**A9Wk Introduction to non-linear regression**

In many situations, we will find that the relationship between the dependent and independent variable is not necessarily linear but a non-linear. In this section, we shall introduce the concept of **non-linear regression** via a simple example.

Before we start, we need to introduce a series of non-linear relationships between the variable y and x and their governing equations. Some of the most popular curves that describe the shape of these relationships are presented in Figures 1 to 7.

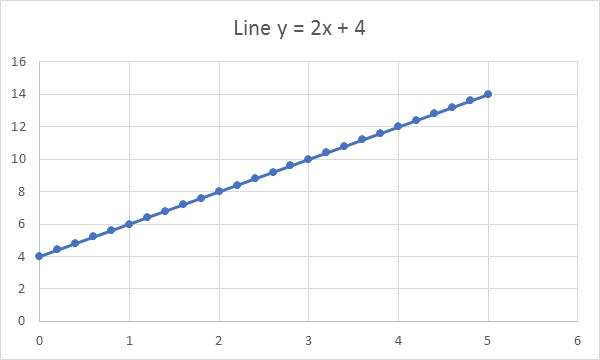


Figure 1 Line y = b0 + b1x

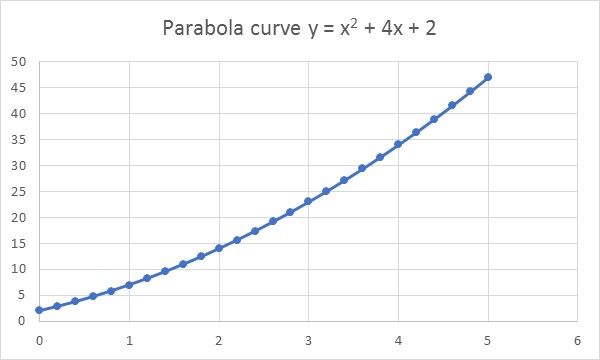


Figure 2 Parabola curve y = b2x2 + b1x + b0

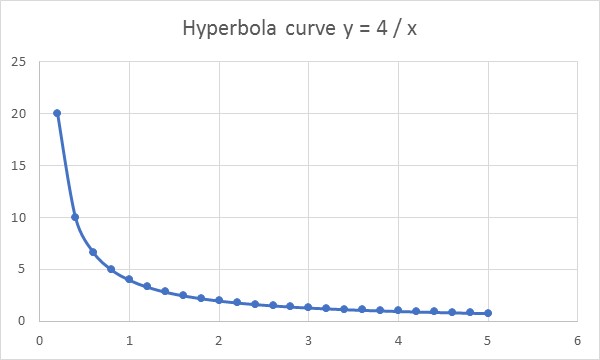


Figure 3 Hyperbola curve y = b0/x

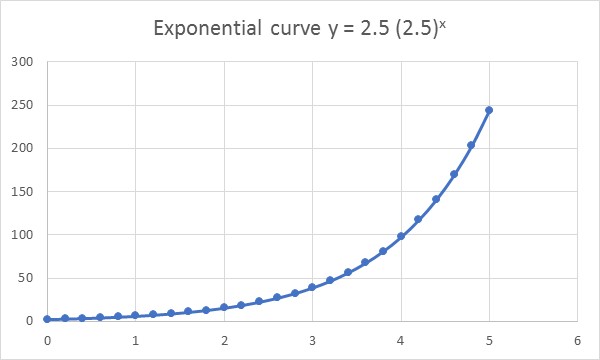


Figure 4 Exponential curve y = b0b1x

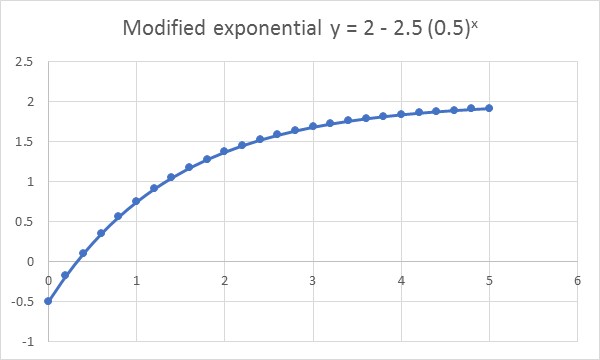


Figure 5 Modified exponential curve y = b2 + b0b1x

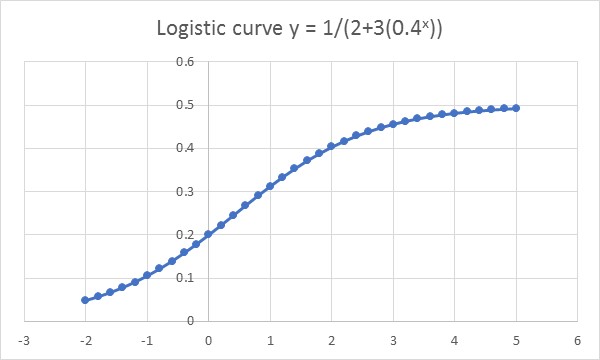


Figure 6 Logistic curve 1/y = b2 + b0b1x

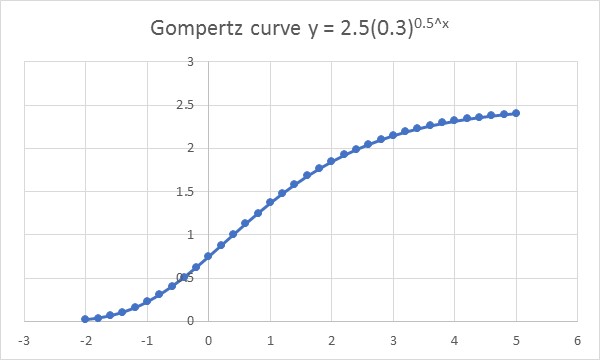


Figure 7 Gompertz curve 

Let’s look at just one of these non-linear relationships. Equation (1) represents the equation of a parabola (or polynomial of degree 2) from figure 2:

(1)

The values of the parameters b0, b1 and b2 can be determined using least squares regression by solving equations (1) – (5).

(2)

(3)

(4)

Where  (5)

Note that to fit a curve to a scatter plot using Excel is quite straightforward. Right click on a data point in the scatter plot and choose the ‘Add Trendline’ option and select the curve you would like to fit.

For example, if we wanted to fit a polynomial of order 2 to the scatterplot then we would choose the Polynomial option and select order 2 as illustrated in Figure 8.

The general equation of a polynomial of order 2 would be: Y = b0 + b1x + b2x2.

Finally, you can ask the Format Trendline > Trendline Options menu to include this equation on the scatterplot together with the value of the coefficient of determination (R-squared). Just click the two boxes at the bottom, left-hand side of the dialogue box.

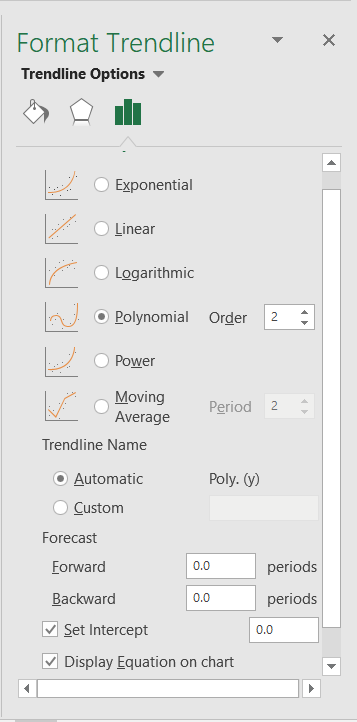


Figure 8

We will use Excel to show how to fit this curve to a data set, calculate the equation of the line, and calculate the coefficient of determination. The data set is not shown here, just the principle of how to use Excel for this purpose.

**Example**

Table 1 provides the sales and price data collected from a range of discount stores selling a particular product but using their own discount policy to price the product. The question is: can we fit an appropriate relationship to predict sales given price?

|  |  |
| --- | --- |
| Price, x | Sales (£000’s), y |
| 0.30 | 100.00 |
| 0.40 | 95.00 |
| 0.50 | 93.00 |
| 0.58 | 90.20 |
| 0.60 | 90.00 |
| 0.65 | 88.00 |
| 0.70 | 85.00 |
| 1.10 | 86.00 |
| 1.15 | 83.00 |
| 1.40 | 82.00 |
| 1.80 | 80.00 |
| 2.60 | 81.00 |

Table 1

The solution to this problem consists of identifying the type of relationship between the two variables. Figure 9 illustrates graphically the relationship between sales and price which is a scatter plot for sales (y) against price (x) illustrating a possible non-linear relationship between the variables y and x.

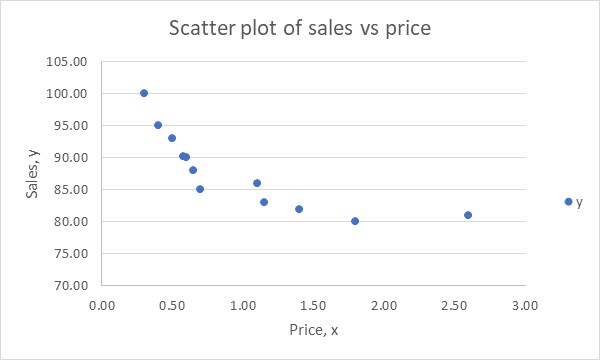


Figure 9

From Figure 9 we may suggest that the relationship between the two variables is given by model 1 or model 2.

Model 1: Line fit y = b0 + b1x

Model 2: Curve fit y = b0 + b1/x

If the relationship was non-linear we may still use linear regression as long as we are able to transform the non-linear data to a linear form. The parameters b0 and b1 in model 1 and model 2 can be estimated using the methods described in previous sections and we will use the Data Analysis > Regression method to calculate these values (include residual plot and normal probability plot). Figure 8.32 and 8.33 presents the ANOVA results for both models.

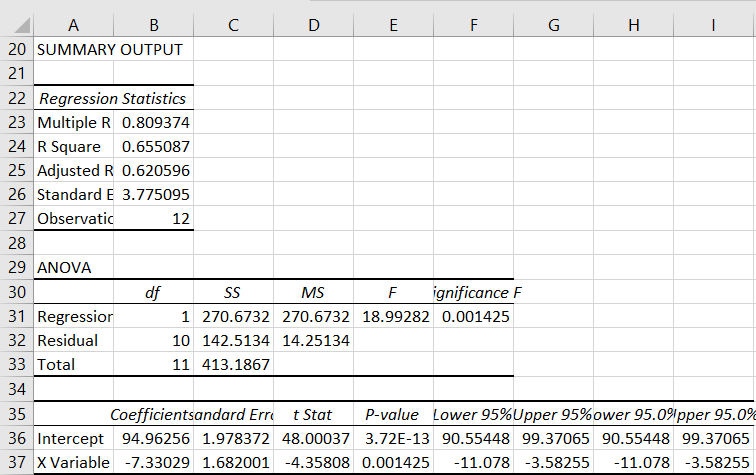


Figure 10 Regression ANOVA table for model 1: y = b0 + b1x

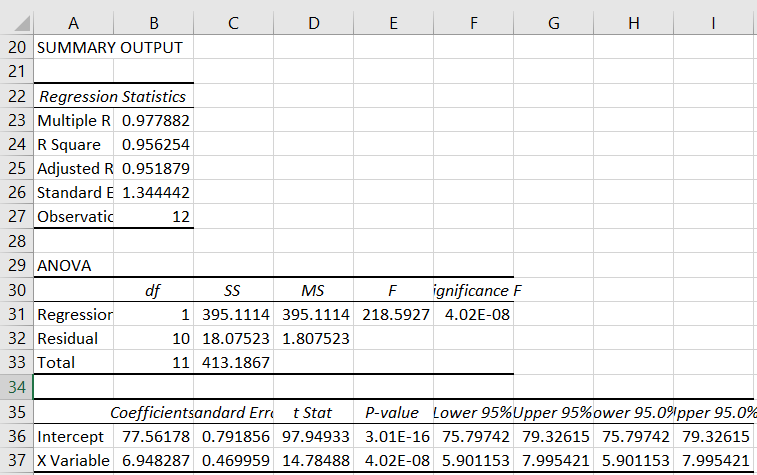


Figure 11 Regression ANOVA table for model 2: y = b0 + b1/x

Table 2 shows the results of applying least squares regression for model 1 and 2. The results show that model 2 represents a better fit to the data set than model 1.

|  |  |  |
| --- | --- | --- |
| Model | Equation | COD |
| 1 |  | 0.66 |
| 2 |  | 0.96 |

Table 2

**>>**

From Table 2, we can see that for the non-linear model 96% of the variations in one variable are explained by variations in another, whilst for the linear model only 66% of variations are explained by the model. Clearly, we are better off using the non-linear model: .

Note that in model 1 the regression is fitted to variable y and variable x and in model 2 the x variable has been transformed to 1/x and the regression is fitted to variable y and variable 1/x.

To complete the solution, you would then need to analyse the model 2 ANOVA table results to check on whether or not the model 2 parameter terms (b0, b1) are significant contributors to the value of the independent variable (y). This would be done using the Student’s t-test (or F-test).

From Figure 11, the two parameter values b0 and b1 are significant contributors to the value of the y variable (p = 3×10-13<0.05 for b0 and p = 4×10-8<0.05 for b1). The final step in the analysis process is to check the model assumptions. From the Data > Data Analysis > Regression results we requested the residual and normal probability plots. Figures 12 to 13 compare the results for model 1 and 2.

Model 1 

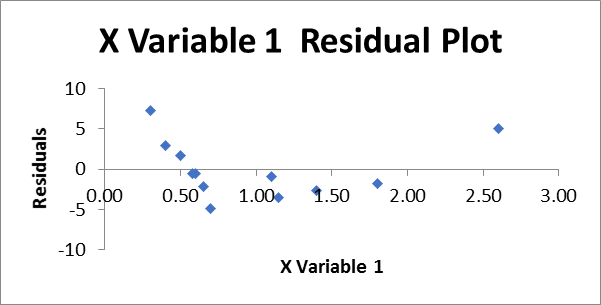


Figure 12

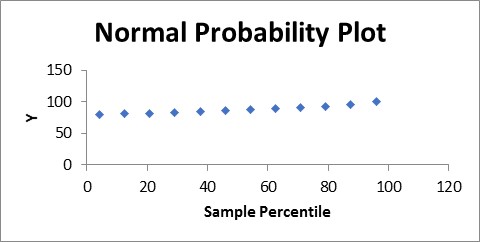


Figure 13

Model 2 

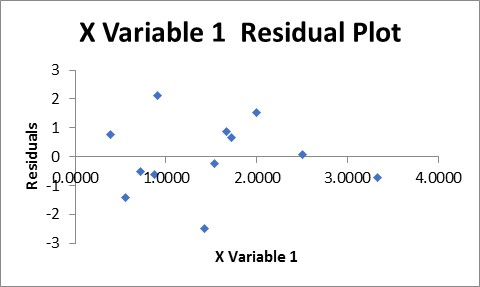


Figure 14

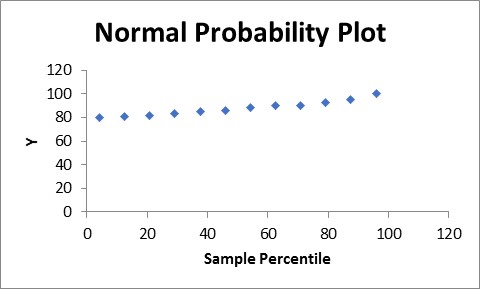


Figure 157

Figure 8.12, unlike Figure 14, shows some pattern, indicating that model 2 is better suited to model this relationship. Model 1 effectively violates the linearity assumption, but model 2 does not. It is, therefore, better suited to this data set.

We can also see from Figure 12 that the variance might to be growing in model 1. If this was the case, it violates the constant variance of errors assumption, again confirming better suitability of model 2.

On the other hand, Figure 12 and Figure 14, which are the normal probability plots for model 1 and 2, indicate that neither of the two models violate the normality of errors assumption. However, due to the other two violations provided above, we maintain that model 2 is the appropriate one for this relationship.