CHAPTER 13: UNDERSTANDING RELATIONSHIPS

1 Making connections

When we use quantitative data we are often seeking to demonstrate that there is a link between one set of data and other. We might want to investigate what effect a major historical event had on the price of food or whether married men more use more words on a daily basis than their wives.

In Table 1 we have data about the price of wheat and the price of oats between 1830 and 1839. The prices were originally in pounds and shillings, but have been decimalised here for clarity. What is the relationship between wheat prices and oats prices. When wheat prices are high are oat prices high too? From the data alone is difficult to see for sure.

On the graph in Figure 1 we have plotted the wheat price against the barley price for each year. We are see that there is a pattern of sorts as the numbers plots sort of line up. But how do we describe this pattern in more detail? One way is to calculate the Pearson product-moment correlation coefficient of the data. The Pearson product-moment correlation coefficient is a number between 1 and -1. This number is referred to as r or `Pearson's r'.

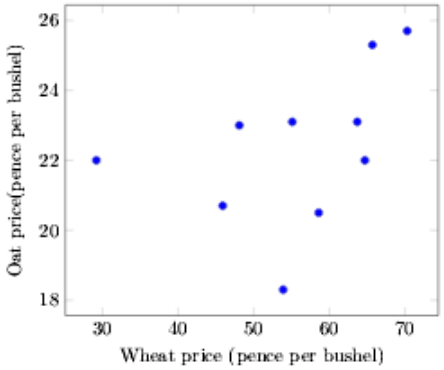
2 Calculating the Pearson product-moment correlation coefficient

The method used to calculate the correlation coefficient [1])There are other correlation coefficients which are not Pearson product-moment correlation coefficients, e.g. The Spearman's rank). It is useful to create a scatterplot so that you have a rough idea of what the correlation co-efficient might be. You will also be able to identify outliers, that is an observation which does not seem to fit the overall pattern. Additionally you may find that your

Table 1: Wheat and oats prices

Year	Wheat	Oats
1830	63.7	23.1
1831	65.7	25.3
1832	58.6	20.5
1833	53.9	18.3
1834	45.9	20.7
1835	29.2	22
1836	48.1	23
1837	55.1	23.1
1838	64.7	22.4
1839	70.3	25.7

Figure 1: Wheat and oats prices, 1830-1839



relationship is U-shaped or S shaped in which case there may be a relationship but the calculation of r will not reveal this. See Anscombe's Quartet for more about this

STAGE 1: Notice that we have calculated the mean of both the wheat prices (x) and the oat prices (y).

STAGE 2: We need to take the variances of wheat (column 7) and oats (column 8) and divide them by the number of observations:

Wheat

$$\frac{1326}{10} = 132.6$$

Oats

$$\frac{44.1}{10} = 4.41$$

STAGE 3: We then take the square roots to find the Standard Deviations

Wheat: = √218.8=11.52

Oats: = √44.1=2.10

We now have the four numbers we need to calculate r.

1. The SD of the wheat: 11.52

r = correlation coefficient. This will be number between -1 and +1.

x = the price of wheat

y = the price of oats

n = number of observations. In this case n=10 because there are ten pairs-- wheat and oat prices were observed each year.

SDx The standard deviation of the wheat

Sdy The standard deviation of oats

2. The SD of the oats: 2.10
3. The number of observations: 10 (note that this is 10 and not 20 as there are ten years of paired data).
4. The Sum of Column 6: 107.7

STAGE 4

We now have a value of r of 0.442. Your value of r will be between -1 and +1. If you have a value of r which is more than 1 or less than -1 you have made an error in your calculation. The value of the r helps us

$$r = \frac{\frac{1}{n}(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + \ldots (x_n - \overline{x})(y_n - \overline{y})}{SD_x SD_y}$$

Table 2: Table to calculate correlation coefficient

Year	x	x²	y	y²	xy
1830	63.7	4057.69	23.1	533.61	1471.47
1831	65.7	4316.49	25.3	640.09	1662.21
1832	58.6	3433.96	20.5	420.25	1201.3
1833	53.9	2905.21	18.3	334.89	986.37
1834	45.9	2106.81	20.7	428.49	950.13
1835	29.2	852.64	22	484	642.4
1836	48.1	2313.61	23	529	1106.3
1837	55.1	3036.01	23.1	533.61	1272.81
1838	64.7	4186.09	22.4	501.76	1449.28
1839	70.3	4942.09	25.7	660.49	1806.71
Sum of columns	ΣX= 555.2	ΣX²=32150.6	ΣY= 224.1	ΣY²=5066.19	ΣXY=12548.9

Figure 2 Karl Pearson 1857-1936. Pearson developed numerous statistical tests including the Pearson Product Moment Correlation Coefficient

to make a judgement about the strength of the association between wheat prices and oat prices.

Table 3 Interpreting value of r

2.2 Interpreting r

Notice that I am careful to use the word association rather than relationship. What we are looking for is to find the nature of the cause and effect in any association we might observe. It seems reasonable to conclude that there is a moderate association between wheat prices and oat prices. Table 3 is a useful guide to interpreting the r value.

2.3 Things to remember

Fortunately you can use a spreadsheet or a statistical package to work out the correlation co-efficient for a large amount of data.

The Correlation co-efficient is a bivariate test. This means that each observation has two parts. In this case each year has two (a pair of) prices -the price for wheat and the price for oats.

Very importantly the correlation measures association and not causation. We have been able to demonstrate a moderately weak association between wheat prices and oat prices but we cannot say that a rise in oat

r		How the association between would work between wheat prices and oat prices.
1	Perfect positive association	If r =1 Wheat and oat prices go up and down together in the same direction.
0.8	Strong positive association	
0.6		
0.4	Moderately weak positive association	The association between oat and wheat prices is r = 0.442
0.2		
0	No association	If r= 0 There is no association between wheat prices and oat prices.
-0.2		
-0.4	Moderately weak negative association	
-0.6		
-0.8	Strong negative association	
-1	Perfect negative association	If r = -1 Wheat prices go up when oat prices go down and oat prices go up when wheat prices go down.

prices is caused by a rise in wheat prices based on the correlation co-efficient alone. We will be addressing the issues of causation in Chapter 18.

3 Exercises

Calculate the correlation coefficient for the UK price and US price for the best-selling books in Table 4 .

Calculate the correlation coefficient for the distance and times taken in days. [2]

4 References

[1]The Pearson product-moment correlation coefficient is often known simply as the `Correlation coefficient' and is referred to as such in this chapter.

[2]This data comes from From Pliny the Elder (AD 23-AD 79), cited in Lionel Casson (1951) Speed of the sail of ancient ships Transactions of the American Philological Association 82, pp. 136-148

Table 4 Comparison of US and UK prices of best-selling fiction on major retailer's website

Title	Author	UK price	US price (converted to £)
Lover reborn	JR Ward	8.4	13.4
I've got you number	Sophie Kinsella	8	12.8
Betrayal	Danielle Steel	8	12.8
The Patchwork Heart	Jane Green	8.4	13.5
Lone Wolf	Jodi Picoult	9.2	14.7
The Thief	Clive Cussler	9.6	15.4
Death comes to Pemberly	P D James	10.3	16.5

Table 5 Distance and voyage length in the Roman Empire

Voyage	Distance (Nautical Miles)	Length of Voyage
Ostia-Africa	270	2
Messina-Alexandria	830	6
Ostia-Gibraltar	935	7
Ostia-Hispania Citerior	510	4
Messina-Alexandria	830	7
Ostia-Provincia Narbonensis	345	3
Puteoli-Alexandria	1000	9

CHAPTER 14: Predicting new observations from known data

1 Introduction

When it gets cold outside I turn on the central heating in my house. If I am using the heating then I am using more gas. Therefore there is a relationship between the temperature outside and the amount of gas I use. Suppose that each day I collect two sets of data: 1) the outside temperature and 2) the amount of gas I use. Over a period of time I will be able to predict the amount of gas I use just by taking the temperature.

Notice that this only works one way. A change the outdoor temperature affects the amount of gas I use, but using more or less gas will not increase or decrease the outside temperature.

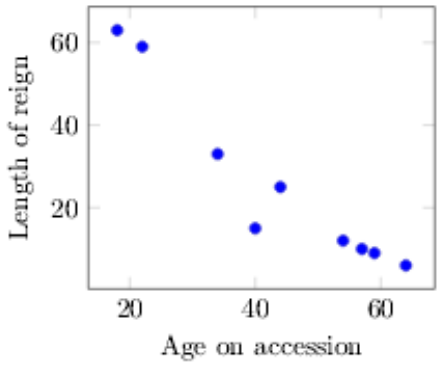


Figure 1 Relationship between age on accession and length of reign

2 Predicting reign length of British monarchs

shows the age that British monarchs were when they came to the throne and how long they reigned for. It would be reasonable to suppose that monarchs who came to the throne younger would have reigned longer and this seems to be the case. Only monarchs since George I in 1714 are included as no kings or queens since this time have been deposed or died violently. Edward VIII is excluded as he abdicated his throne and the present Queen Elizabeth II is excluded on the grounds that she is still alive.

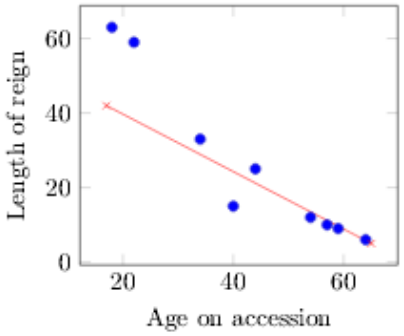
Table 1 Reign of British Monarchs 1714-1952

'	Age on Accession to throne	Reign
George I	54	12
George II	34	33
George III	22	59
George IV	57	10
William IV	64	6
Victoria	18	63
Edward VII	59	9
George V	44	25
George VI	40	15

'

If we plot these figures onto a scatter graph (Figure 1) we can see the relationship between age on accession and reign is negatively correlated. The older a person becomes king or queen the shorter their reign.

Figure 2 Relationship between age on accession and length of reign with approximate `good fit' line drawn by eye



We can see the rough pattern of the dots and draw a 'good fit' line. I have drawn such a line to create Figure 2.

3 Prediction

We can use the line to make a *prediction* about how long a monarch will reign if the only information we have is their age. We can do this by finding the age on the *x* axis, finding where it intersects the best-fit line, and reading off answer off the *y* axis. For example, We can see that a monarch who becomes king or queen at the age of 50 can expect to reign for 20 years. So by looking at a known pattern of data we can predict the value of one variable simply by knowing the value of one other variable. To use another example, if I have lots of data on the relationship between exam performance at school and exam performance at university I can use this data to predict the university exam performance of an individual student, simply by knowing about their performance at school.

4 Simple linear regression

I have drawn the line in Figure 2 by hand. If you were to perform this exercise you might put the line in a slightly different place. And because the line is in a different place you will get a slightly different answer when you made your prediction about the length of time a monarch will reign for. This section introduces a technique called **simple linear regression.** This technique is called simple linear regression to distinguish it from other forms of regression analysis. This is the only form of regression covered in this book. The simple regression analysis uses a mathematical technique to predict instead of trying to predict by eye.

5 Finding the regression line

The regression line is not a number, but a simple equation which describes the line of the best fit. It is most easily done using computer software, but it is useful to work through it manually to see how it works. Importantly it essential to understand the following:

We have designated age of accession to the throne as *x* and the length of reign as *y*. This is not a co-incidence. X is the information we have and Y is what we are trying to predict. It is important to get these the right way round. If we were looking at predicting how much electricity we are going use today based on our observation of the weather, the weather is our *x* variable and the gas bill (which we are trying to predict) is the *y*.

We will use this formula for simple linear regression.

$$y=mx+b$$

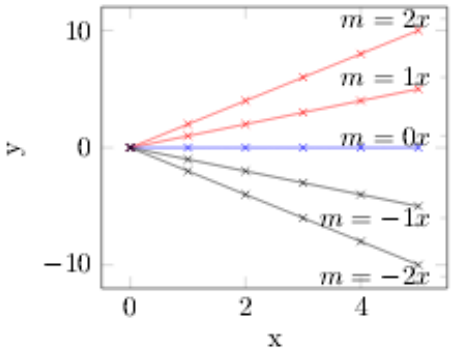
y is what we wish to find out. In the this case, the length of time a monarch reigned for.
m is the **gradient** of the slope. The gradient is simply how steep a slope is. *m*=2 is steeper than *m*=1. If *m*=0 there is no slope. If *m* is a minus number then the slope is negative. You may find Figure 3 a helpful illustration

You may sometimes find the formula presented with different letter, e.g. $y=ax+b$ or with Greek letters Alpha *a* and Beta *β*. *e.g.* $y=ax+\beta$

b is the **intercept**. This is the value of *y* when x is zero. In this case this is how long we would expect a monarch to reign if they came to the throne at the age of zero. You may find that Figure 4 is a helpful illustration of this.

x is our predictor variable. This is the variable we are using to explain *y*. This is why the equation takes the form $y=mx+b$. You may find that Figure 5 is a helpful illustration of this. In order to find the value of *y* we need to find *m* and *b*.
The best way to start is by drawing a table (see Table 2). It will give us the numbers we need to calculate *y* and *m*.

Figure 3Various values of mx



Firstly to find the gradient of the slope *m*

$$m=\frac{n\Sigma(xy)-\Sigma x\Sigma y}{n\Sigma(x^2)-(\Sigma x)^2}$$

Figure 4 Various values of b(the intercept)

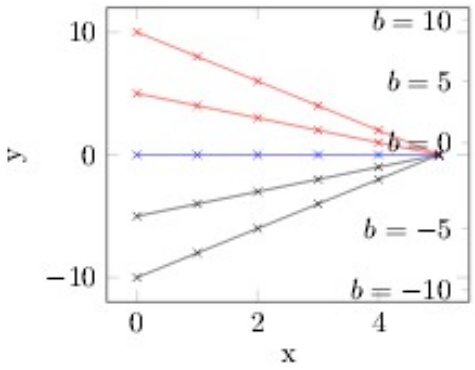
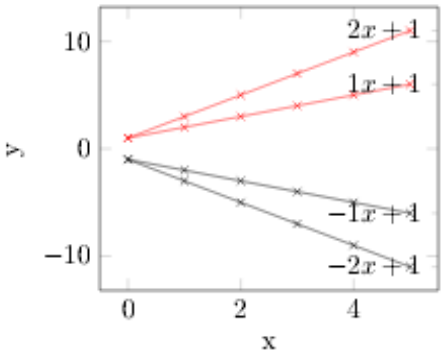


Figure 5Various values of $mx+b$



n is the number of observations. In this case we have nine monarchs so

Table 2 Reign of British Monarchs

	Age on Accession to throne (x)	Reign (y)	xy	x2
George I	54	12	648	2916
George II	34	33	1419	1849
George III	22	59	1298	484
George IV	57	10	570	3249
William IV	64	6	384	4096
Victoria	18	63	1134	324
Edward VII	59	9	351	3481
George V	44	25	1100	1936
George VI	40	15	600	1600
Totals	Σx=401	Σy=232	Σxy=7684	Σx²=19935

Therefore:

$n=9$

$\Sigma(xy)=7684$

$\Sigma x = 401$

$\Sigma y = 232$

Σ^2

$$m = \frac{9\times 7684 - (401\times 232)}{9\times 19935 - (401\times 401)}$$

$$= \frac{-23,876}{18,614} = -1.28$$

Therefore $m = -1.28$.

Secondly we need to find the intercept. The formula for the intercept is

$$b=\Sigma y - m(\Sigma x)n$$

Using the numbers from our table this makes

$$b=(232)-(-1.28)\times(401)\times 9=232-$$
$$-153.39=745.39=82.8$$

Now we have our m and b values we are able to predict how long a king or queen will reign for when coming to the throne.

Remember:

$$y=mx+b$$

Now replace the letters with the values we have

$$\textit{Length of reign} = -1.28 \times \textit{age at accession} + 82.8$$

So how can we this data to predict future events? If someone comes to the throne age 50

$$-1.28\times50+82.8=18.8 \textit{years}$$

If at the age 20

$$-1.28\times20+82.8=57.2 \textit{years}$$

This regression line has been plotted in Figure 6.

6 The limitations of simple linear regression

The main limitation of the regression equation is that it is a generalisation derived from a lot of data points. It can shed light on a general pattern, but it will never be 100% accurate. To use the example again of predicting university exam performance from school exam performance some students will do as well as predicted, some better and some less well. A small number will do a lot better or a lot worse than predicted.

6.1 Predicting the correct way round

There are important cautions to be aware about when using any form of regression analysis. Firstly unlike the correlation analysis we looked at in Chapter 13 the regression equation only works one way. We can use the same regression equation to calculate how long we might expect a monarch to reign when we know their age on coming to the throne. In other cases it is obvious from cause and effect that the equation can only work one way round. We use our central heating more when the temperature decreases but using the central heating does not make the outside temperature decrease. In the case of Table 6" we can calculate a regression equation for the relationship between the distance from London to a particular city and the price. However it is not possible to predict the city from the price. Even if I form a regression equation to predict the distance from the price this does not tell us what direction from London the price refers to. Paris is a similar distance from London as York to the north, Amsterdam to the East and Swansea to the west.

6.2 Outliers

Secondly outliers are an important issue in regression analysis. If you are using a statistical analysis software package it will probably alert you any outliers, but they are usually easy to spot on the scatterplot like in Figure 6. Outliers are distant from the regression slope and may have a disproportionate influence on the slope. Looking back at our scatterplot you might have noticed that one of the observations is quite a bit further away from the line that the others. The observation marked in Figure 7 is that of George VI who became king at the age of 40, but reigned for just 15 years. We can read off on the graph how long we might have expected George VI to reign. If we follow age 40 up to the regression line we can see that a monarch who comes to the throne age 40 could usually expect a reign of over 30 years. How best to deal with outliers is a matter of judgement. Sometimes researchers remove outliers from the analysis in order to make their equation more reliable and increase the value of r^2.

This is sometimes appropriate, but it does need to be justified. Generally speaking we do well not to remove outliers unless we know why they are outliers or that they are sufficiently rare to be able to dismiss them as `freak' observations.

6.3 Scope

Life expectancy is increasing so using the life expectancy of eighteenth and nineteenth century kings and queens is likely to underestimate future life expectancy.

It is not advisable to use the regression equation when the x values are outside the scope of the original observations. William IV was 64 years only when he became king. The regression line `predicts' that any future monarch who came to the throne in their seventies would actually have a negative reign which is of course impossible.

7 How reliable is our equation? Calculating r^2

It is all very well to use the regression equation to predict how long a monarch might reign, but how reliable is it? Does our regression line equation explain the relationship between a monarch's age on accession and the length of the their reign, perfectly? Reasonably well? Hardly at all? In a similar way to the correlation co-efficient we can use an $r2$ calculation to find out how well the regression equation explains the relationship.

The $r2$ value is between 0 and 1. A value of 1 indicates that the regression equation explains the relationship perfectly. A value of zero indicates that the regression equation does not explain the relationship at all.

R^2 is calculated as follows

$$r^2 = \frac{\textit{Sum of squares} - \textit{Estimated Sum of Squares}}{\textit{Sum of squares}}$$

Stage 1:

Figure 6 Relationship between age on accession and length of reign with

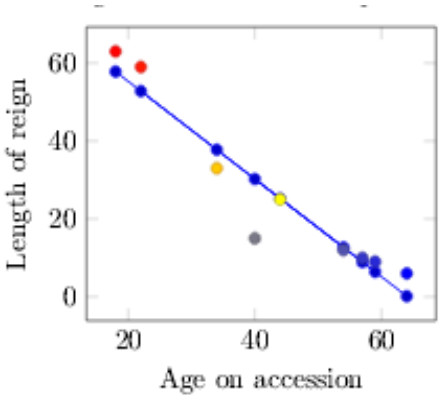
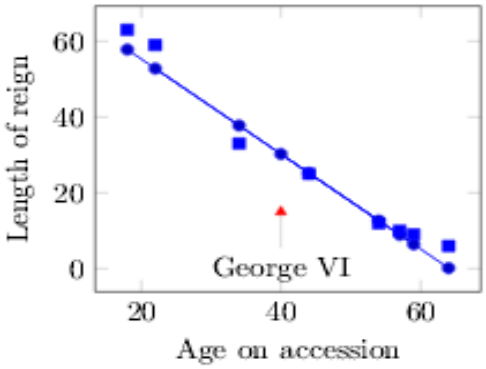


Figure 7 Relationship between age on accession and length of reign with regression



First we need to to calculate the sum of squares (see Table 3).

1. Calculate the mean of the *y* values. *y* is the length of a monarch's reign. We call this ȳ.

   ȳ= 25.78
2. Calculate the difference between each *y* and .
3. Then square the differences
4. Add together the squared differences.This gives us the total Sum of Squares, which is 3769.56.

Stage 2: Secondly we need to calculate the Estimated Sum of Squares (ESS) (see Table 4).

1. Find the estimated *y* values. These are the values of *y* (the length of reign) which would be the case if the regression equation was perfect. We call this ŷ or y-hat.
2. To so this we need to take each *x* value (that is the age on accession to the throne) and use the regression equation to find out what the *y would be* if the regression equation worked. For example George I was 54 when he became king.
3. Our regression equation is

   $$\textit{Length of reign} = (-1.28 \times \textit{age at accession}) + 82.8$$

So for the case of George I the equation is

$$\textit{length of reign} = -1.28 \times 54 + 82.8 = 13.68$$

So ŷ = 13.68

In other words we would expect George I to reign for 13.68 years based in his accession at the age of 54.

1. Do the same for each king and queen.
2. Find the differences between the actual reign (*y*) and the estimated length of reign. (ŷ)
3. Square all these differences.
4. Now add together all the squares of the differences. This comes to 378.72.

$$r^2 = \frac{3769.56 - 378.72}{3769.56} = 0.90$$

Now we calculate the r^2

0.9 is near to 1 which shows that the regression equation is a very strong, though not perfect predictor of reign length. Sometimes r^2 is expressed at a percentage, so we would express 0.9 as 90%. Put in simple terms this indicates that 90% of reign length can be explained by the age at which a monarch came to the throne.

Results summary for predicting monarch reign

We can put a summary of results into a table (Table 5).

8 Exercises

1. Table 6" shows the lowest available prices to travel to different world cities from London and the approximate distance in kilometres. Longer distance flights are more expensive that shorter flights. Create a scatterplot for cheapest price (on the y vertical axis) and distance (on the x horizontal axis)
   1. Calculate the best fit regression line.
   2. Compare your regression line with that of the length of monarchs'

reigns. Which line reflects the data better?

2. Study Table 7 of British Prime ministers since 1940. The table shows the age at which each Prime Minister came into office and how long they were in office for.
    1. Draw a scatterplot with the age the person became Prime minister on the *x* axis and the length of time they served as Prime Minister on the *y* axis. Compare the Prime Minister's scatterplot to the scatterplot of the King and Queens. How do the plots differ? Why do the plots differ?
3. Table 9 shows the distance between Bridgetown, Barbados and the Third class train fare in 1910. [1]
    1. Calculate the regression equation for the relationship between distance (the predictor variable) and the cost in cents (the variable to be predicted).
    2. Calculate *r*2
4. Examine the data in Table 8
    1. Draw a scatterplot of the data with price discount on the *y* axis anrom Rome on the *x* axis.
    2. Calculate the regression equation for Wheat price discount and Distance from Rome. (Note that the numbers are all negative). [2]
    3. Peter Temin [3] reports that this regression equation was rejected by reviewers of Roman history journals as a fluke. Why might experts in Roman history be sceptical of this data?

References

1. Data from: Jim Horsfield (2001) *From the Caribbean to the Atlantic: A Brief History of the Barbados Railway,* St. Austell: Paul Catchpole}
2. These figures were published are from Kessler D. and P. Temin. 2008. 'Money and Prices in the Early Roman Empire' in William V. Harris (ed.) The Monetary Systems of the Greeks and Romans (Oxford).
3. Peter Temin (2006) The Economy of the Early Roman Empire, The Journal of Economic Perspectives 21, pp. 133.151

Table 3 Table for calculating the Sums of Squares

y	ȳ	y−ȳ	(y−ȳ)²
12	25.78	-13.78	189.83
33	25.78	7.22	52.16
59	25.78	33.22	1103.71
10	25.78	-15.78	248.93
6	25.78	-19.78	391.16
63	25.78	37.22	1385.5
9	25.78	-16.78	281.5
25	25.78	-0.78	0.6
15	25.78	-10.78	116.16
	Sum of squares (SS)		
			3769.56

Table 4 Table for calculating Estimated Sum of Squares (ESS)

x	y	ŷ	y−ŷ	(y−ŷ)²
54	12	13.68	-1.68	2.82
34	33	39.28	-6.28	39.44
22	59	54.64	4.36	19.01
57	10	9.84	0.16	0.03
64	6	0.88	5.12	26.21
18	63	59.76	3.24	10.5
59	9	7.28	1.72	2.96
44	25	26.48	-1.48	2.19
40	15	31.6	-16.6	275.56
		Estimated Sum of squares		378.72

Table 5 Summary

Equation	r	r²
-1.28+ age at accession+ 82.8	0.95	0.9

Table 6 Airfares and distances from London to selected destinations worldwide

City	Cheapest price from London £ (y)	Distance from London (km) (x)
Paris	37	341
New York	354	5586
Las Vegas	495	8423
Mexico City	645	8943
Doha	369	5219
Johannesberg	524	9039
Sydney	761	16991
Auckland	765	18329
Hong Kong	504	9646
Barbados	558	6771
Amsterdam	63	359

Table 7: British Prime Ministers since 1940

Prime Minister	Age became Prime Minister	Time as PM (years)
Gordon Brown	56	3
Tony Blair	44	10
John Major	47	7
Margaret Thatcher	54	11
James Callaghan	64	3
Harold Wilson (second term)	58	2
Edward Heath	54	4
Harold Wilson (first term)	48	6
Alec Douglas-Home	60	1
Harold Macmillan	63	6
Anthony Eden	58	2
Winston Churchill (second term)	77	4
Clement Atlee	62	6
Winston Churchill (first term)	66	5

Table 9 Distances and third class train fares to from Bridgetown, Barbados c.1910

Miles from Bridgetown	Station	Third class fare (cents)
2.5	Rouen	4
5.5	Bulkeley	6
7	Windsor	8
9	Carrington	12
10	Sunbury	14
11	Bushby Park	16
13	Three Houses	16
16	Bath	20
20	Bathsheba	24
24	St Andrews	24

Table 8 Relationship between distance from Rome and wheat prices. (c.150BC to AD 80)

Region	Distance from Rome (km)	Distance from Rome "discount"
Sicily	427	-1.5
Spain (Lusitania)	1363	-2.5
Po Valley	1510	-3
Asia Minor (Pisidian Antioch)	1724	-3.13
Egypt (Fayum)	1953	-4
Palestine	2298	-3.25