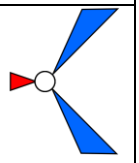
16 AA

bncdoc.id	B16
bncdoc.author	Marsh, Catherine
bncdoc.year	1988
bncdoc.title	Exploring data.
bncdoc.info	Exploring data. Sample containing about 38183 words from a book (domain: applied science)
Text availability	Worldwide rights cleared
Publication date	1985-1993
Text type	Written books and periodicals
David Lee's classification	W_ac_soc_science

<16/c>  Key: **Footprint ConEn1** **Footprint ConEn2** **Footprint ConEn3**	clearly in the plot of residual Y versus fitted Y than in the original scatterplot. What to do if the relationship is curved is discussed fully in the next chapter. Predicting morbidity from mortality Let us return now to the policy question: how well do **death rates** predict **sickness**, the condition that requires the cash? Forster is right that the correlation between the variables is far from perfect; the data points are spread quite widely to either side of the line. We want residuals to be small. They represent the failures of prediction; the larger they are, the worse the fit. But how are we to evaluate their size in any particular instance? The general answer that we first came across in section 6.7 is: the residual spread should be small in comparison with the original spread. In this case the residual d Q is 20 (figure 10.7), compared with an original d Q of 26 (derived from column 2 of figure 10.2); the spread has been reduced to around three-quarters its original size. Is this a big reduction? First consider two extreme cases which delineate the boundaries. If the association was perfect, all the points would lie on the line, and there would be no spread in the residuals; 100 per cent of the original spread would be accounted for, and we would describe the two variables as being perfectly correlated. If there was no relationship at all, the spread of the residuals would be the same as originally, and there would thus be no reduction; we would say that there was no correlation between the two variables. Reducing the spread by 25 per cent is not a very dramatic reduction; it is nearer zero than 100 per cent. So regional **death rates** do not predict regional self-reported **sickness rates** very well. Critics of the decision to include **death rates** in the RAWP formula have implicitly argued that they should correlate nearly perfectly with **sickness rates**; in the light of such evidence, they have advocated dropping them (Forster 1977; Barr and Logan 1977). However, an inadequate indicator may be preferable to no indicator; failure to include **death rates** in the formula would mean that funds were even less related to need than at present. Perhaps self-reports of **sickness** should be substituted for death rates in the RAWP formula. However, we have no guarantee that the **sickness indicators** themselves are reliable. Moreover, attempts to classify self-reported data into **disease categories** are notoriously unreliable, and a disease-specific measure of need is required since **the costs** of **treatment** differ so widely. Every currently available indicator of **health** need is inadequate in one way or another. **Treatment** rates for different **diseases** reflect
	availability of the treatment
	rather than prevalence of **the disease**. Self-reported **sickness** is of unknown validity and unobtainable within reliable **disease categories**. **Death rates** seem not to be a very good indicator of **disease rates** either. Is the attempt to distribute resources according to need therefore impossible? Perhaps **death rates** will have to suffice until a more serious attempt is made by **health care** professionals to ascertain the regional variations in **disease** independently of **treatment**. Conclusion In this chapter we have looked at the extent to which chronic sickness rates as reported on the GHS are predictable from death rates. The full relationship was first examined

	by means of a useful pictorial device: the scatterplot. This relationship was next summarized by fitting a straight line. The strength of the association between the mortality and morbidity was given by the slope of this line, indicating how much sickness goes up for each point increase in the death rate. The spread in overall sickness rates was compared with the spread in residual sickness rates to obtain a measure of the degree of correlation around the line. Those who have been exposed to the confirmatory techniques of linear regression may recognize the analogous measures. In regression, the criterion for line fitting is to minimize squared residuals, the 'least-squares' rule 4 of section 10.5. It is the variance, not the midspread, which is broken down into a fitted ('explained') and residual ('unexplained') component. The correlation is given by the Pearson's correlation coefficient r. Regression techniques have their advantages: the lines can be derived in one fell swoop from a formula, and do not require iterating. Moreover, if the residuals are well behaved (Gaussian, without freak values) then the calculation of the likely error associated with the coefficients is fairly straightforward. However, error terms are often highly non-Gaussian, and outliers from the line are the rule rather than the exception. Regression techniques, because they set out to make the squared distances of the residuals from the line as small as possible, can be unduly influenced by a few exceptional data points. They are therefore much less resistant than the techniques introduced in this chapter. The slope, b, is mathematically very close to the proportion difference, d. It is an asymmetric measure, just like d: the slope depends on which variable is treated as the response variable, just as d depends on which way the percentages have been run; exercise 10.2 has been set to illustrate this. Symmetric methods of line fitting exist, such as principal components analysis and factor analysis, which summarize the extent to which two variables cluster together (rule 2 of section 10.5), but they are not