WEEK 8 RELATIONSHIPS BETWEEN VARIABLES

AIMS

To introduce the idea of *association* between variables and describe some methods of assessing and measuring the strength of any such association. In particular to examine the use of scatterplots and correlation coefficients. To contrast a measure of agreement with measures of association. To introduce the idea of the regression model, particularly linear and logistic regression.

OBJECTIVES

At the end of this Unit you should be able to:

- Explain what is meant by the term "association".
- Draw a scatterplot (scattergram) and assess qualitatively the strength and direction of any association between the variables.
- Explain and interpret both Pearson's and Spearman's correlation coefficient values and judge their significance.
- Choose the correlation measure most appropriate for any given set of data.
- Explain the difference between association and agreement and be able to describe the principle behind kappa.
- Outline the shortcomings of correlation particularly in the context of causation.
- Describe the basic idea underlying linear regression analysis, and correctly interpret the significance of the results from a linear regression analysis.
- Explain the property of adjustment.
- Explain the principal difference between linear and logistic regression.

Reading: Bland: Sections 11.9; Section 11.10 (last two paragraphs only); Section 12.4 (ignore equations and concentrate on testing the significance of the coefficient, towards the end of the section).

or Bowers-2: Chapter 8 (ignore the computer application sections).

Introduction

Up to now we have focussed on the application of descriptive and inferential statistics to a single variable. For example, the variable "bone mineral density" and how it differs between two groups of women (Figure 7.1), or the variable "stump pain" and how it differs between placebo and treatment groups (Figure 6.5). The methods we have used so far are, not surprisingly, known as **univariable statistics**.

In this unit we turn to ways of investigating connections between *two* (or more) variables. For obvious reasons such methods of analysis are known collectively as **multivariable** statistics. We want to focus on two multivariable procedures, **correlation** and **regression**. The former procedure enables us to describe the strength and direction of the **association** between two variables, and assumes no *causality*. By this we mean that changes in the value of either variable do *not* necessarily lead to or cause changes in the other variable - simply that the two variables seem to move *together* in some way. In regression analysis we do assume that changes in one variable are *caused* by changes in one or more other variables. We use regression analysis to model and subsequently analyse such situations. We will start with the concept of *association*.

Measuring association

When we describe two variables as being "associated" we mean that the variables show some sort of connection. For example, high values of one variable tend generally to be associated with high values of another variable and low values with low values. For example, cigarette smoking and coffee consumption. This form of association is said to be *positive*.

Alternatively, we may find that high values of one variable tend generally to be associated with *low* values of the other and vice versa. For example, wine consumption and annual income. This form of association is said to be *negative*.

Scatterplots

In situations where we are interested in exploring a possible connection between two variables, it is nearly always useful to produce a **scatterplot** of the data (also known as scattergrams or scatter diagrams). It doesn't matter which variable is plotted on which axis if we are only interested in exploring possible association between the two variables. (Note however, that if we are examining a causal model, i.e. where one variable *depends* on another, the dependent variable is plotted on the y or vertical axis, the "causing" variable on the x or horizontal axis).

If any association exists between the two variables it will usually be possible to discern a "pattern" in the scatter of points. For example, the points may appear to scattered around an imaginary straight line or around a regular curve. The closer the scatter is to the straight line or curve, the closer the association between the variables. In the case of a straight line scatter, if the scatter slopes upwards from left to right this is indicative of a positive association, if down from left to right, of a negative association.* No discernible pattern indicates (probably) weak or no association. The strength of any association is judged by how close to some imaginary straight line the points lie.

Figure 8.1 shows a scatter plot of % mortality from aortic aneurysm and number of hospital episodes per year in 18 UK hospitals. This scatter, which slopes down from left to right, indicates moderate *negative* association. Hospitals which have recorded a higher number of episodes experience a lower mortality rate, and vice versa. It would be possible to draw an imaginary straight line through these points although most of the points would not lie particularly close to it. In fact, there is some evidence of a "curvy" association.

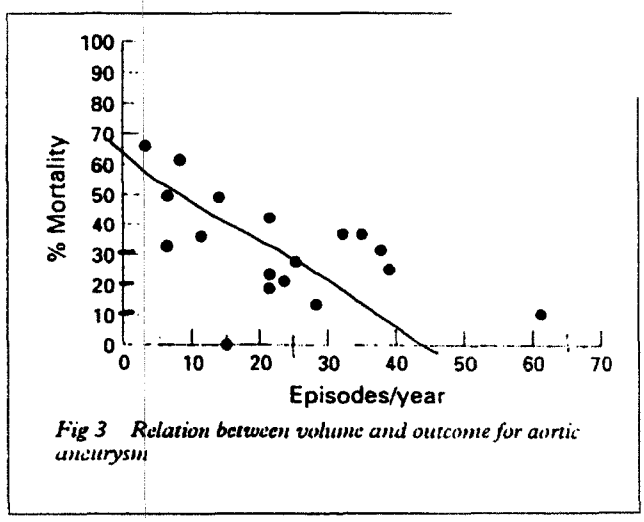


Fig 3 Relation between volume and outcome for aortic aneurysm

Figure 8.1 Scatterplot of % mortality from aortic aneurysm and annual number of episodes treated by 18 UK hospitals. Quality in Health Care, 4, 1995.

As a further example Figure 8.2 shows a scatterplot of body mass index as reported by patients and as measured in a clinic. This scatter reveals a strong *positive* association. Notice how closely to the straight line the points lie,

* We are not interested here in "curvy" patterns which may be indicative of a non-linear (not a straight-line) association, the implications of which we cannot deal with in this course.

indicative of a strong *linearassociation.*

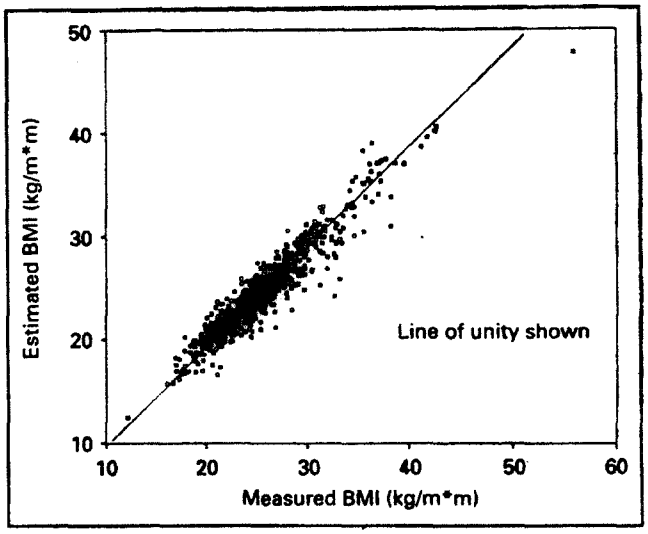


Figure 1. Scatter plot. Estimated BMI: using patient reported data. Measured BMI: using staiometer and scales.

Figure 8.2 **Scatterplot of patient reported and clinic measured body mass index. British J of General Practice, 48, 1998.**

Q. 8.1 Figure 8.3 is a scatterplot of suicide rate and the use of calcium channel blockers from a cross-section study across 152 Swedish municipalities. (a) Comment on the direction and strength of any association between the two variables; (b) The authors had a causal relationship in mind. What do you think it was?

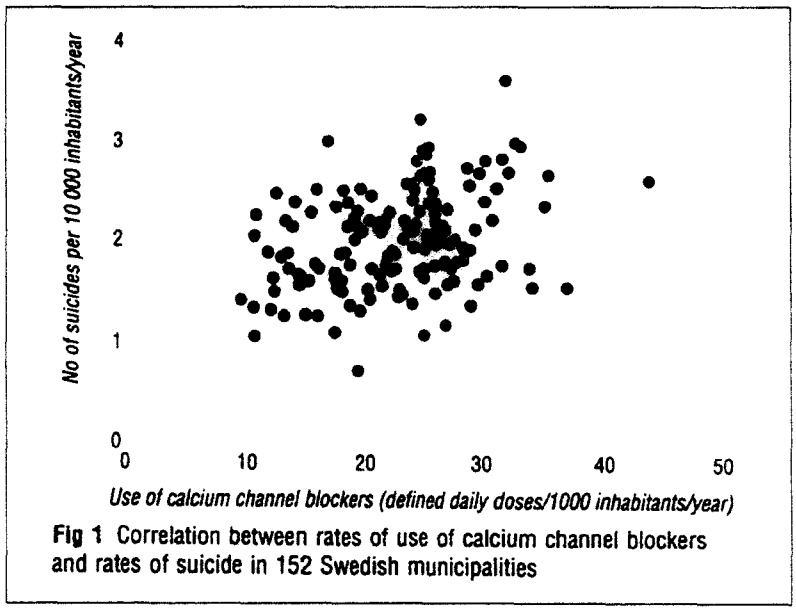


Fig 1 Correlation between rates of use of calcium channel blockers and rates of suicide in 152 Swedish municipalities

Figure 8.3 **Scatterplot of suicide rate and the use of calcium channel blockers in 152 Swedish municipalities. BMJ, 316, 1998.**

Correlation coefficients

A scatter plot can only give us a qualitative idea of the strength of an association between two variables. However, a single quantitative measure of **linear** association is provided by the **correlation coefficient**. That is, it is a measure of how close the points come to lying exactly on a straight line. A correlation coefficient can take any value from –1 to +1. Negative values mean a negative association, positive values a positive association. The closer the value is to –1 or +1, the stronger the association. A value close to zero indicates that the two variables have a weak or no association.

In practice we use sample data to calculate the value of the *sample* correlation coefficient with which we can estimate the *population* correlation coefficient, usually designated ρ (pronounced ro), and calculate confidence intervals for it. There are a number of different correlation coefficients, but we have space only to consider the two most widely encountered in the clinical literature – Pearson's product moment correlation coefficient r, and Spearman's rank order correlation coefficient r_s.

Pearson's product moment correlation coefficient

Pearson's correlation coefficient is the most widely used measure of correlation. It measures the strength of *linear* association between two *metric* variables. This measure is really only valid if the data is metric continuous and approximately Normally distributed, but in practice this requirement is not usually applied too stringently. However, if the data is not continuous and Normal (for example, if there are outliers) then the estimated correlation coefficient may be misleading and caution is required in interpreting the results. We can calculate a confidence interval for the true population correlation coefficient or perform a hypothesis test on it.

Q. 8.2 The table in Figure 8.4 is from a study into the medical record validation of maternally reported birth characteristics and pregnancy-related events among the mothers of children attending a child cancer clinic. The table shows the Pearson correlation between gestational age, as reported by the mother and as recorded in medical records, for a number of specific demographic subgroups (ignore the last column). Which estimated correlation appears to be: (a) the strongest? Is it positive or negative? (b) The weakest? (c) Which correlation coefficient is estimated the most precisely? The least precisely? Explain. (d) Do you think the association between gestational age as recorded by the mother and

from medical records is effected by the birth order of the child in question? Explain.

TABLE 3. Validity and reliability of gestational age within specific demographic subgroupings among participating members from a United States and Canadian cooperative clinical trials group and matched controls, 1983–1988

	Correlation of gestational age	98% CI*	Kappa statistic†
All gestational ages	0.839	0.817–0.859	0.62
Case/control status			
Cases	0.849	0.813–0.878	0.63
Controls	0.835	0.805–0.861	0.61
Education			
<High school	0.694	0.553–0.797	0.51
High school	0.833	0.790–0.868	0.63
>High school	0.835	0.804–0.861	0.62
Household income			
<$22,000	0.791	0.734–0.837	0.59
$22,000–$34,999	0.882	0.849–0.908	0.62
≥$35,000	0.843	0.800–0.877	0.65
Unknown	0.745	0.641–0.823	0.60
Time (years) from delivery to interview			
<2	0.896	0.862–0.921	0.64
2–3.9	0.821	0.784–0.852	0.63
4–5.9	0.828	0.775–0.869	0.61
6–8	0.852	0.734–0.920	0.42
Maternal age (years)			
<25	0.822	0.773–0.861	0.64
25–29	0.889	0.862–0.912	0.63
30–34	0.760	0.694–0.813	0.57
≥35	0.888	0.824–0.930	0.64
Birth order			
First born	0.880	0.853–0.903	0.67
Second born	0.815	0.778–0.846	0.57
≥Third born	0.632	0.416–0.781	0.52
Maternal race			
White	0.846	0.822–0.866	0.64
Other	0.782	0.680–0.855	0.42

* CI, confidence interval.
† Three categories, <38, 38–41, ≥42 weeks.

Figure 8.4 The correlation coefficients of gestational age, as reported by the mother and as recorded in medical records, for a number of specific demographic subgroups in a child-cancer study. Amer J Epidemiology, 145, 1997.

Spearman's rank correlation coefficient

If the data for either or both of the variables, although metric continuous, cannot be said to be approximately Normally distributed, or if it is metric discrete or ordinal, then Spearman's correlation coefficient (denoted r_s) is appropriate. Essentially it approximates Pearson's coefficient by applying the same calculation to the sample data after it has first been *ranked*.

Q. 8.3 Figure 8.5 shows Spearman's correlation coefficients between breast size and a number of other factors, according to oral contraceptive use, from a study into endogenous hormone levels and oral contraceptive use. The subjects were Swedish female university students. Amer J of Epidemiology, 45, 1997.

(a) Identify the type of the seven variables in the table with which breast size is correlated, and suggest the most appropriate correlation coefficient for each. (b) With which variable, in which group, is breast size most strongly and significantly associated? And the most weakly associated? What directions are these associations? The p-values in the table are to test the null hypothesis that the true, population, correlation coefficient is zero. (c) With which variable(s) is breast size not significantly associated among those who have never used oral contraceptives? Explain.

TABLE 6. Spearman rank correlations (r_S) between breast sizes in healthy Swedish female university students, according to oral contraceptive use and body mass index, height, weight, family history of breast cancer, age at menarche, and age, 1993–1994*

	Oral contraceptive use							
	Never users (n = 20)		Former users (n = 20)		All nonusers (n = 40)		Current users (n = 25)	
	r_s	p	r_s	p	r_s	p	r_s	p
Body mass index†	0.47	0.038	0.71	<0.001	0.53	<0.001	0.27	0.196
Height	0.25	0.286	0.42	0.068	0.32	0.046	−0.06	0.783
Weight	0.47	0.037	0.55	0.011	0.50	0.001	0.24	0.248
Family history of breast cancer in a first or second degree relative	−0.17	0.473	−0.41	0.071	−0.24	0.131	0.01	0.951
Age at menarche	−0.23	0.329	0.03	0.889	−0.06	0.735	−0.08	0.707
Waist:hip ratio	−0.00	0.985	0.35	0.133	0.18	0.256	0.25	0.226
Age	0.09	0.718	−0.20	0.396	−0.04	0.814	0.15	0.480

* Values from measurements taken during menstrual cycle days 5–10 were used.
† Weight (kg)/height (m)^2.

Figure 8.5 Spearman's correlation coefficient between breast size and a number of other factors, from a study into endogenous hormone levels and oral contraceptive use. Amer J of Epidem, 45, 1997.

Q. 8.4 Figure 8.6 shows a scattergram of % body fat against age for a sample of 18 healthy male adults aged between 23 and 61. (a) The sample Pearson correlation coefficient (with its p-value) is one of the following. Which one and why?

(i) - 0.790 (0.043); (ii) 0.425 (0.079); (iii) - 0.13 (0.048); (iv) 0.99 (0.001) .

(b) Is the Pearson correlation coefficient an appropriate measure of association for these data?

(c) Could we predict the likely % body fat for a person aged 35 knowing only the value of the correlation coefficient between % body fat and age? Explain your answer.

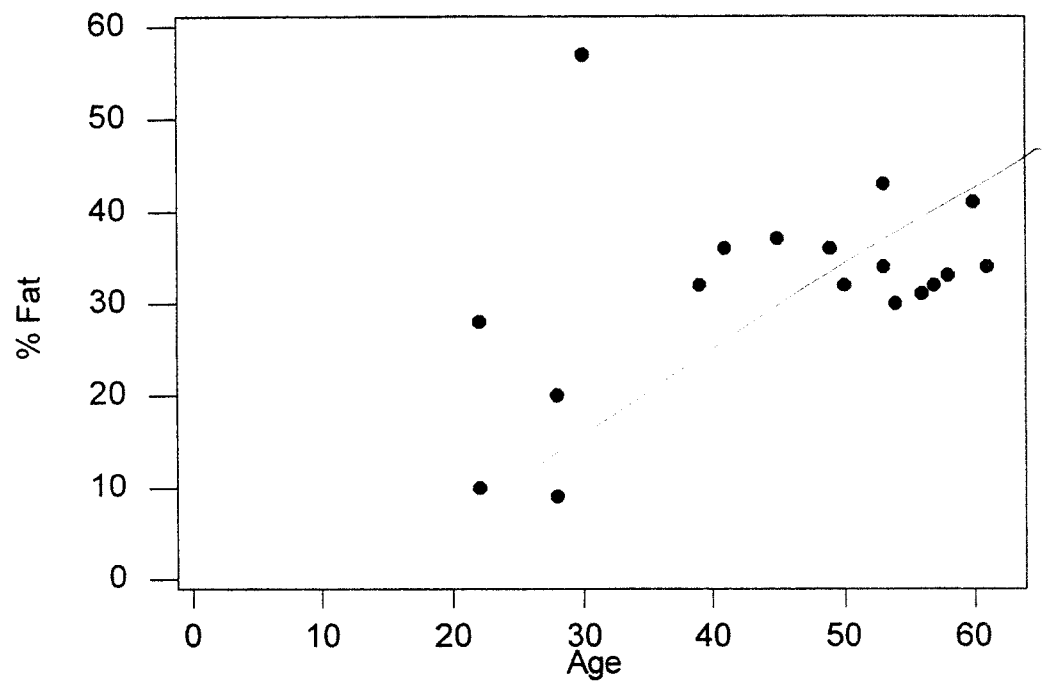


Figure 8.6 Scatterplot of % body fat versus age for 18 healthy male adults

When you use a correlation coefficient you should bear in mind three principal limitations:

- It only measures linear association. Thus even if the data is clustered very closely to a curve, a calculated correlation coefficient may show little or no association.

- It is sensitive to sample size - the bigger the sample the bigger the correlation coefficient. For example you might find that with a sample size of 100 the p-value for a particular correlation coefficient is 0.08, whereas with a sample of 200, the p-value is 0.04, and this is not simply because a bigger sample is probably more representative of the population.

- It cannot be used for prediction purposes, since it offers no information of a causal nature.

Association and agreement

One further point is worth making. Association is not the same as **agreement**. Association measures the degree to which two sets of values tend to move *together*. Agreement measures the degree to which the values are actually the *same*. Two variables can be closely associated without their values necessary agreeing. To illustrate the difference, suppose a trainee paramedic assesses the Glasgow Coma Scale score of 10 RTA patients*. His supervisor simultaneously scores the same patients. Their scores are:

Patient	1	2	3	4	5	6	7	8	9	10
Trainee	5	9	3	7	8	5	4	9	7	5
Supervisor	4	10	2	5	9	4	2	8	6	5

We can see that when the supervisor scores high the trainee also tends to score high. When the supervisor scores low the trainee tends to score low. The two sets of scores are strongly and *positively* associated. Pearson's $r = +0.950$, with a p-value of < 0.000, so the association is highly significant. However only one score out of the ten (or 10%) agree exactly, so agreement is very poor. And don't forget that we would have *expected* them to agree by chance on a few of them anyway, even if they had to take a wild guess without even seeing the patients.

* The Glasgow Coma Scale (from 0 to 16) is used to assess the seriousness of head injury, such as those sustained in a road traffic accident (RTA). Scoring system is: 13-15 = mild injury; 9-12 = moderate injury; ≤ 8 = severe injury.

Kappa

What we need is a measure which adjusts the observed agreement for the number of agreements which are due to chance alone. This is what Cohen's **kappa** (written as ***k***) does. Kappa measures the proportion of scores which agree (i.e. fall in the same category) *adjusted* for the proportion which could be expected to agree by *chance*. Kappa is properly known as the **chance-corrected proportional agreement statistic**.

Kappa can vary between 0 (no agreement) and 1 (perfect agreement). Values of kappa may be assessed with the help of the table below. Only values for kappa of about 0.60 or more indicate good agreement.

κ	Strength of agreement
< 0.20	poor
0.21 -0.40	fair
0.41 - 0.60	moderate
0.61 - 0.80	good
0.81 - 1.00	very good

Assessing agreement with kappa

Strictly speaking Kappa is a measure of agreement between two *nominal* variables, but many problems can be "nominalised". For example, since a score of ≤ 8 is a critical cut-off point in the Glasgow Coma Scale, let's transform the GCS data in the table above into scores of ≤ 8 (labeled S for Serious) or ≥ 9 (labeled NS for Not-serious). The revised table below shows the result.

Patient	1	2	3	4	5	6	7	8	9	10
Trainee	S	NS	S	S	S	S	S	NS	S	S
Supervisor	S	NS	S	S	NS	S	S	S	S	S

Now we can see that 7 out of the 10 scores agree, an apparent agreement level of 70%. However, when expected agreement is taken into account, Kappa equals 38%, so agreement is still only.

It is possible to calculate confidence intervals for kappa, which enables some estimate of its precision to be made. In general, variables which are found to be firmly *associated* will usually show good *agreement*, and vice versa, although, since this is not invariably the case, the methods are not interchangeable and correlation should not be used as a proxy for agreement.

One limitation of kappa is that it is sensitive to the proportion of subjects in each category, in other words, to prevalence. The consequence of this is that kappas from different studies should not be compared if the prevalences are not the same.

Q. 8.5 Figure 8.4 above, shows not only the correlation for gestational age as recalled by mother and from medical records, for a number of demographic sub-groups, it also gives the value of the kappa statistic for the degree of agreement between the two gestational ages for the same sub-groups. (a) Which sub-group displays, (i) the best agreement; (ii) the worst agreement. (b) Does the sub-group with the highest correlation also have the highest value of kappa?

Regression analysis

Linear regression

If clinical researchers wish to analyse *causal* relationships between variables they commonly turn to the techniques of regression analysis. The most popular is perhaps the *logistic* regression model but we will start here with the **linear regression model**. since the basic ideas are more than likely already familiar to you.

Suppose we believe that age and % body fat are related in such a way that increases in age bring about or cause increases in body fat. Moreover the relationship is *linear*, i.e. a scatterplot of the points would show the sample values scattered around a straight line. The implications of this are that an increase in age of one year whether from 20 to 21 or from 63 to 64, brings about the *same* increase in % body fat. Note that although inspection of the scatterplot in Figure 8.6 suggests that the association between the two

variables is *positive*, it doesnot provide any evidence about the possibility of a *causal* relationship between them. The causal nature of any relationship is established prior to any analysis and is based on historical observation, biological processes, considerations of plausible cause and effect, theoretical developments, insight based on practice, and so on.

We can format this linear relationship as an equation:

$$\% \text{ body fat} = \beta_1 + \beta_2 \times \text{Age}$$

This should be familiar to you from GCSE Maths as the equation of a straight line, where β_1 is the intersect coefficient and β_2 is the slope coefficient. This equation is known as the **linear regression equation**. The variable on the left-hand side of the equation is referred to as the *outcome* or *dependent* variable, and must be continuous metric and Normally distributed. The variable on the right-hand side is known as the *independent* or *explanatory* variable or the *factor*, and can be of any type.

If there really is a causal relationship between body fat and age (the question the researchers wish to address) then β_2 has to have a non-zero value. If it was 0, then it wouldn't matter what value age took, once it was multiplied by 0 it would disappear from the equation and could have no influence on body fat.

So the job of the researcher is, first to estimate the values of the two coefficients using the sample data, and second to assess the significance of β_2 using a confidence interval or hypothesis test. If either the confidence interval for β_2 includes 0 or the p-value for β_2 is ≥ 0.05, then changes in age do not effect % body fat (see Units 6 and 7).

If we apply a suitable computer program to the age/body fat data it is easy enough to calculate β_1 and β_2 to be 16.334 and 0.349 respectively. The estimated regression equation is thus

$$\% \text{ body fat} = 16.334 + 0.349 \times \text{Age}$$

The constant coefficient β_1 is of little interest and is usually ignored. For β_2 the computer also calculates a p-value of 0.079 together with a 95% confidence interval of (-0.045 to 0.742). On both counts β_2 is not statistically significant and, from this data anyway, we would conclude that age did not influence body fat, and thus that there was no relationship between the two variables. The regression line is shown plotted through the data in Figure 8.7.

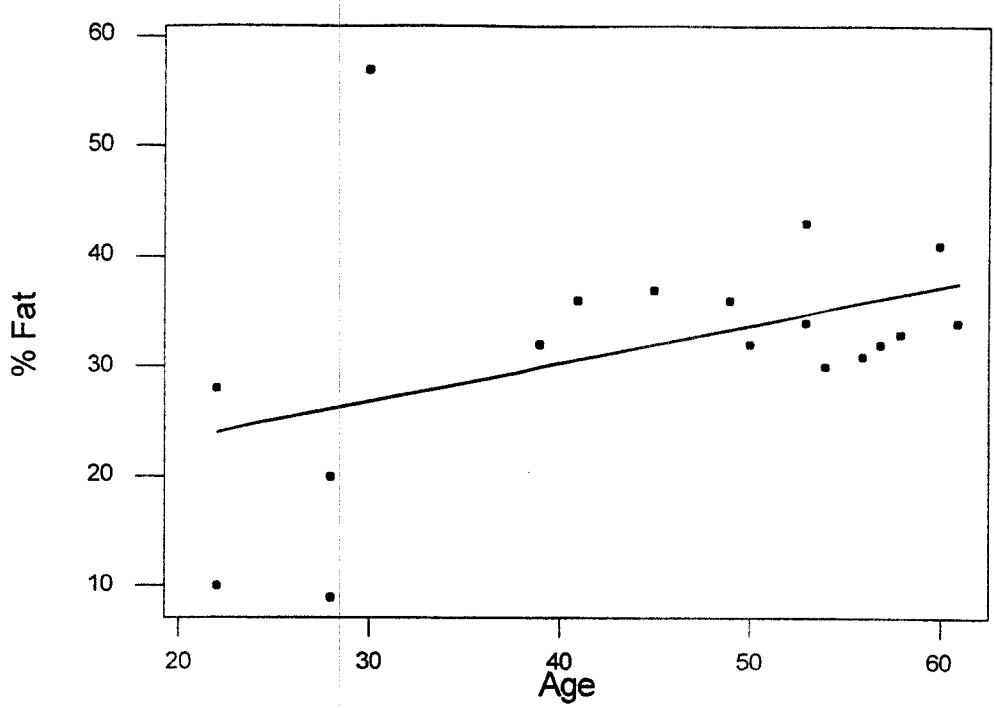


Figure 8.7 Estimated linear regression line between % body fat and age for 18 healthy adults.

Q. 8.6 If two groups of subjects are one year apart in mean age, how much more % body fat will the older of the two have on average (assuming the relationship is statistically significant)? What is this value?

In practice, regression equations will contain more than variable on the right hand side and these may be a mixture of nominal ordinal and metric.

As an example, Figure 8.8 is from a study into the effect of chronic hypertension in women on their risk of producing small-for-gestational-age babies.

The 2185 subjects were recruited from five pre-natal clinics in France between August 1991 and May 1993. The table reports the results of a linear regression analysis, in which the dependent variable is *birthweight* (g). The independent variables are a mixture of continuous, ordinal and nominal variables. The table also provides, for each independent variable, its estimated coefficient value, β, the associated p-value, and the standard error SE (see Unit 6 for a note on standard error as a measure of the preciseness of the estimate - smaller is better).

TABLE 3. Effect of chronic hypertension on mean birth weight values, multiple linear regression (n = 1,938), France, 1991–1993

Independent variable	β* ± SE†	p value
Chronic hypertension (yes vs. no)	−161 ± 48	0.0009
Smoking (yes vs. no)	−113 ± 24	<0.00001
Weight at initial visit (kg)	8 ± 1	<0.00001
Mother's height (cm)	9 ± 2	<0.00001
Age (years)	1 ± 2	0.76
Multiparous (yes vs. no)	120 ± 21	<0.00001
Ethnic group of origin		
North African vs. Western European	108 ± 37	0.004
Sub-Saharan African vs. Western European	−140 ± 52	0.007
Other origin vs. Western European	19 ± 33	0.56
Educational level		
Primary school or below vs. university	−43 ± 31	0.16
Secondary school vs. university	−65 ± 25	0.008
Technical school vs. university	−50 ± 33	0.13

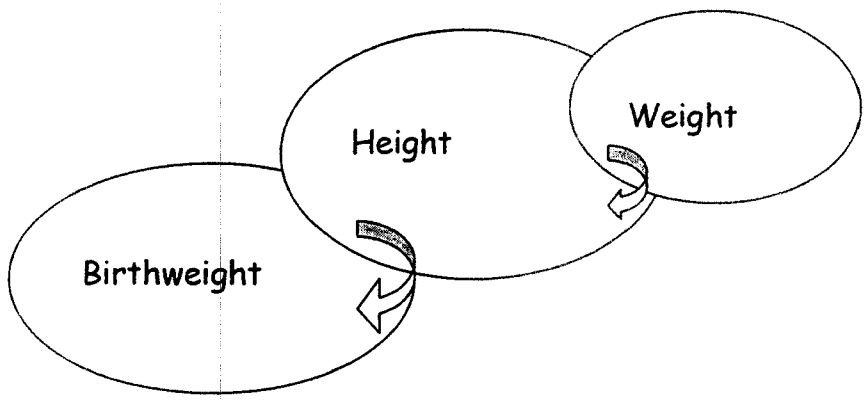
* β, partial regression coefficients adjusted for the other predictors and gestational age at delivery.
† SE, standard error.

Figure 8.8 Results from a linear regression analysis into the effects of chronic hypertension in mothers on the birthweight of their babies according to a number of risk factors (the independent variables. The dependent variable is mean birthweight (g). Amer J of Epidemiology, 145, 1997.

Q. 8.7 (a) Which variables are significant influences on mean birthweight? (b) By how much and in what direction does having chronic hypertension have on mean birthweight? (c) What is the average difference in birthweight of babies with mothers who smoke compared to babies whose mothers don't smoke? (d) If two otherwise similar mothers differ in weight at initial visit by 1kg, who's baby is likely to be heavier and by how much?

One very attractive feature of the regression model is that it measures the effect of each independent variable on the dependent variable after **adjusting** for the influence any other variable(s) might have. For example, in the birthweight study above, mothers' Height and Weight are likely to be closely linked, so that a change in either brings about an associated change in the other, and hence a change in birthweight. The idea is illustrated in the Venn diagram below. It is difficult therefore to disentangle their separate effects on

Birthweight. However, this is what the regression analysis achieves. So the value of 8 for the mothers' Weight coefficient in Figure 8.8 measures the "pure" effect of a change in a mothers' Weight on the change in Birthweight, after any contribution made by the Height variable has been eliminated.



Another reason for the enthusiasm of clinical researchers for the regression model is that it enables what is called the **confounding** problem to be addressed. You will learn about confounding elsewhere so it will not be discussed any further here.

Logistic regression

In the linear regression model described above, the dependent variable is required to be metric continuous and Normally distributed. In clinical research, the dependent or outcome variable is more often *dichotomous*, i.e. it can only take two possible values, for example, alive or dead, malignant or benign, case or control, treated with active drug, treated with placebo, and so on. These two states are usually scored as 0 and 1. If we wish to apply regression analysis to such studies we need to turn to the **logistic regression model**. The maths of this is somewhat more complicated than that for the linear regression model, so we cannot do much more than give a brief summary here*.

The most popular use of logistic regression analysis is to determine *odds ratios* for risk factors. We have already seen an example of this in Unit 5. The cross-section study from which Figure 5.3 is taken (reproduced here for convenience as Figure 8.9) uses a sample of 890 women aged 18 to 35 to investigate possible risk factors for genital chlamydia. The odds ratios and their 95% confidence intervals shown in the table are produced directly by a computer logistic regression analysis in which the outcome variable is "Has genital chlamydia (No =

* For the mathematically minded the logistic regression equation is $Y = e^{(\beta 0 + \beta 1 \times X1 + \beta 2 \times X2 + ...)}$

0, Yes = 1)". The independent risk factors are the variables shown in the table, e.g. age, marital status, number of partners, etc. The numeric values of the coefficients are not of as much interest in logistic regression as they are in linear regression and are not usually shown.

Table 2 Demographic and behavioural characteristics of 879* women participating in study—comparison of those positive for chlamydia infection with those negative for infection

Risk factor	% (No) of women with positive result	Odds ratio
Age group (n=848):		
≤20	10.6 (9/85)	8.64 (2.28 to 32.8)
21-25	3.8 (8/210)	2.89 (0.76 to 11.0)
26-30	0.9 (3/331)	0.67 (0.13 to 3.34)
≥31	1.4 (3/222)	1
Marital status (n=822):		
Married	0.6 (1/170)	0.19 (0.02 to 1.45)
Cohabiting	3.1 (8/260)	1.00 (0.41 to 2.49)
Single	3.1 (12/392)	1
No of partners in past year (n=812):		
0-1	1.7 (11/630)	1
≥2	4.9 (9/182)	2.93 (1.19 to 7.18)
One or more new partners in past 3 months (n=782):		
No	2.4 (16/671)	1
Yes	4.5 (5/111)	1.93 (0.69 to 5.38)
Ever had sexually transmitted disease (n=818):		
No	2.3 (14/616)	1
Yes	3.5 (7/202)	1.54 (0.61 to 3.88)
Ever had termination of pregnancy (n=831):		
No	2.6 (15/575)	1
Yes	2.7 (7/256)	1.05 (0.42 to 2.61)
Genitourinary symptoms at present (n=807):		
No	2.4 (11/467)	1
Yes	3.2 (11/340)	1.33 (0.53 to 2.99)

*Total is not always 879 owing to missing data.

Figure 8.9 Results from a logistic regression analysis. The dependent variable is genital chlamydia. The logistic regression model is used because it produces odds ratios for each independent risk factor - as shown here. BMJ, 315, 1997.

The regression model is one of a family of statistical methods known collectively as **multivariable models**. Other examples are proportional and Poisson regression models, which we cannot consider here. In all of these there is only one outcome variable. A second important class of methods embraces what is known as the **multivariate models**. In these models there may be more than one dependent variable. Examples are factor analysis, principal component analysis, and cluster analysis. We cannot consider these any further either.

Unit 8 Relationships between variables
Solution to examples

Q. 8.1 (a) The association appears to be positive but not particularly strong.
(b) That it is differences in levels of calcium channel blockers prescribed in various municipalities that are the cause of differences in suicide levels. In municipalities with low prescribing rates, the suicide rate is also generally low, and vice versa.

Q. 8.2 (a) That for women who gave birth less than two years ago; is positive, r = 0.896;
(b) That for women whose child was ≥ third born, r = 0.632.
(c) The most precise is that with the narrowest confidence interval, i.e. that for all gestational ages, whose 95% confidence interval is (0.817 to 0.859). The least precise is that with the widest confidence interval, i.e. that where the child is ≥ third born, whose 95% confidence interval is (0.416 to 0.781).
(d) Yes, because the correlation between maternally reported gestational age and medical record reported gestational age gets weaker as the child in question goes from being the first born (0.880), to being second born (0.815), to being ≥ third born (0.632).

Q. 8.3 (a) All are metric continuous except Family history of breast cancer in 1st or 2nd degree relative (which is nominal). All except family history are therefore appropriate for Pearson's r. The fact that the authors used Spearman's r_s suggests that they had doubts about the Normality of the distribution of the data and played safe. The Family history data does not lend itself to either Pearson's r or Spearman's r_s since this data will be nominal (yes/no). Most appropriate would be the point-biserial correlation coefficient. This is well suited to the situation where one variable is metric continuous and the other dichotomous, as here, but space permits nothing further on this.
(b) the strongest correlation of breast size is with body mass index among "Former users", r = + 0.71, p-value < 0.001; the weakest is with Waist/hip ratio among "Never users", r = 0.00, p-value = 0.985; (c) height, family history, age at menarche, waist/hip ratio, and age (all p-values ≥ 0.05); only bmi and weight have significant correlations with breast size.

Q. 8.4 (a) Correct answer is (ii). The association is definitely positive, which rules out (i) and (iii). Also it is not very strong - the points don't cluster narrowly around any imaginary straight line - this rules out (iv) which represents a very strong association.
(b) Yes, since age and % blood fat are both metric continuous - we have to assume that both are N distributed.

(c) No. Knowing that age = 35 and r = 0.425 does not enable us to predict % body fat. For that we would need a causal relationship equation.

Q. 8.5 (a) Agreement between maternally reported and clinical recorded gestational age is (i) highest for mothers whose first child this was (k = 0.67); (ii) worst for both mothers for whom the birth of the child in question was between 6-8 years ago, and for mothers whose race is recorded as Other, k = 0.42 for both.
(b) Highest correlation between maternal and clinic values of gestational age is 0.896 – where time since birth of child in question is < 2 years, although agreement is relatively good (k = 0.64), it is not the best, as we know from the answer to (a)(i) above.

Q. 8.6 The older person will on average have 0.349% more body fat than the younger. This is the same as the value of the age coefficient in the regression equation above. In other words, the size of this coefficient is the value the *dependent* variable will increase by when the *independent* variable increases by one unit of measurement. Since Age is measured in units of 1 year, its coefficient measures the increase (or decrease if the coefficient is negative) in % body fat for a one-year increase in age.

Q. 8.7 (a) All except Age; Other origin v. Western European; Primary school education only compared to university education; and Technical school education compared to university education (all p-values ≥ 0.05).
(b) We get the answer from the first row of the table, which shows that women with chronic hypertension have babies with a mean birthweight 161g lower than non-hypertensive women.
(c) On average babies with smoking mothers have a birthweight 113g lighter.
(e) The heavier woman's baby will be 8g heavier than the lighter woman's baby.