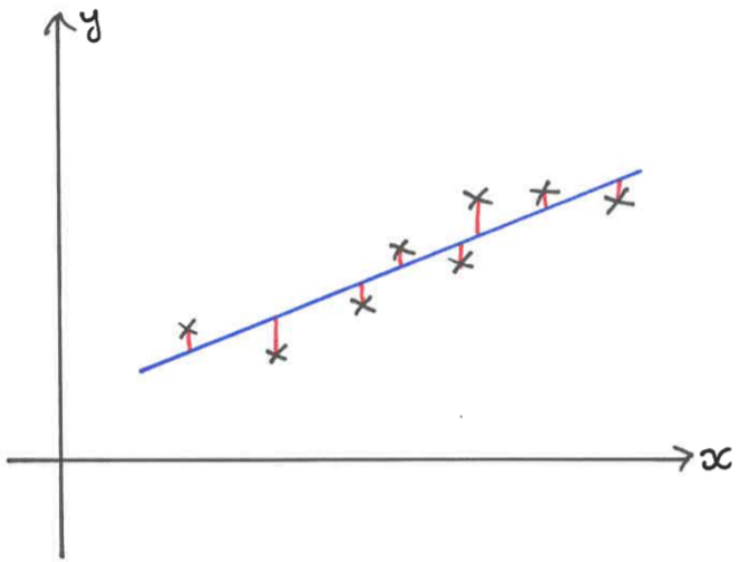


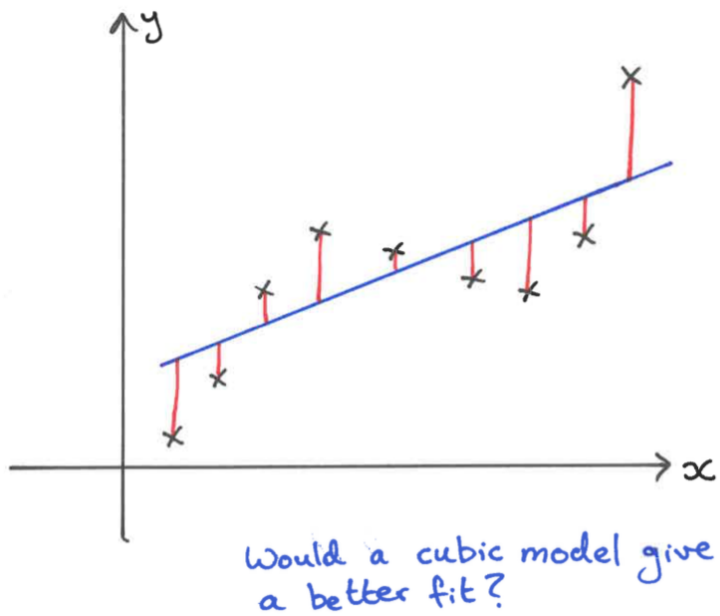


SOLVER returns fitted values of $m = 0.997$ and $c = 3.446$.

Which model to choose?

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





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 $SS_{tot} = \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2$

Notice that SS_{res} is the sum of the square residuals that we had to calculate and minimise during curve fitting anyway!

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 means a poorer fit.

The fraction

$$\frac{SS_{res}}{SS_{tot}}$$

scales the relative size of the residuals appropriately. Essentially, it means that the greater the variance in the y-values of the data set (i.e. the larger the value of SS_{tot}, the larger the squared residuals (the error between the data and the model) can be whilst still being considered a "good" fit.

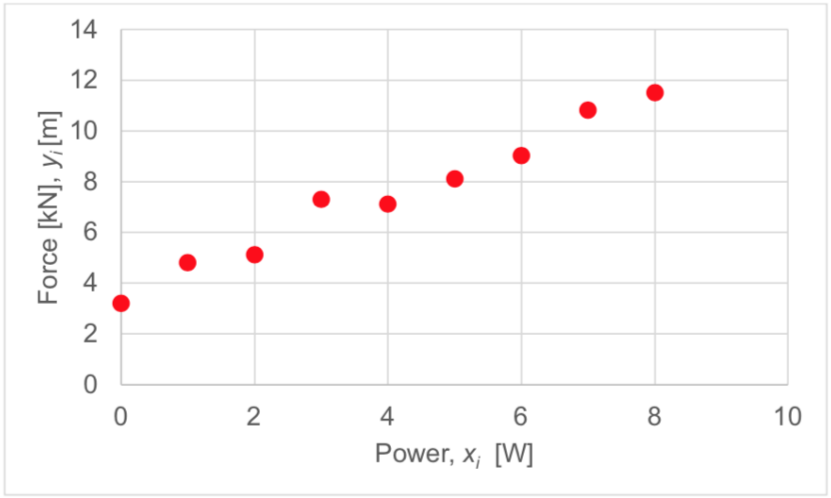
Which model to choose?

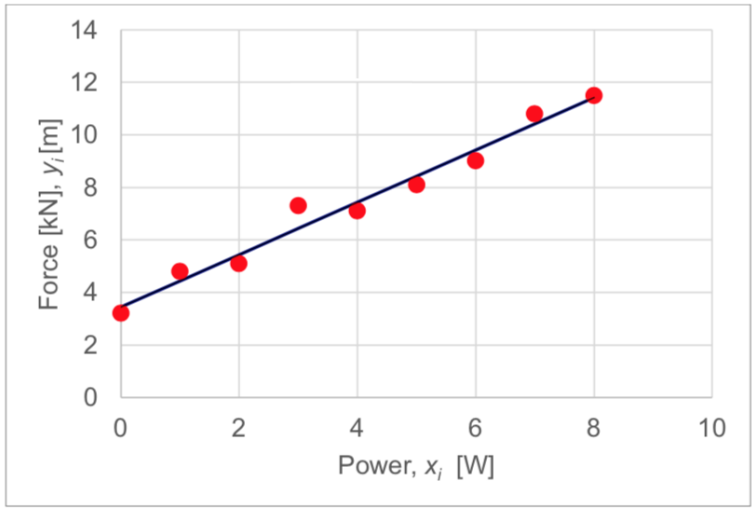
So when we are presented with a data set, we might have an idea from the context of what kind of model to fit (for example, in the tutorial questions for this week we might know that the data is taken from a probability distribution that might be roughly normal or log-normal).

Otherwise, we could create a scatter plot of the data, and make an educated guess from the plot of what model(s) might be suitable. Then we would undertake a curve-fitting procedure in each case to find the *best version of each model*, and finally compare the resulting R^2-value to evaluate the best version of each model and determine which one produced the overall best result.

Example (III/IV)

Let's return to our mechanical press example.





We have used Solver to fit the linear function:

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

and found "best fit" parameter values of $m = 0.997$ and $c = 3.446$.

Example (IV/IV)

Clearly, this straight line appears to fit the data well. We can confirm this by finding the R^2 value.

> We will use the Excel workbook
> `MMADLectureEx_CurveFittingAndRSquared_PART2.xlsx`

And we find a value of $R^2 = 0.9741$, indicating a close fit.

Note: for simple built-in functions (linear, exponential, logarithmic, quadratic), Excel can automatically fit a model and calculate R^2. In this case, we can use this to check our results by right-clicking on the graph and selecting "add trendline".

MMADLectureEx_CurveFittingAndRSquared_PART3.xlsx

	A	B	C	D	E	F
1	**Power, x_i [W]**	**Force, y_i [kN]**	**Model [kN]**	**Square Displacement [kN^2]**		**y_i-mean(y_i)**
2	0	3.2	=$F$13*A2+$F$14	=(C2-B2)^2		=(B2-$B$11)^2
3	1	4.8	=$F$13*A3+$F$14	=(C3-B3)^2		=(B3-$B$11)^2
4	2	5.1	=$F$13*A4+$F$14	=(C4-B4)^2		=(B4-$B$11)^2
5	3	7.3	=$F$13*A5+$F$14	=(C5-B5)^2		=(B5-$B$11)^2
6	4	7.1	=$F$13*A6+$F$14	=(C6-B6)^2		=(B6-$B$11)^2
7	5	8.1	=$F$13*A7+$F$14	=(C7-B7)^2		=(B7-$B$11)^2
8	6	9.0123	=$F$13*A8+$F$14	=(C8-B8)^2		=(B8-$B$11)^2
9	7	10.8	=$F$13*A9+$F$14	=(C9-B9)^2		=(B9-$B$11)^2
10	8	11.5	=$F$13*A10+$F$14	=(C10-B10)^2		=(B10-$B$11)^2
11	**Mean of Data:**	=AVERAGE(B2:B10)	**Sum of Square Residuals:**	=SUM(D2:D10)	**Sum of Squares:**	=SUM(F2:F10)
12					The parameters to optimise:	
13					**m**	0.997076564970542
14					**c**	3.44639300594564
15			**Sum of Square Residuals (SSres)**	=D11		
16			**Sum of Squares (SStot)**	=F11		
17			**R^2**	=1-D15/D16		

1. Create a third column containing the y-values predicted by the model (in this case $y = mx + c$).
2. Use SOLVER to obtain the values of the parameters m and c that minimise the sum of the square residuals SS_{res}.
3. Then calculate the sum of the squares SS_{tot}.
4. Finally calculate the R^2 value.

Application: modelling the spread of COVID-19 in the UK

A common technique in mathematical epidemiology involves dividing a population into "compartments": groups such as people who have had the disease and recovered, those who currently have it, etc. Advanced models may break down the population further by factors such as age or geographic location.

If we could use curve fitting to fit such a model to the data documenting the spread of coronavirus in the UK, we would be able to estimate factors including the R_0 "reproduction value" that was guiding UK government policy during the height of the pandemic.

Application: modelling the spread of COVID-19 in the UK

The simplest epidemic model is the **SIR** (susceptible - infected - recovered) model. This separates your population into three groups:

1. The **Susceptible** group S, have not yet contracted the disease, but they may catch it in the future.
2. The **Infectious** group I, are those who currently have the disease and are capable of transmitting it to others.
3. The **Recovered** group R, are those who previously had the disease and have now recovered or passed away.

Application: modelling the spread of COVID-19 in the UK

You can then construct three ODE's (differential equations) describing how the number of individuals in these three groups changes, according to two parameters:

- **Infection rate** β controls how rapidly infectious individuals cause susceptible individuals to catch the disease (so S goes down and I goes up).
- **Recovery rate** γ controls how long it takes for infectious individuals to recover (so I decreases and R increases).

Application: modelling the spread of COVID-19 in the UK

We may then obtain a numerical solution to the SIR equations using the Euler method:

$$t_{n+1} = t_n + 1$$

$$S_{n+1} = S_n - \frac{\beta}{N} S_n I_n$$

$$I_{n+1} = I_n + \frac{\beta}{N} S_n I_n - \gamma I_n$$

$$R_{n+1} = R_n + \gamma I_n$$

where N is the population of the UK and t_n is the number of days after the initial date of January 31st 2020.

Application: modelling the spread of COVID-19 in the UK

In an SIR model, the basic reproductive number R_0 that you may have heard about on the news is given by:

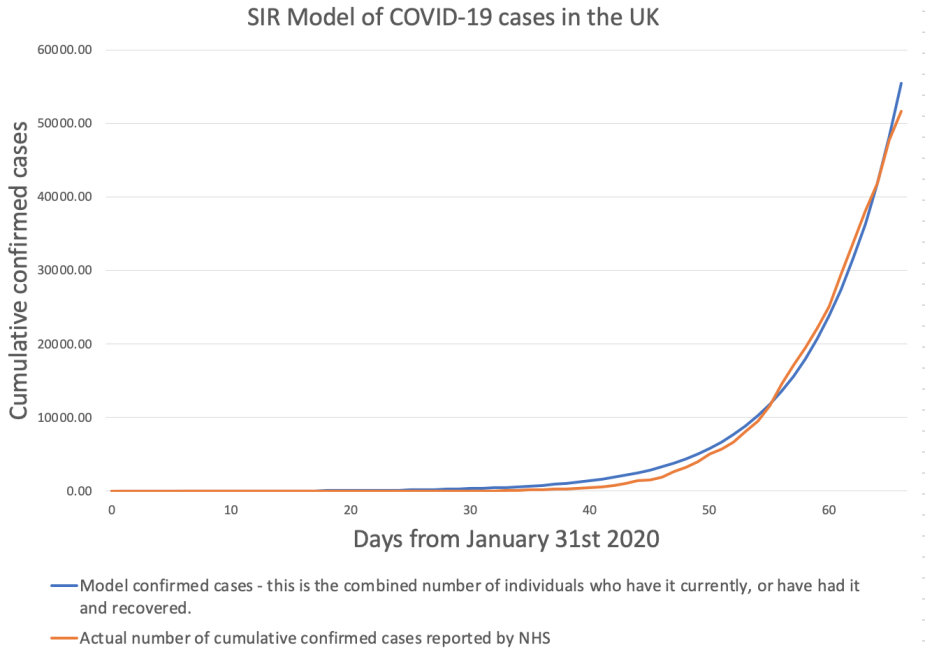
$$R_0 = \frac{\beta}{\gamma}$$

This value describes how rapidly the disease grows. If it is greater than 1 then the disease is transmitting faster than people are recovering, so it will spread.

This value was reported to be the basis of UK government policy, so based on the daily number of confirmed cases of COVID-19 in the UK between January 31st 2020 and April 7th 2020, let's try to fit an SIR model, determine the best values of β and γ and thus calculate an estimate of R_0

Application: modelling the spread of COVID-19 in the UK

In the Excel spreadsheet, `SIR_model_of_the_UK.xlsx`, we have used SOLVER to vary the values of β and γ and fit the daily confirmed cases announced by the NHS to the combination of I (currently infected) and R (previously infected and now recovered) in the model.



You can see that we can get a close fit with a predicted R_0 value of about 1.65. Do you think this estimate is reliable? (Consider that only hospitalised cases were likely to be tested at that time.)