

# A Tale of Six Cities

---

*Neville Hunt*

Coventry University.

---

---

## ◆INTRODUCTION ◆

---

ONE OF THE interesting aspects of teaching statistics to students from different disciplines is to observe their differing attitudes to the subject. We are probably all familiar with the mathematician who will calculate a veritable feast of summary statistics with immense (and often inappropriate) accuracy - but cannot express in words what any of them actually means. I find the opposite problem with students from a humanities background who fail to appreciate that statistics involves an essentially quantitative approach to data analysis. Only recently an Economics student submitted a piece of coursework to me which claimed to be an analysis of the housing market in Coventry. Not one single figure was to be found in the whole report. Instead it was littered with comments such as: "houses with a garage cost more than houses without a garage, on average"! Did we need data to tell us that?

The crisis of interpretation seems to be most accentuated in the field of regression analysis. My recent experience of examining many hundreds of A Level Statistics projects has revealed the widespread habit of students to collect bivariate data with the sole objective of establishing a (hopefully) significant correlation between their two variables. Often the equation of the regression line is thrown in for good measure, yet with no intention of using it for prediction purposes. Every time I teach regression to students I emphasise the importance of interpreting the regression coefficients in the context of the real world; every time I mark their work I am disappointed.

---

## ◆RULES OF THUMB◆

---

In recent years, prompted by a newspaper article by Robert Leedham in *The Weekend Guardian* newspaper, I have adopted the approach of referring to the regression equation as a "rule of thumb". In this article several examples of newspaper cuttings were given, of which these are three:

- As a rough rule of thumb, an increase of \$100 billion a year in investment demand may raise the world interest rate by about one per cent. **Sunday Telegraph**
- As a rough rule of thumb, withdrawal symptoms are thought to last a month for every year on tranquillisers. **The Guardian**
- The electricity industry says, as a rough rule of thumb, a one degree temperature fall ups demand by 400-600 megawatts (each megawatt being one million watts). **The Guardian**

What people want to know is not **whether** a fall in temperature increases demand for electricity, but **how big** is the increase. After some discussion the students recognise that these statements can only be made with any precision if a regression analysis is conducted.

---

## • A CASE STUDY •

---

An exercise that I have found to be successful in applying the "rule of thumb" concept is that of establishing a relationship between the distance "as the crow flies" between two cities, and the actual distance by road. The crow distance is much easier and quicker to measure on a map, so a rule of thumb for converting a crow distance into a road distance would be useful.

The cities chosen for the study are the 42 cities of England, including Sunderland which gained its city status in 1992.

The exercise itself is supplied in worksheet form overleaf, which readers are welcome to photocopy for their own use. It incorporates a number of skills:



## ◆RANDOM SAMPLING◆

The choice of six as the number of cities is guided by the desire to obtain a sufficient number of data for a viable analysis without making the exercise over-tedious. There are  ${}^6C_2 = 15$  possible journeys between the six cities chosen.

The other fringe benefit is that, since  ${}^{42}C_6 = 5245786$ , it would be extremely unlikely for two students working *independently* to choose the same six cities!

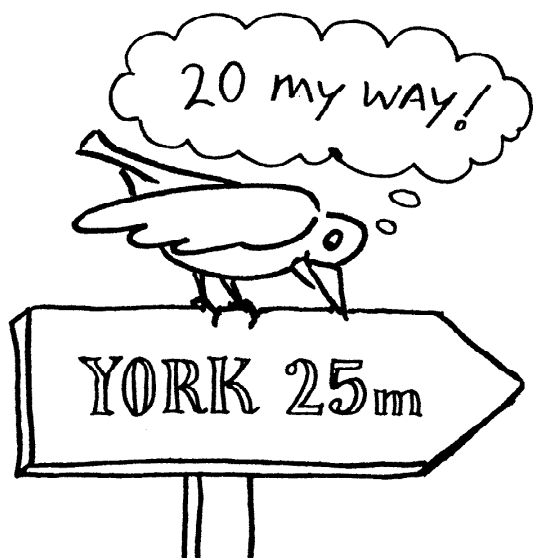
There are obviously many ways of making the selection random. Perhaps the simplest is to use the random number facility on a calculator to generate a random number between 0 and 1, then multiply it by 42, rounding up to the nearest integer. Thus, for example,  $0.104 \times 42 = 4.368 \rightarrow 5 \rightarrow$  Cambridge.

---

## ◆USE OF SECONDARY DATA◆

The road and crow distances may be obtained from any road atlas or map of England. Some atlases supply a convenient table of shortest road distances between principal towns and cities. Unfortunately there is little agreement between different atlases, either about the definition of “principal” or regarding the actual road distances themselves. In any case, most atlases mark the distance between neighbouring places on the map itself, so calculation of the total road distance is not difficult. The distances used in the later example were taken from the 1992 AA Road Atlas.

Some students may have access to proprietary software such as Autoroute Express which offers facilities for measuring both crow distance and road distance using different types of road. In terms of getting a “feel” for the data this method is less beneficial.



## ◆MEASUREMENT◆

The main measuring task is to determine the crow distances. For this a map showing the whole country on a single page will be needed. This will undoubtedly incur a loss of accuracy due to the smallness of the scale. The crow distance can be measured with a ruler in millimetres and will then need scaling into miles. The worksheet allows space for both the scaled and unscaled measurements.

---

## ◆GRAPH PLOTTING◆

It is often not good practice to insist on the origin appearing on a scatter plot, but this is an occasion where it is helpful if it does. Any errors in measurement, scaling or recording ought to be self-evident on the scatter plot. There may well be genuine outliers if one of the cities selected is particularly inaccessible - Plymouth, for example.

---

## ◆REGRESSION ANALYSIS◆

At an elementary level it might be appropriate simply to fit a straight line to the data “by eye”. Given the very high expected correlation this should be relatively easy.

The best approach would be to obtain the least squares regression equation. This gives opportunity for more advanced analysis.

Students may find it strange that the estimated constant coefficient of the regression line is nonzero. If they calculate a 95% confidence interval for this parameter they should be reassured to find that it includes zero.

When it comes to prediction there is an opportunity to bring home to students the difference between mean value prediction and single value prediction. Minitab helpfully supplies a *confidence* interval for the *mean*  $y$  at given  $x$  and a (wider) *prediction* interval for a *single*  $y$  at given  $x$ . Here the distinction is between:

- a confidence interval for the mean road distance between cities 145 miles apart as the crow flies, and
- a prediction interval for the road distance between Sheffield and Bristol, which are 145 miles apart as the crow flies.

---

## ◆INTERPRETATION◆

The constant coefficient of the regression line will undoubtedly generate some discussion. If it is positive, it could be argued that the ring roads and one-

way systems of modern cities require drivers to travel several miles before they actually make any progress towards their destination! If it is negative students might consider what errors the estimate is subject to - measurement, processing, sampling, etc.

Provided that the constant coefficient is sufficiently small to be regarded as zero (and experience suggests that this is normally the case), the slope coefficient gives us the scaling-up factor from crow to road. I despair with those of my students who persist in telling me that the rule of thumb is to “multiply by 1.2483” surely, “add on a quarter or “increase by 25%” would be more meaningful?

### ◆EXAMPLE DATA ◆

Suppose that the six cities selected at random were Cambridge, Carlisle, Hull, Manchester, Nottingham and Southampton. The crow and road distances are given in the table below and a Lotus 1-2-3 scatter plot is provided.

Journey	Crow distance	Road distance
Cambridge to Carlisle	227	260
Cambridge to Hull	110	141
Cambridge to Manchester	132	161
Cambridge to Nottingham	74	87
Cambridge to Southampton	110	131
Carlisle to Hull	129	173
Carlisle to Manchester	99	119
Carlisle to Nottingham	156	190
Carlisle to Southampton	287	340
Hull to Manchester	80	99
Hull to Nottingham	67	93
Hull to Southampton	203	254
Manchester to Nottingham	59	70
Manchester to Southampton	181	231
Nottingham to Southampton	140	167

The correlation between road and crow distances is 0.995, although the linearity is so self-evident as to make correlation superfluous.

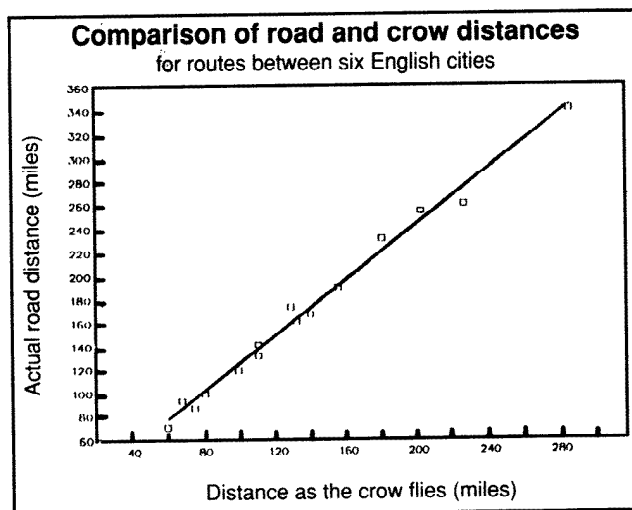
The estimated regression equation is:  
 $\text{road} = 7.37 + 1.17 \text{ crow}$

The rule of thumb is therefore “add on 17%, then a further 7 miles”.

Further analysis could involve calculation of confidence/prediction intervals, although with only 15 points we cannot expect the intervals to be particularly precise.

Statistic	Estimate	St.Error	95% C.I.
constant	7.368	5.029	(-3.495, 18.231)
crow	1.17112	0.03342	(1.099, 1.243)

The constant coefficient of 7.368 is evidently not



significantly different from zero. If we regard the constant term as negligible, the crow coefficient indicates that the road distance is 17% (or approximately one sixth) further than the crow distance, a simpler rule of thumb.

The predicted road distance between Sheffield and Bristol is 177 miles, given their crow distance of 145 miles. The 95% prediction interval is from 159 miles to 195 miles. The actual road distance of 184 miles falls within this interval.

By comparison, the 95% confidence interval for the mean road distance between English cities 145 miles apart is from 172 miles to 182 miles, substantially more precise, as we should expect. There is no reason at all why the Sheffield to Bristol distance of 184 miles should fall within this interval.

### ◆ CONCLUSION ◆

Many statistical investigations have hidden pitfalls or complications which serve only to distract from the underlying message. Data collected to demonstrate a relationship between two variables often display a disappointingly small correlation, or are surprisingly non-linear. The investigation advocated here has been found to produce satisfying results with a clear and patently useful application. It demands quantitative interpretation. Although it is best tackled within a regression context, it could be tackled at a more elementary level using line fitting by eye.

#### References

- AA Road Atlas of Great Britain and Ireland, Tiger Books International, 1992.
- Autoroute Express, Ordnance Survey Digital Mapping, NextBase Ltd, 1988.