1 Introduction

SUMMARY

A very large number of clinical studies with human subjects have and are being conducted in a wide range of settings. The design and analysis of such studies demands the use of statistical models in this process. To describe such situations involves specifying the model, including defining population regression coefficients (the parameters), and then stipulating the way these are to be estimated from the data arising from the subjects (the sample) who have been recruited to the study. This chapter introduces the simple linear regression model to describe studies in which the measure made on the subjects can be assumed to be a continuous variable, the value of which is thought to depend either on a single binary or a continuous covariate measure.

Associated statistical methods are also described defining the null hypothesis, estimating means and standard deviations, comparing groups by use of a z- or t-test, confidence intervals and p-values. We give examples of how a statistical computer package facilitates the relevant analyses and also provides support for suitable graphical display.

Finally, examples from the medical and associated literature are used to illustrate the wide range of application of regression techniques: further details of some of these examples are included in later chapters.

INTRODUCTION

The aim of this book is to introduce those who are involved with medical studies whether laboratory, clinic, or population based, to the wide range of regression techniques which are pertinent to the design, analysis, and reporting of the studies concerned. Thus our intended readership is expected to range from health care professionals of all disciplines who are concerned with patient care, to those more involved with the non-clinical aspects such as medical support and research in the laboratory and beyond.

Even in the simplest of medical studies in which, for example, recording of a single feature from a series of samples taken from individual patients is made, one may ask questions as to why the resulting values differ from each other. It may be that they differ between the genders

Regression Methods for Medical Research, First Edition. Bee Choo Tai and David Machin.

and/or between the different ages of the patients concerned, or because of the severity of their illnesses. In more formal terms we examine whether or not the value of the observed variable, y, depends on one or more of the (covariate) variables, often termed the x's. Although the term covariate is used here in a generic sense, we will emphasize that individually they may play different roles in the design and hence analysis of the study of which they are a part. If one or more covariates does influence the outcome, then we are essentially claiming that part of the variation in y is a result of individual patients having different values of the x's concerned. In which case, any variation remaining after taking into consideration these covariates is termed the residual or random variation. If the covariates do not have influence, then we have not explained (strictly not explained an important part of) the variation in y by the x's. Nevertheless, there may be other covariates of which we are not aware that would.

Measurements made on human subjects rarely give exactly the same results from one occasion to the next. Even in adults, height varies a little during the course of the day. If one measures the cholesterol levels of an individual on one particular day and then again the following day, under exactly the same conditions, greater variation in this than that of height would be expected. Any variation that we cannot ascribe to one or more covariates is usually termed random variation, although, as we have indicated, it may be that an unknown covariate may account for some of this. The levels of inherent variability may be very high so that, perhaps in the circumstances where a subject has an illness, the oscillations in these measurements may disguise, at least in the early stages of treatment, the beneficial effect of treatment given to improve the condition.

STATISTICAL MODELS

Whatever the type of study, it is usually convenient to think of the underlying structure of the design in terms of a statistical model. This model encapsulates the research question we intend to formulate and ultimately answer. Once the model is specified, the object of the corresponding study (and hence the eventual analysis) is to estimate the parameters of this model as precisely as is reasonable.

Comparing two means

Suppose a study is designed to investigate the relationship between high density lipoprotein (HDL) cholesterol levels and gender. Once the study has been conducted, the observed data for each gender may be plotted in a histogram format as in Figure 1.1.

These figures illustrate a typical situation in that there is considerable variation in the value of the continuous variable HDL ranging from approximately 0.4 to 2.0 mmol/L. Further, both distributions tend to peak towards the centre of their ranges and there is a suggestion of a difference between males and females. In fact the mean value is higher at $\overline{y}_F$ = 1.2135 for the females compared with $\overline{y}_M$ = 1.0085 mmol/L for the males.

Formal comparisons between these two groups can be made using a statistical significance test. Thus, we can regard $\overline{y}_F$ and $\overline{y}_M$ as estimates of the true or population mean values μ_F and μ_M. The corresponding standard deviations are given by s_F=0.3425 and s_M=0.2881 mmol/L, and these estimate the respective population values σ_F and σ_M. To test the null hypothesis of no difference in HDL levels between males and females, the usual procedure is to assume HDL within each group has an approximately Normal distribution of the same standard deviation. The null hypothesis, of no difference in HDL levels between

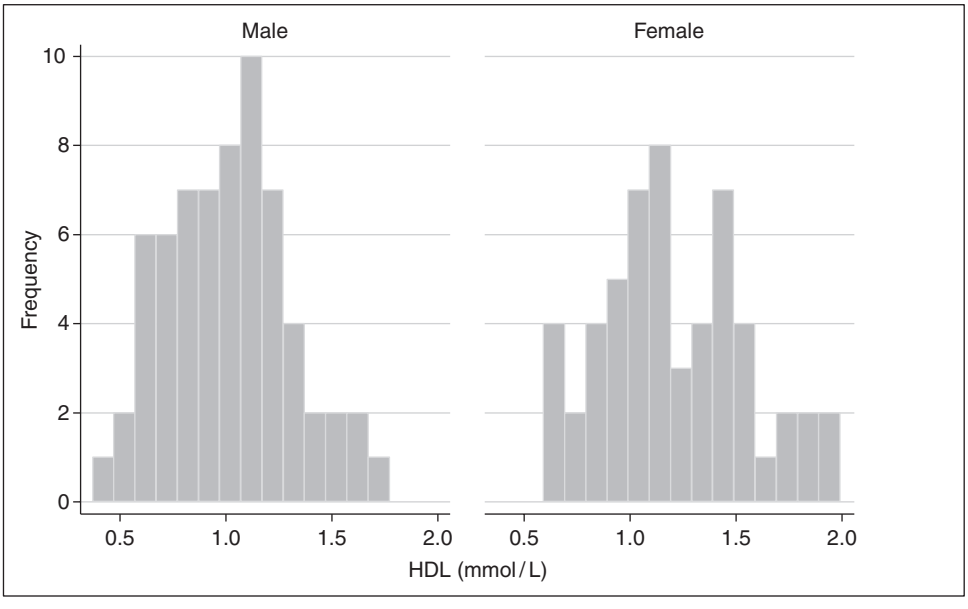


Figure 1.1 Histograms of HDL levels in 65 males and 55 females (part data from the Singapore Cardiovascular Cohort Study 2)

the sexes, is then expressed by H_0: $\mu_F = \mu_M$ or equivalently H_0: $\mu_F - \mu_M = 0$. The statistical test, the Student's *t*-test, is calculated using

$$t = \frac{(\bar{y}_F - \bar{y}_M) - (\mu_F - \mu_M)}{s_{Pool}\sqrt{\frac{1}{n_F} + \frac{1}{n_M}}} \tag{1.1}$$

where $n_F = 55$ and $n_M = 65$ are the respective sample sizes, and the expression for s_{Pool} is given in *Technical details* provided at the end of this chapter on page 16. In large samples, and when the null hypothesis is true, that is if $\mu_F - \mu_M = 0$ in equation (1.1), t has a standard Normal distribution with mean 0 and standard deviation 1.

For the data of Figure 1.1, $s_{Pool} = 0.3142$ and so, if the null hypothesis is true,

$$t = \frac{(1.2135 - 1.0085) - 0}{0.3142\sqrt{\frac{1}{55} + \frac{1}{65}}} = 3.5612 \text{ or } 3.56$$

As the sample sizes are large, to determine the statistical significance this value is referred to the standard Normal distribution of Table T1 where the notation z replaces t. The value in the table corresponding to $z = 3.56$ is 0.99981. The area in the two extremes of the distribution is the p-value $= 2\ (1 - 0.99981) = 0.00038$ in this case. This is very small probability indeed, and so on this basis we would reject the null hypothesis of no difference in HDL between the sexes. Thus, we conclude that there is a real difference in mean HDL levels between women and men estimated by 1.2135–1.0085 = 0.2050 mmol/L. These same calculations are repeated using a statistical package in Figure 1.2 in which the Student's *t*-test is activated by the command (`ttest`).

ttest hdl, by (gender)

Two-sample t test with equal variances

Group	Obs	Mean	SD
male	65	$\overline{y}_M$ = 1.0085	0.2881
female	55	$\overline{y}_F$ = 1.2135	0.3425
		s_{Pool} = 0.3142 (Pooled)	
diff		$\overline{y}_F - \overline{y}_M$ = 0.2050 95% CI 0.0910 to 0.3190	

diff = mean (female) – mean (male), t = 3.56

Ho: diff = 0, Ha: diff != 0, Pr (|T| > |t|) = 0.0005.

Note

In the main body of the text in place of Pr (|T| > |t|) = 0.0005 we use the abbreviated notation p-value = 0.0005.

Figure 1.2 Edited command and annotated output for the comparison of HDL levels between 65 males and 55 females using the *t*-test (part data from the Singapore Cardiovascular Cohort Study 2)

A feature of the computer output is that a 95% confidence interval (CI) for the true difference between the means, that is for $\delta = \mu_F - \mu_M$, is given. In this situation, but only when the total study size is large, the 100 $(1 - \alpha)$% CI for the mean difference between the male and female populations takes the form:

$$(\overline{y}_F - \overline{y}_M) - \left[z_{\alpha/2} \times SE(\overline{y}_F - \overline{y}_M)\right] \text{ to } (\overline{y}_F - \overline{y}_M) + \left[z_{\alpha/2} \times SE(\overline{y}_F - \overline{y}_M)\right] \quad (1.2)$$

where $z_{\alpha/2}$ is taken from the z-distribution and the standard error (SE) of the difference between the means is denoted by $SE(\overline{y}_F - \overline{y}_M) = \sqrt{\frac{s_F^2}{n_F} + \frac{s_M^2}{n_M}}$.

If $\alpha = 0.05$, which corresponds to $\gamma = 0.975$ in the figure of Table T1, then $z_{0.025} = 1.96$. This value is one that is highlighted in Table T2. Equation (1.2) is adapted to the situation when the sample size is small by use of equation (1.6), see *Technical details*.

The actual 95% confidence interval quoted in Figure 1.2 suggests that, although the observed difference between the means is 0.21 mmol/L, we would not be too surprised if the true difference was either as small as 0.09 or as large as 0.32 mmol/L.

Linear regression

The object of this text is to describe statistical models so we now reformulate the above example in such terms. Consider the following equation

$$HDL = \beta_0 + \beta_1 Gender \quad (1.3)$$

where we code $Gender = 0$ for males and $Gender = 1$ for females.

For the males, equation (1.3) becomes $HDL_M = \beta_0 + \beta_1 \times 0 = \beta_0$, whereas for the females $HDL_F = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$. From these, the difference between females and males is

```
regress hdl gender
------------------------------------------------------------------------------
         hdl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cons |   b0 = 1.0085
      gender |   bG = 0.2050   0.0576     3.56   0.0005    0.0910    0.3190
------------------------------------------------------------------------------
```

Figure 1.3 Edited commands and annotated output for the comparison of HDL levels between 65 males and 55 females using a regression command (part data from the Singapore Cardiovascular Cohort Study 2)

$HDL_F - HDL_M = (\beta_0 + \beta_1) - \beta_0 = \beta_1$. Thus, the difference in HDL between the sexes corresponds precisely to this single parameter or regression coefficient. If model (1.3) is fitted to the data, then the estimates of β_0 and β_1 are denoted by b_0 and b_1. Thus, we estimate the true or population difference between the sexes β_1 by the estimate b_1 obtained from the data collected in the study. In practice we may denote b_1 in such an example by either b_{Gender} or b_G to make the context clear.

The commands and output using a statistical package for this are given in Figure 1.3. This uses the command (**`regress`**) followed by the measurement concerned (**`hdl`**) and the name of the covariate within which comparisons are to be made (**`gender`**). The results replicate those of Figure 1.2, in that $b_G = 0.2050$ mmol/L exactly equals the difference between the two means obtained previously while $b_0 = \overline{y}_M = 1.0085$ mmol/L. Although this agreement is always the case, in general the approach using a regression model such as equation (1.3) is more flexible and allows more complex study designs to be analyzed efficiently.

Suppose the investigators are more interested in the relationship between HDL and body-weight of the individuals, and examine this using the scatter diagram of Figure 1.4(a). As we noted earlier, there is considerable variation in HDL ranging from 0.4 to 2.0 mmol/L. Additionally the body-weights of the individuals concerned varies from approximately 40 to 100 kg. There is a tendency for HDL to decline with increasing weight, and one objective of the study may be to quantify this in some way. This may be achieved by assuming, in the first instance, that the decline is essentially linear. In which case, a straight line may be drawn through (usually termed 'fitted to') the data in some way.

In this context, the straight line is described by the following linear equation or linear model:

$$HDL = \beta_0 + \beta_1 Weight \tag{1.4}$$

In this equation, β_0 and β_1 (in practice better denoted β_W or β_{Weight}) are constants which, once determined, position the line for the data as in the panel of Figure 1.4(b). The method of fitting the regression model to such data is given in *Technical details*. Although this fitted line suggests that there is indeed a decline in HDL with increasing weight, there are individual subjects whose HDL values are quite distant from the line. Thus, there is by no means a perfect linear relationship and hence fitting the linear model to take account of weight has not explained all the variation in HDL values between individuals. It is usual to recognize this lack-of-fit by extending the format of equation (1.4) to add a residual term, ε, so that:

$$HDL = \beta_0 + \beta_W Weight + \varepsilon \tag{1.5}$$

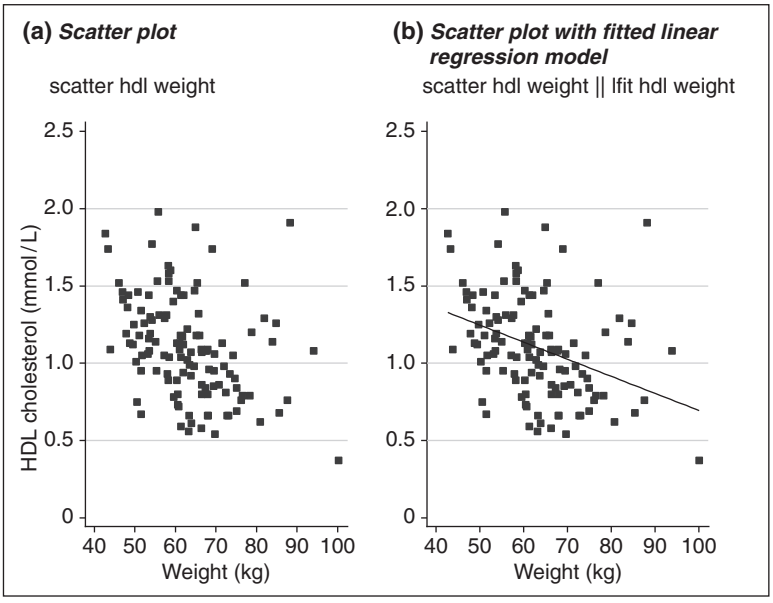


Figure 1.4 Edited command to produce (a) the scatter plot of HDL against weight, and (b) the same scatter plot with the corresponding linear regression model fitted (part data from the Singapore Cardiovascular Cohort Study 2)

Here, ε represents the residual or random variation in HDL remaining once weight has been taken into account. This noise (or error) is assumed to have a mean value of 0 across all subjects recruited to the study, and the magnitude of the variability is described by the standard deviation (SD), denoted by $\sigma_{Residual}$.

Expressed in these terms, the primary objective of the investigation will be to estimate the parameters, β_0 and β_W. In order to summarize how reliable these estimates are, we also need to estimate $\sigma_{Residual}$. Once again we write such estimates as b_0, b_W and $s_{Residual}$ to distinguish them from the corresponding parameters. For brevity, $\sigma_{Residual}$ and $s_{Residual}$ are often denoted σ and s, respectively.

The command required to fit equation (1.5) to the data is (**`regress hdl weight`**). This and the associated output are summarized in Figure 1.5. The fitting process estimates $b_0 = 1.7984$ and $b_W = -0.0110$, and so the model obtained is HDL = 1.7984 − 0.0110 *Weight*. This implies that for every 1 kg increase in weight, HDL declines on average by 0.0110 mmol/L. The p-value = 0.0001 suggests that this decline is highly statistically significant. In *Technical details*, we explain more on the Analysis of Variance (*ANOVA*) section of this output, and merely note here that $F = 18.12$ in the upper panel is very close to $t^2 = (-4.26)^2 = 18.15$ in the lower. In fact, algebraically, $F = t^2$ exactly in the situation described here—the small discrepancy is caused by rounding error in the respective calculations.

As we have seen, not all the variation in HDL has been accounted for by weight so this suggests that other features (covariates) of the individuals concerned may also influence these levels. In fact, a previous analysis suggested a difference between males and females in this respect. Figure 1.6(a) plots the data from the same 120 individuals but indicates which are male and which female. Treating these as distinct groups, then two fitted lines (one for the males and one for the females) are superimposed on the same data as illustrated in Figure 1.6(b). From this latter panel one can see that the line for males is beneath that for

regress hdl weight

Analysis of Variance (ANOVA)

Source	SS	df	MS	F	P>F
Model	1.7175	1	1.7175	18.12	0.0001
Residual	11.1819	118	0.0948		
Total	12.8994	119			

Fitted model

hdl	Coef	SE	t	P>| t |	[95% CI]
cons	$b_0 =$ 1.7984				
weight	$b_w =$ –0.0110	0.0026	–4.26	0.0001	–0.0161 to –0.0059

Note

The command (`regress hdl weight`) replaces the command (`lfit hdl weight`) which had produced the fitted line of Figure. 1.4(b) as it provides the actual estimates of the regression coefficients.

Figure 1.5 Command and annotated output for the regression of HDL on weight (part data from the Singapore Cardiovascular Cohort Study 2)

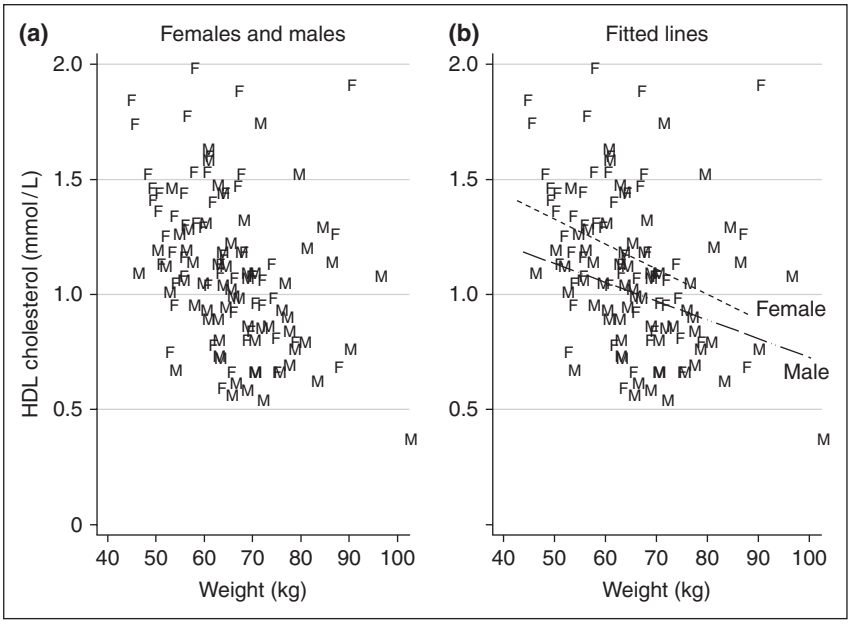


Figure 1.6 (a) Scatter plot of HDL and weight by gender in 120 individuals, and (b) the same scatter plot with linear regression lines fitted to the data for each gender separately (part data from the Singapore Cardiovascular Cohort Study 2)

females but, for both genders, the HDL declines with weight. Thus, some of the variation in HDL is accounted for by weight and some by the gender of the individuals concerned. Nevertheless, a substantial amount of the variation still remains to be accounted for.

Types of dependent variables (*y*-variables)

In the previous sections we have described models for explaining the variation in the values of HDL. Typically when using statistical models, HDL would be described as the dependent variable and, for general discussion purposes, it is usually termed the *y*-variable. However, the particular *y*-variable concerned may be one of several different data types with a label used for the variable name which is context specific. Thus, we refer to HDL rather than *y* in the above example, and note that this is a continuous variable taking non-negative values. Although we have indicated in Figure 1.1 that this may have an approximately Normal distribution form, this will not always be the case. In such a situation, a transformation of the basic variable may be considered. For a continuous variable this is often the logarithmic transformation. Thus, in a model we may consider the dependent variable as (say) $y = \log(\text{HDL})$ rather than HDL itself.

We will see below, and in later chapters, that the underlying dependent variable may also take a binary, multinomial, ordered categorical, non-negative integer or time-to-event (survival) form. In each of these situations the *y*-variable for the modeling may differ in mathematical form from that of the underlying dependent variable.

SOME COMPLETED STUDIES

As we have indicated there are countless ongoing studies, and many more have been successfully completed and reported, that will use regression techniques of one form or another for analysis. To give some indication of the range and diversity of application, we describe a selection of published medical studies which span those conducted on a small scale in the laboratory to large clinical trials and epidemiological studies. These examples include some features that we also draw upon in later chapters.

Example 1.1 Linear regression: Interferon-λ production in asthma exacerbations

Busse, Lemanske Jr and Gern (2010, Figure 6) reproduce a plot of the percentage reduction in FEV_1 in individuals with asthma and in healthy volunteers against their generation of interferon-λ. These data were first described by Contoli, Message, Laza-Stanca, *et al*. (2006, Figure 2f) who state: 'Induction of IFN-λ protein by rhinovirus in BAL cells is strongly related to severity of reductions in lung function on subsequent *in vivo* rhinovirus experimental infection. IFN-λ protein production in BAL cells infected *in vitro* with RV16 was significantly inversely correlated with severity of maximal reduction from baseline in FEV_1 (forced expiratory volume in 1 s) recorded over the 2-week infection period when subjects were subsequently experimentally infected with RV16 *in vivo* ($r = 0.65$, $P < 0.03$).' Their results with the information extracted from Busse, Lemanske Jr and Gern (2010, Figure 6) are reproduced in Figure 1.7.

Expressed in terms of a regression model, using the command (**`regress FEV1 Interpgml`**), the increasing slope of $b_{IFN\text{-}\lambda} = 0.1011$ per unit increase in pg/mL is statistically significant, p-value $= 0.017$.

We will return to this example in Chapter 2 but note here that, as there are two groups of individuals concerned (those with asthma and healthy individuals), the potential influence of this second covariate (type of subject) in addition to IFN-λ protein production must be considered.

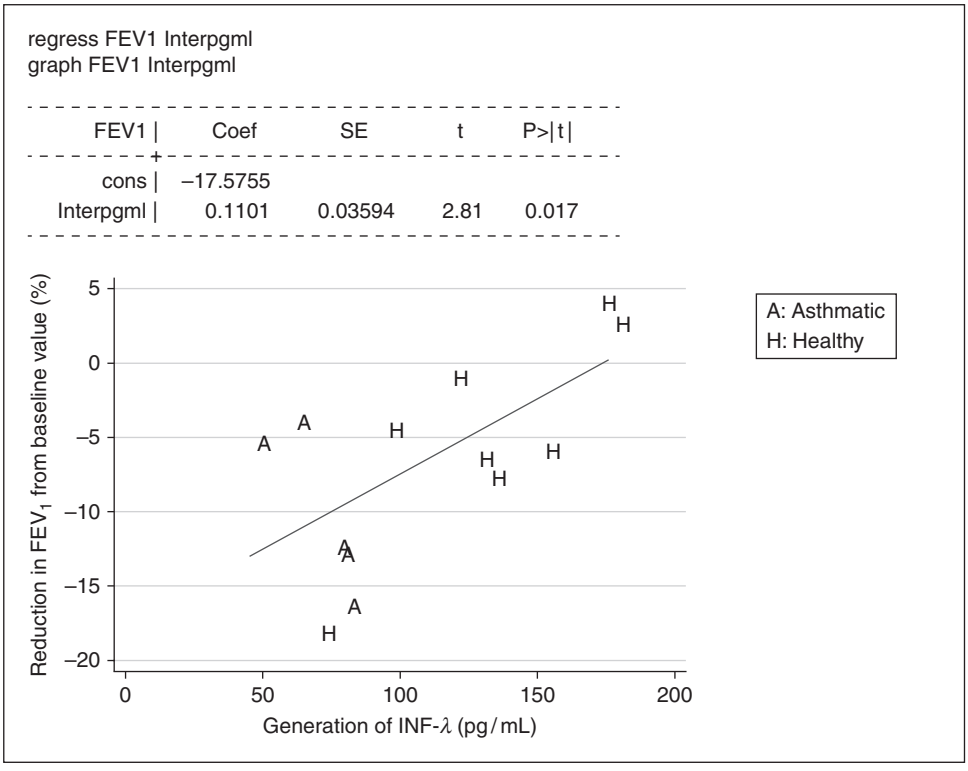


Figure 1.7 Annotated commands and output to investigate the response to human rhinovirus infection in eight healthy individuals and five with asthma (information from Busse, Lemanske Jr and Gern, 2010, Figure 6)

	Units	Estimated regression coefficient	95% CI	*p*-value
Design variable				
TNF-α score	(ng/L)	0.20	0.13 to 0.27	< 0.001
Other covariates				
log(triacylglycerol)	log(mmol/L)	0.15	0.01 to 0.30	0.041
MAP	mmHg	0.01	0.002 to 0.20	0.011
Duration of diabetes	years	0.02	0.01 to 0.02	< 0.001
Total cholesterol	mmol/L	0.12	0.04 to 0.21	0.006

Figure 1.8 Association between log(ACR) and inflammatory variables in a multiple regression analysis (after Ng, Fukushima, Tai, *et al.*, 2008, Table 2)

Example 1.2 Multiple linear regression: Activation of the TNF-α system in patients with diabetes

Ng, Fukushima, Tai, *et al.* (2008) investigated whether activation of the TNF-α system may potentially exert an effect on the albumin:creatinine ratio (ACR), expressed in g/kg, in patients with type 2 diabetes. In a multiple regression equation summarized in Figure 1.8, they used the logarithm of ACR, that is log(ACR), as the *y*-variable, TNF-α score as the key covariate and

log(triacylglycerol), mean arterial pressure (MAP), duration of diabetes and total cholesterol as the other covariates. They concluded that: '… log(ACR) was significantly associated with TNF-α score, with a unit change in TNF-α score resulting in a 0.20 unit change in log(ACR) … .'

As we have noted, $y = \log(\text{ACR})$ rather than ACR itself was used as the dependent variable. Further, there was a principal covariate, TNF-α score, and four potentially influencing covariates: log(triacylglycerol), MAP, duration of diabetes and total cholesterol. In fact Ng, Fukushima, Tai, *et al*. (2008, Table 1) recorded a total of 18 potential covariates. Most of these were screened out as not influencing values of log(ACR) using a variable selection process (see Chapter 7) to leave the final model to summarize the results containing only the five covariates listed in Figure 1.8. In situations such as this when there is a principal or design covariate specified, then reporting details of the simple linear regression, here log(ACR) on TNF-α score without adjustment by the other covariates, is recommended. This information then enables the reader to judge how much the presence of the other (here four) covariates in the full model influence the magnitude of that regression coefficient (here $\beta_{TNF\text{-}\alpha}$) of principal interest.

Example 1.3 Multiple logistic regression: Intrahepatic vein fetal blood samples

Figure 1.9 includes part of the data collated by Chinnaiya, Venkat, Chia, *et al*. (1998, Table 6) giving the number of fetal deaths according to different puncture sites chosen for sampling from both normal and abnormal fetuses. Here the event of concern is a fetal death at some stage following the fetal blood sampling of whatever type. This is clearly a binary 0 (alive), 1 (dead) variable. There were 52 fetal deaths among the 292 sampled using the intrahepatic vein (IHV) technique: an odds of 52:240. For percutaneous umbilical cord sampling (PUBS), the odds were 20:50. Comparing the two fetal sampling techniques gives an odds ratio: $OR = \dfrac{20/50}{52/240} = 1.85$, suggesting a greater death rate using PUBS. An even greater

Fetal loss	Intrahepatic vein (IHV)	Percutaneous umbilical cord sampling (PUBS)	Cardiocentesis	Total (%)
< 2 weeks of procedure	21	7	11	39 (10.2)
> 2 weeks of procedure	10	6	2	18 (4.7)
Postnatal death	21	7	2	30 (7.9)
Total fetal deaths (%)	52 (17.8)	20 (28.6)	15 (75.0)	87 (22.8)
Live-births	240	50	5	295
Total sampled	292	70	20	382
Odds	52:240	20:50	15:5	87:295
Odds Ratio (*OR*)	1	1.85	13.85	
(95% CI)	–	1.01 to 3.36	4.82 to 39.79	

Figure 1.9 Fetal loss following fetal blood sampling according to different puncture sites in both normal and abnormal fetuses (Source: Chinnaiya, Venkat, Chia, *et al.*, 1998, Table 6. Reproduced with permission of John Wiley & Sons Ltd.)

risk is apparent when cardiocentesis is used with $OR = \frac{15/5}{52/240} = 13.85$ when compared with IHV.

In Figure 1.9 more detail of when the deaths occurred is given so that the *y*-variable of interest may be the ordered (4 – level) categorical variable fetal loss and live-birth rather than the binary variable death or live-birth. As loss following blood sampling may be influenced by gestational age of the fetus as well as clinical indications including pre-term rupture of membranes, hydrops fetalis, and the number of needle entries made, then these variables may need to be accounted for in a full analysis.

Example 1.4 Poisson regression: Hospital admissions for chronic obstructive pulmonary disease (COPD)

Maheswaran, Pearson, Hoysal and Campbell (2010) evaluated the impact of a health forecast alert service on admissions for chronic obstructive pulmonary disease (COPD) in the Bradford and Airedale region of England. Essentially, the UK Meteorological Office (UKMO) provides an alert service which forecasts when the outdoor environment is likely to adversely affect the health of COPD patients. This alert enables the patients to take appropriate action to keep themselves well, and thereby potentially avoid a hospital admission. In brief, general practitioner (GP) groups providing primary medical care chose to participate or not in the evaluation, and those that did registered their COPD patients with the UKMO. Registered patients were given an information pack which included details about the automated telephone call they would receive should bad weather trigger an alert. The number of hospital admissions was subsequently noted over the two winter periods 2006–7 and 2007–8. A summary of the study findings is given in Figure 1.10.

Winter (December to March)	Number of GP practices	Admissions 2006–7	Admissions 2007–8	Ratio of Admissions (2007–8) / (2006–7) *R*	Ratio of Admissions (2007–8) / (2006–7) 95% CI
Exposure category					
None	51	203	195	0.96	0.79 to 1.18
Part winters	10	57	55	0.96	0.65 to 1.42
Whole winters	25	183	172	0.94	0.76 to 1.16
Total	86	443	422		
Adjusted for exposure category			$R_{Adjusted\ Category} = 0.98$		0.78 to 1.22
Exposure scale					
0	51	203	195	0.96	0.79 to 1.18
0.01 – 0.2	7	42	35	0.83	0.52 to 1.34
0.21 – 0.4	10	98	82	0.84	0.62 to 1.13
0.41 – 0.6	3	17	27	1.59	0.83 to 3.11
0.61 – 0.8	7	42	42	1.00	0.64 to 1.57
> 0.8	7	36	38	1.06	0.65 to 1.71
Total	85	438	419		
Adjusted for exposure scale			$R_{Adjusted\ Scale} = 1.11$		0.80 to 1.52

Figure 1.10 Admissions for chronic obstructive pulmonary (COPD) disease in Bradford and Airedale, England by category of general practice (GP) exposure to the Meteorological Office Forecast Alert Service (Source: Maheswaran, Pearson, Hoysal and Campbell, 2010, Table 1. Reproduced with permission of Oxford University Press.)

The GP practices concerned each comprise a very large number of patients, so that among them COPD represents a rare event. In which case, the unit for analysis is the number of hospital admissions, *h*, from each practice rather than the ratio of this number to the size of the practice concerned. As a consequence Poisson regression methods (see Chapter 5) were used for analysis. These models took account of the GP practice concerned as an unordered categorical variable, the particular winter (2006–7 or 2007–8) as binary, and either the exposure category or the exposure scale, both of which were treated as ordered numerical variables with equal category divisions. In Figure 1.10 the admission rate ratio adjusted for the exposure category, $R_{Adjusted\ Category}$=0.98 (95% CI 0.78 to 1.22), implies that admissions in 2007–8 were 2% lower in GP practices that participated relative to practices that did not. In contrast, the admission rate ratio adjusted for the exposure scale, $R_{Adjusted\ Scale}$=1.11 (95% CI 0.80 to 1.52) implies that admissions in 2007-8 were 11% higher in GP practices that participated and entered all their COPD patients into the forecasting system, relative to practices that did not. The wide confidence intervals, both of which cover the null hypothesis ratio of unity, indicate that the value or otherwise of the warning system has not been clearly established.

This study provides an example of a multi-level or clustered design in that the GPs agree to participate in the study but it is the admission to hospital or not of their individual patients that provides the outcome data. However, patients treated by one health care professional tend to be more similar among themselves than those treated by a different health care professional. So, if we know which GP is treating a patient, we can predict, by reference to experience from other patients, slightly better than chance, the outcome for the patient concerned. Consequently the patient outcomes for one GP are positively correlated and so are not completely independent of each other. Due note of the magnitude of this intra-cluster correlation (ICC) is required in the design and analysis processes. Multi-level designs are discussed further in Chapter 11.

Example 1.5 Cox proportional hazards regression: Nasopharyngeal cancer

Wee, Tan, Tai, *et al*. (2005) conducted a randomized trial of radiotherapy (RT) versus concurrent chemo-radiotherapy followed by adjuvant chemotherapy (CRT) in patients with nasopharyngeal cancer. The trial recruited 221 patients, 110 of whom were randomized to receive RT (the standard approach) and 111 CRT. The Kaplan-Meier estimates of the overall survival times of the patients in the two groups are given in Figure 1.11(a). The estimated hazard ratio (*HR*) of 0.50, calculated using a Cox proportional hazards regression model (see Chapter 6), indicates a survival advantage to those receiving CRT.

For nasopharyngeal cancer patients it is well known that, for example, their nodal status at the time of diagnosis has considerable influence on their ultimate survival. This is shown for those recruited to this trial in Figure 1.11(b) by a hazard ratio, *HR*=0.55, which indicates considerable additional risk for those with N3 nodal status. However, despite nodal status being an important influence on subsequent prognosis, a Cox proportional hazards regression model including the treatment received and nodal status gave a *HR*=0.51 (95% CI 0.31 to 0.85, *p*-value=0.009) in favor of CRT. All elements of which are very close to the respective components of the caption included within Figure 1.11(a). This suggests that the benefit of CRT over RT remains for all patients irrespective of their nodal status. In a final report it is not important to show Figure 1.11(b) as the considerable risk associated with N3 nodal status was known at the design stage of the trial. This is why the 95% CI and *p*-value are omitted from the caption.

In general, despite a covariate being a strong predictor of outcome, adding this to the regression model may not substantially change the estimated value of the 'key' regression coefficient, which in the case just discussed is that corresponding to the randomized

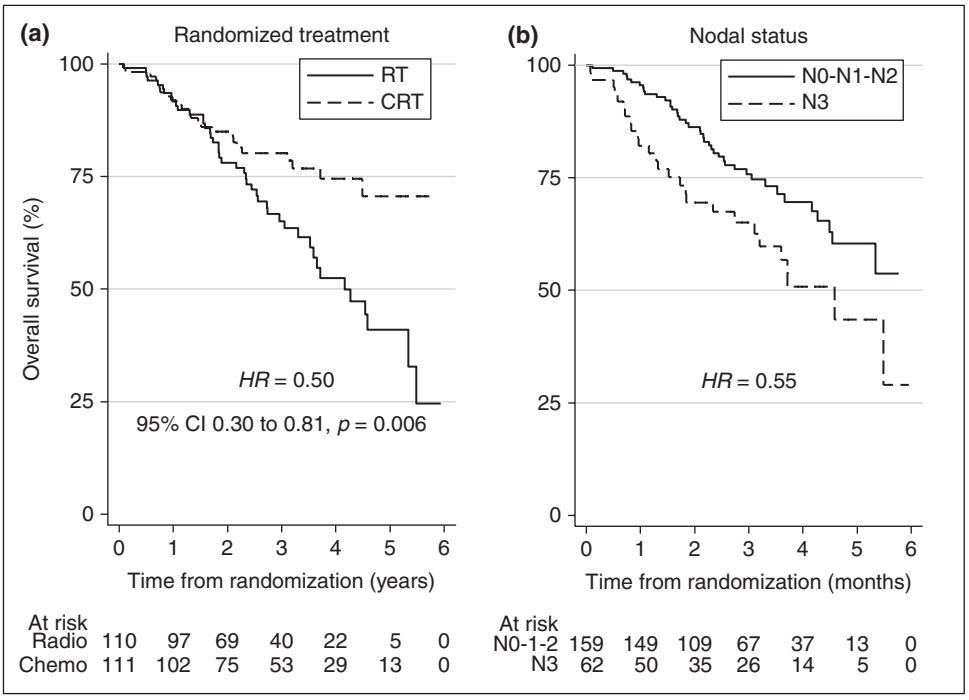


Figure 1.11 Kaplan-Meier estimates of the overall survival of patients with nasopharyngeal cancer by (a) randomized treatment received, and (b) nodal status at diagnosis (data from Wee, Tan, Tai *et al.*, 2005)

treatment given. Thus, the main concern here is the measure of treatment difference, and the role of the covariate is to see whether or not our view of this measure is modified by taking note of its presence. The aim in this situation is not to study the influence of the covariate itself.

Example 1.6 Repeated measures: Pain assessment in patients with oral lichen planus

Figure 1.12 illustrates the results from a longitudinal repeated measures design in which the pain experienced by patients with oral lichen planus (OLP) is self-recorded over time. The patients were recruited to a randomized trial conducted by Poon, Goh, Kim, *et al.* (2006) comparing the efficacy of topical steroid with topical cyclosporine for healing their OLP. In general, pain levels diminished over a 12-week period but with little difference between the treatments observed. The plots illustrate some of the difficulties with such trials. For example, although a fixed assessment schedule was described in the protocol, practical circumstances dictated that these were not strictly adhered to. Also the number of patients returning to the clinics involved declined as the period from initial diagnosis and randomization increased. Typically there are a large number of data items and considerable variation in pain levels recorded both in the same patient and between different patients. In addition, successive data points within the same patient are unlikely to be independent of each other. The use of fractional polynomials (see Chapter 11) allows flexible (not just linear) regression models to be used to describe such data.

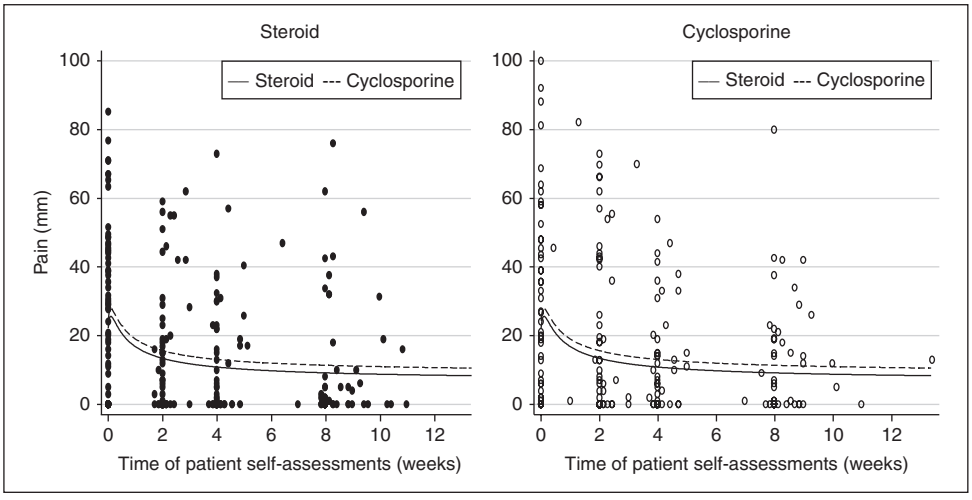


Figure 1.12 Visual Analog Scale (VAS) recordings of pain experience from OLP by treatment group. The fractional polynomial curve for the comparator treatment is added to each panel to facilitate visual comparisons between treatments (Source: Poon, Goh, Kim, *et al.*, 2006, Fig. 2. Reproduced with permission of Elsevier.)

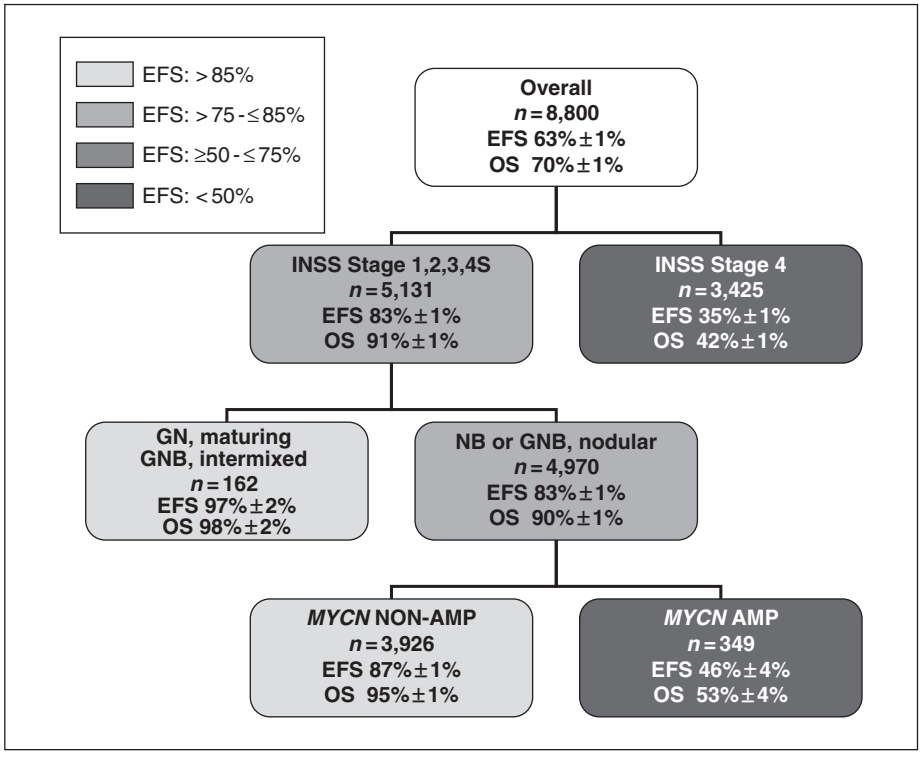


Figure 1.13 Section of the regression tree analysis of 5-year event-free (EFS) and overall (OS) survival of 8,800 young patients with neuroblastoma (Source: Cohn, Pearson, London, et al., 2009, Fig. 1A. Reproduced with permission of Springer.)

Example 1.7 Regression trees: International Neuroblastoma Risk Group

Cohn, Pearson, London, *et al*. (2009) used a regression tree approach to develop an international neuroblastoma risk group (INRG) classification system for young patients diagnosed with neuroblastoma (NB). Figure 1.13 shows their 'top-level' split which divides the 8800 patients using the International Neuroblastoma Staging System (INSS) into two quite distinct prognostic groups (Stage 1, 2, 3, and 4S versus Stage 4) with, for example, event-free survival (EFS) of 83% and 35% respectively, at five years from diagnosis. Two branches follow the first of these groups; one for the small group comprising ganglioneuroma (GN), maturing ganglioneuroblastoma (GNB), and intermixed types (EFS 97%) which is a terminal node or leaf; the other for those with neuroblastoma (NB) or GNB nodular (EFS 83%). This latter group is then further branched into those with MYCN non-amplified (EFS 87%) and amplified status (EFS 46%), which are then subsequently divided again, as is the intermediate node of INSS Stage 4 patients, until terminal nodes are reached. This branching process identified a total of 20 terminal nodes of homogeneous patient groups ranging in size from 8 to 513 with a median size of 59, and comprising 5-year EFS rates ranging from 19% to 97%, with a median of 61%.

Further reading

A general introduction to medical statistics is Swinscow and Campbell (2002), which concentrates mainly on the analysis of studies, while Bland (2000), Campbell (2006) and Campbell, Machin and Walters (2007) are intermediate texts. Altman (1991) and Armitage, Berry and Matthews (2002) give lengthier and more detailed accounts. All these books cover the topic of regression models to some extent, while Mitchell (2012) focuses specifically on using Stata for regression modeling purposes.

Machin and Campbell (2005) focus on the design, rather than analysis, of medical studies in general while Machin, Campbell, Tan and Tan (2009) specifically cover sample size issues. A useful text is that of Freeman, Walters and Campbell (2008) on how to display data.

More advanced texts which address aspects of regression models specifically include Clayton and Hills (1993), Dobson and Barnett (2008), Everitt and Rabe-Hesketh (2006) and Kleinbaum, Kupper, Muller and Nizam (2007). Collett (2002) specifically addresses the modeling of binary data, while Collett (2003) and Machin, Cheung and Parmar (2006) focus on time-to-event models. Diggle, Liang and Zeger (1994) cover the analysis of repeated measures data in detail as does Rabe-Hesketh and Skrondal (2008).

Altman DG (1991). *Practical Statistics for Medical Research*. London, Chapman and Hall.

Armitage P, Berry G and Matthews JNS (2002). *Statistical Methods in Medical Research*. (4th edn). Blackwell Science, Oxford.

Bland M (2000). *An Introduction to Medical Statistics*. (3rd edn). Oxford University Press, Oxford.

Campbell MJ (2006). *Statistics at Square Two: Understanding Modern Statistical Applications in Medicine*. (2nd edn). Blackwell BMJ Books, Oxford.

Campbell MJ, Machin D and Walters SJ (2007). *Medical Statistics: A Commonsense Approach: A Text Book for the Health Sciences*, (4th edn) Wiley, Chichester.

Clayton D and Hills M (1993). *Statistical Models in Epidemiology*, Oxford University Press, Oxford.

Collett D (2002). *Modelling Binary Data*, (2nd edn) Chapman and Hall/CRC, London.

Collett D (2003). *Modelling Survival Data in Medical Research*, (2nd edn), Chapman and Hall/CRC, London.

Diggle PJ, Liang K-Y and Zeger SL (1994). *Analysis of Longitudinal Data*. Oxford Science Publications, Oxford.

Dobson AJ and Barnett AG (2008). *Introduction to Generalized Linear Models*. (3rd edn), Chapman and Hall/CRC, London.

Everitt BS and Rabe-Hesketh S (2006). *A Handbook of Statistical Analysis using Stata*. (4th edn), Chapman and Hall/CRC, London.
Freeman JV, Walters SJ and Campbell MJ (2008). *How to Display Data*. BMJ Books, Blackwell Publishing, Oxford.
Kleinbaum G, Kupper LL, Muller KE and Nizam E (2007). *Applied Regression Analysis and Other Multivariable Methods*. (4th edn), Duxbury Press, Florence, Kentucky.
Machin D and Campbell MJ (2005). *Design of Studies for Medical Research*. Wiley, Chichester.
Machin D, Campbell MJ, Tan SB and Tan SH (2009). *Sample Size Tables for Clinical Studies*. (3rd edn). Wiley-Blackwell, Chichester.
Machin D, Cheung Y-B and Parmar MKB (2006). *Survival Analysis: A Practical Approach*. (2nd edn). Wiley, Chichester.
Mitchell MN (2012). *Interpreting and Visualizing Regression Models Using Stata*. Stata Press, College Station, TX.
Rabe-Hesketh S and Skrondal A (2008). *Multilevel and Longitudinal Modeling Using Stata (2nd edn)*, Stata Press, College Station, TX.
Swinscow TV and Campbell MJ (2002). *Statistics at Square One*. (10th edn), Blackwell, BMJ Books, Oxford.

TECHNICAL DETAILS

Student's *t*-test

For a given set of continuous data values $y_1, y_2, \ldots, y_n$ observed from n patients, the mean, $\overline{y}$, and standard deviation, s, are calculated as follows: $\overline{y} = \dfrac{\sum y_i}{n}$ and $s = \sqrt{\dfrac{\sum (y_i - \overline{y})^2}{n-1}}$. These provide estimates of the associated parameters μ and σ.

For the data from the n_M males and n_F females of Figure 1.1, each gender provides a mean and standard deviation which we denote as $\overline{y}_M$, s_M and $\overline{y}_F$, s_F, respectively. To make a statistical comparison between these groups, it is often assumed that s_M and s_F are each estimating a common standard deviation σ estimated by s_{Pool}. In which case the two estimates s_M and s_F are combined as follows: $s_{Pool} = \sqrt{\dfrac{(n_M - 1)s_M^2 + (n_F - 1)s_F^2}{(n_M - 1) + (n_F - 1)}}$. For the data concerned $s_{Pool} = \sqrt{\dfrac{(55-1)0.3425^2 + (65-1)0.2881^2}{(55-1)+(65-1)}} = 0.3142$. This is then used in equation (1.1), which we repeat here for convenience: $t = \dfrac{(\overline{y}_F - \overline{y}_M) - (\mu_F - \mu_M)}{s_{Pool}\sqrt{\dfrac{1}{n_F} + \dfrac{1}{n_M}}}$.

As we noted previously, if the sample sizes n_M and n_F are sufficiently large then, under the assumption that the null hypothesis is true, that is H_0: $(\mu_F - \mu_M) = 0$, t can be regarded as having a Normal distribution with mean 0 and standard deviation 1. Once the value of t is calculated from the data, this can be referred to Table T1 to obtain the corresponding p-value.

In circumstances when the sample sizes n_M and n_F are not large, or there is some concern as to whether they are large enough, then use of Table T4, rather than Table T4, is necessary to obtain the p-value. Table T4 requires the degrees of freedom, df, which in this situation is, $df = (n_F - 1) + (n_M - 1) = n_F + n_M - 2$. In the above example, $df = (55 - 1) + (65 - 1) = 118$ is large and corresponds to the ∞ (infinity) row of Table T4. This row suggests that for a p-value to be less than $\alpha = 0.05$ would require t to exceed 1.960. Had the degrees of freedom

been smaller, say $df=18$, then for a p-value to be less than $\alpha=0.05$ would require, from Table T4, that t exceeds 2.101. In general, as the degrees of freedom decline, more extreme values are required for t than z in order to declare statistical significance.

In the situation of small degrees of freedom, the confidence interval (CI) of equation (1.2) for the difference between two means is modified to become:

$$(\overline{y}_F - \overline{y}_M) - [t_{df,\alpha/2} \times SE(\overline{y}_F - \overline{y}_M)] \text{ to } (\overline{y}_F - \overline{y}_M) + [t_{df,\alpha/2} \times SE(\overline{y}_F - \overline{y}_M)] \quad (1.6)$$

Thus $z_{\alpha/2}$ is replaced by $t_{df,\alpha/2}$ and the necessary values are taken from Table T4 of the t-distribution with the appropriate degrees of freedom.

Linear regression

In general terms, the linear regression models of equations (1.3) and (1.4) are expressed for individual 'i' in the study as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (1.7)$$

where y_i is termed the dependent variable and x_i the independent variable or covariate.

Although y is a continuous variable in the case of HDL, in other circumstances the dependent variable concerned may be binary, ordered categorical, non-negative integer or a time-to-event (survival). The covariates gender and weight are respectively binary and continuous but other types of covariates such as categorical or ordered categorical variables may be involved.

In general the data of, for example Figure 1.4(a), can be regarded as a set of n pairs of observations (x_1, y_1), (x_2, y_2), … , (x_n, y_n). To fit the linear model to the data when y is continuous requires estimates of the regression coefficients β_0 and β_1, which are given by

$$b_0 = \overline{y} - b_1 \overline{x} \text{ and } b_1 = \frac{S_{xy}}{S_{xx}} \quad (1.8)$$

where $S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y})$ and $S_{xx} = \sum (x_i - \overline{x})^2$. In addition, the residual standard deviation, $\sigma_{Residual}$ (more briefly σ) is estimated by

$$s_{Residual} = \sqrt{\frac{S_{yy} - b_1^2 S_{xx}}{(n-2)}} \quad (1.9)$$

where $S_{yy} = \sum (y_i - \overline{y})^2$. Thus, with the three equations included in (1.8) and (1.9), we have estimates b_0, b_1, and $s_{Residual}$ (or s) for the corresponding unknown parameters, β_0, β_1, and σ.

Once we have the estimates for b_0 and b_1 then, for any subject i with covariate value x_i, we could predict their y_i by $Y_i = b_0 + b_1 x_i$. Clearly, in choosing b_0 and b_1 we wish that the resulting Y_i will be close to the observed y_i and hence make our prediction error $(y_i - Y_i)$ as small as possible. The estimates of equation (1.8) result from choosing values of b_0 and b_1 which minimize the sum of squares, $\Sigma (y_i - Y_i)^2$. This leads b_0 and b_1 to be termed the *ordinary least-squares* (OLS) estimates of the *population* parameters β_0 and β_1.

The model of equation (1.7) once fitted to the data gives for the *i*th individual $y_i = b_0 + b_1 x_i + e_i = Y_i + e_i$ where the observed residual, e_i is the estimate of the true residual, ε_i. The residuals are the amount that the observed value differs from that predicted by the model, and represent (in this case) the variation not explained after fitting a straight line to the data. The e_i from all n subjects are usually assumed to have a Normal distribution with an average value of zero.

Once the problem under study is expressed by means of a statistical model, then the null hypothesis is expressed as, for example, H_0: $\beta_1 = 0$. If the null hypothesis were indeed true, then this implies that the covariate x in equation (1.7) explains none (strictly at most an unimportant amount) of the variation in y.

To test whether β_1 is significantly different from zero, the appropriate t-test is:

$$t = \frac{b_1 - 0}{SE(b_1)} \tag{1.10}$$

where $SE(b_1) = s_{Residual} / \sqrt{S_{xx}}$.

We note from Figure 1.3 that, when using (**`regress hdl gender`**) to test the null hypothesis of no difference in means between the genders, $b_G = 0.2050$ and this has standard error, $SE(b_G) = 0.05756$. In this case the degrees of freedom, $df = n - 2$; the '2' appears here as it is necessary to estimate 'two' parameters, β_0 and β_G, in order to fit the linear regression line. Finally, the calculated value of t is referred to Table T4 to assess the statistical significance, although in practice the p-value will usually be an integral part of the computer output.

The associated expression for $100(1 - \alpha)\%$ CI for the estimate of the regression slope β_1 is

$$b_1 - [t_{df,\alpha/2} \times SE(b_1)] \text{ to } b_1 + [t_{df,\alpha/2} \times SE(b_1)] \tag{1.11}$$

where $t_{df,\alpha/2}$ is replaced by $z_{\alpha/2}$ in large sample situations. In practice, most computer software packages will adjust for sample size automatically.

Predicting a mean value of *y* for a particular *x*

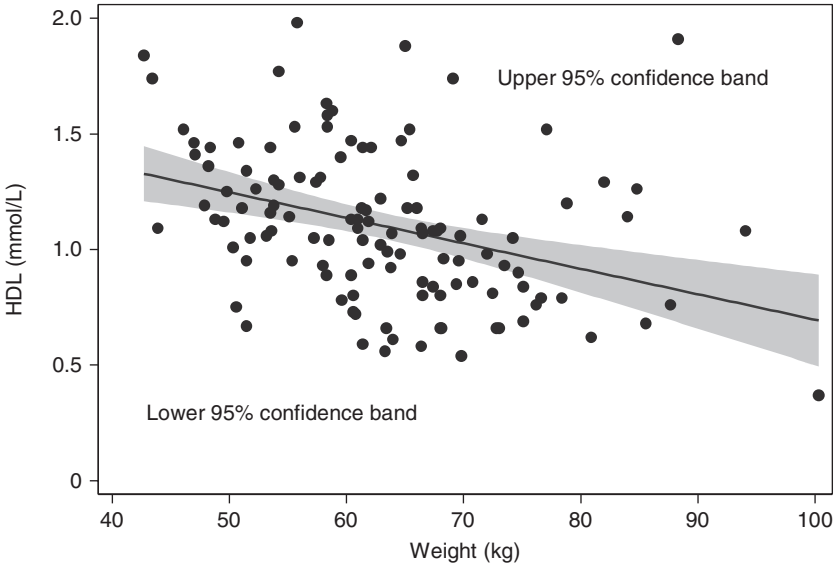
For a specific value of an individual's weight $x = x_0$ (say) of Figure 1.5, we can obtain the estimated *mean* value of HDL for all individuals of that weight as $Y_0^{Mean} = b_0 + b_1 x_0$. This can be shown to have a standard error given by $SE(Y_0^{Mean}) = s_{Residual} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$, the size of which will increase as x_0 makes an increasing departure from $\bar{x}$. This expression can then be used to calculate the $100(1 - \alpha)\%$ CI about the regression line at x_0 given by

$$Y_0^{Mean} - [t_{df,\alpha/2} \times SE(Y_0^{Mean})] \text{ to } Y_0^{Mean} + [t_{df,\alpha/2} \times SE(Y_0^{Mean})] \tag{1.12}$$

The resulting 95% confidence band for the regression of HDL on weight, calculated by evaluating (1.12) over the whole range of the covariate, is given in Figure 1.14(a). It

(a) ***95% confidence band***

```
twoway lfitci hdl weight, stdp
||scatter hdl weight
```



(b) ***ANOVA***

Source	SS	df	MS	F
Model	$SS_{Model} = b_1^2 S_{xx}$	$df_{Model} = 1$	$\frac{b_1^2 S_{xx}}{1}$	$\frac{b_1^2 S_{xx}}{s^2_{Residual}}$
Resid	$SS_{Residual} = S_{yy} - b_1^2 S_{xx}$	$df_{Residual} = (n-2)$	$s^2_{Residual} = \frac{\left[S_{yy} - b_1^2 S_{xx} \right]}{n-2}$	
Total	$SS_{Total} = S_{yy}$	$df_{Total} = (n-1)$		

Source	SS	df	MS	F	P>F
Model	1.7175	1	1.7175	18.12	0.0001
Resid	11.1819	118	0.0948		
Total	12.8994	119			

Figure 1.14 (a) Fitted linear regression line of HDL on weight with the 95% confidence band of the predicted mean values indicated by the shaded area. (b) Analysis of Variance (ANOVA) and associated output to check the goodness-of-fit of the simple linear regression model (part data from the Singapore Cardiovascular Cohort Study 2)

should be noted that the intervals bend away from the fitted line as weight moves away from about 65 kg (actually from the mean value 63.15 kg of these subjects) in either direction.

Predicting an individual's value of *y* for a particular *x*

In contrast to wanting to estimate the *mean* value of y for a particular x_0, we may wish to *predict* the HDL of an *individual* subject with that weight. Although the corresponding $Y_0^{Individual}$ is the same as the predicted mean $Y_0^{Mean} = b_0 + b_1 x_0$ in this situation, we do not require $SE\left(Y_0^{Mean}\right)$ but rather the standard deviation, $SD\left(Y_0^{Individual}\right)$ of the individual values $Y_0^{Individual}$ at that particular value of $x = x_0$. This is given by $SD(Y_0^{Individual}) = s_{Residual}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$. Thus, the distribution of individuals with weight x_0 has a mean value of $Y_0^{Individual}$ and (in broad terms) 95% of those of weight x_0 will have a HDL value which will be within the interval

$$Y_0^{Individual} - 1.96 \times SD(Y_0^{Individual}) \tag{1.13}$$

The difference between (1.12) and (1.13) lies in the use of the predicted Y_0. One purpose is to ask: What is the anticipated value of HDL for (say) a patient with weight $x_0 = 50$ kg and how precise is this estimate? The estimate is $Y_0 = 1.7984 - 0.0110 \times 50 = 1.25$ mmol/L and the 95% CI can be calculated using $SE(Y_0^{Mean})$ in equation (1.12) to give 1.16 to 1.34 mmol/L. The other is to ask: What variation about $Y_0 = 1.25$ mmol/L does one expect to see in different patients of the same 50 kg weight? The answer is provided by evaluating equation (1.13) to give the wider interval 0.64 to 1.86 mmol/L.

Analysis of Variance (ANOVA)

As we have seen in Figure 1.5, the ANOVA part of the computer output following the command (**`regress hdl weight`**) takes the form of Figure 1.14(b), although we have annotated this with some terminology that we have just introduced.

Essentially, ANOVA divides the total sums of squares $SS_{Total} = \sum (y_i - \overline{y})^2$, which represents the total variation of the dependent y-variable for all the subjects concerned, into a part which is explained by the linear model, $SS_{Model} = b_1^2 \sum (x - \overline{x})^2 = b_1^2 S_{xx}$, and that which remains, $SS_{Residual} = S_{yy} - b_1^2 S_{xx}$. The latter is described as the random variation component. If the null hypothesis is true, that is $\beta_1 = 0$, we would then anticipate that b_1 will be close to zero. In this situation $SS_{Model} = b_1^2 S_{xx}$ will be small. To calculate the F-statistic, the sum of squares (SS) for the model and for the residual are each divided by the corresponding degrees of freedom and their ratio obtained. Thus,

$$F = \frac{b_1^2 S_{xx} / df_{Model}}{(S_{yy} - b_1^2 S_{xx}) / df_{Residual}} = \frac{b_1^2 S_{xx} / df_{Model}}{s^2_{Residual}} \tag{1.14}$$

To ascertain the corresponding p-value, statistical tables of the F-distribution are required but these are potentially very extensive as they depend on two sets of degrees of freedom, $df_1 = df_{Model}$ and $df_2 = df_{Residual}$. Thus, only limited tabular entries are reproduced in Table T6. In the example of Figure 1.14(b), $F = \dfrac{1.7175/1}{11.1819/118} = \dfrac{1.7175}{0.0948} = 18.12$ and this has 1 and 118 degrees of freedom. As $df_2 = 118$ is large the tabular entry of infinity (∞) is used, so that with $df_1 = 1$ a value of $F = 6.66$ corresponds to $\alpha = 0.01$ so that we can conclude, as $18.12 > 6.66$, that the p-value < 0.01. However, most statistical packages output the p-value so that the statistical tables are usually unnecessary. In this case a more precise p-value $= 0.0001$ is presented.

In all situations when $df_1 = 1$, that is when only two groups are being compared, the F-test and the t-test of the null hypothesis will give precisely the same p-value. This is because, and only in this situation, the calculated F and the square of the calculated t, that is t^2, are algebraically equivalent. The corresponding t-test will have degrees of freedom equal to df_2 of the F-test. Thus, in the above example, $t = \sqrt{18.12} = 4.26$ with $df = 118$ and Table T4 can therefore in principle be used to obtain the p-value. However, as the degrees of freedom are larger than 30 there is no tabular entry available. All we can deduce from this is that the p-value is less than $\alpha = 0.001$.

In this circumstance, that is when the degrees of freedom of the t-test are large, the t-distribution approaches that of the Normal distribution. Hence the entries of Table T1 can be used to establish the p-value. The largest entry in Table T1 is for $z = 3.99$, with a corresponding two-sided $\alpha = 2(1 - 0.99997) = 0.00006$. Thus, for the calculated $z = 4.26$ the p-value < 0.00006.

Coefficient of determination

The coefficient of determination, R^2, measures how well a regression model performs as a predictor of y by calculating the variation accounted for by the fitted model as a proportion of the total sums of squares. In the case of a simple linear regression with a single covariate it is calculated as

$$R^2 = \frac{b_1^2 S_{xx}}{S_{yy}} \tag{1.15}$$

and which takes values between 0 and 1. For the example of Figure 1.14(b), this gives $R^2 = 1.7175/12.8994 = 0.13$ or 13%

Extending the simple linear model

Although equations (1.1) and (1.3) refer to the same two-group situation, the simple linear regression format of the latter expression can be more easily extended to describe more complex clinical study designs. Specifically, equation (1.7), the more general form of (1.3), can be extended to include other covariates by relabeling x as x_1 and then adding, for example $\beta_2 x_2$, to the right-hand side. In this extended context it often becomes necessary to use the ANOVA approach to test the relevant null hypotheses.

Correlation

In the situation where we are concerned with more than one continuous covariate, then these may not act independently of each other. For example, when examining variation

in HDL one might be concerned whether both body-weight, x_1, and age, x_2, of the individuals play a role. However, the values of such variables may be associated. The strength of the linear association between them is described by the correlation coefficient, ρ, which is estimated by

$$r = \sqrt{\frac{\left[\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]^2}{\sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2)^2}} \tag{1.16}$$

The corresponding test the null hypothesis, H_0: $\rho = 0$, is given by

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \tag{1.17}$$

If the null hypothesis is true, this has a Student's t-distribution with $df = n - 2$, where n is the number of subjects concerned.

In this text we are also concerned with two other applications of the format of equation (1.16). The first is when one wishes to assess the strength of the association between the continuous dependent variable, y, and a continuous covariate x. In such a case, y replaces x_1 while x replaces x_2 in equation (1.16). This is the format which can be used for estimating the correlation between FEV_1 and IFN-λ protein production from the 13 individuals of Example 1.1. The corresponding command is (**`pwcorr FEV1 Interpgml`**) to give $r = 0.6469$. The test of the null hypothesis, $\rho = 0$, of equation (1.17) is $t = \dfrac{0.6469}{\sqrt{\dfrac{1 - 0.6469^2}{13 - 2}}} = \dfrac{0.6469}{0.2299} = 2.81$ and use of Table T4 with $df = 13 - 2 = 11$ gives the p-value < 0.02. An even more precise calculation gives p-value $= 0.017$.

This is the same result as that which followed the command (**`regress FEV1 Interpgml`**) of Figure 1.7, where the increasing slope of $b_{\text{IFN-}\lambda} = 0.1011$ per unit increase in pg/mL was found to be statistically significant, p-value $= 0.017$. This is no coincidence as the test for a null correlation coefficient (H_0: $\rho = 0$) is exactly equivalent to that for testing the null hypothesis: H_0: $\beta_{\text{IFN-}\lambda} = 0$ in this situation.

The other application of the equation (1.16) arises when the dependent variable is measured repeatedly over time on the individuals in the study. For instance, in the clinical trial of Example 1.6, self-reported successive pain assessments are made by the patients with OLP. In this situation, the strength of the association, termed the auto-correlation, between two measures of the *same* variable on *two* occasions, y_1 and y_2, is assessed using equation (1.16) with x_1 and x_2 replaced by y_1 and y_2, respectively. We return to this second application in Chapter 8.

Logarithms and the exponential constant, *e*

Campbell (2006) provides a clear description of logarithms, some of their properties as well as introducing the exponential constant, and upon which we base our description.

Powers

In the simplest situation, if we define $y=x\times x$, then we can more briefly write this as x^2, which we can describe as x to the power 2 or x-squared. This can be extended to multiplying x by itself n times to give $y=x\times x\times x\times \ldots \times x=x^n$.

One result that follows from this is

$$x^m \times x^n = x^{m+n} \tag{1.18}$$

and this holds for *any* values of m and n, not just whole numbers.

One important consequence is that $x^0=1$ because $x^m=x^{0+m}=x^0\times x^m$, and therefore x^0 must equal 1.

The concept of powers can be extended to allow n, m, or both to take fractional and/or negative values. Thus, $y=x^{0.5}$ is equivalent to $y=\sqrt{x}$ the square root of x. This is because $x^{0.5}\times x^{0.5}=x^{0.5+0.5}=x^1=x$ as does the product $\sqrt{x}\times\sqrt{x}=x$.

Further, if $m=1$ and $n=-1$, then $x^{m+n}=x^1\times x^{-1}=x^{1-1}=x^0=1$. Hence $x^1\times x^{-1}=x\times x^{-1}=1$ and so $x^{-1}=1/x$ or the reciprocal of x.

Logarithms

If $y=x^n$, then the definition of a logarithm of y, to what is termed the base of x, is the power that x has to be raised to in order to get y. This is written as $n=\log_x(y)$ or 'n equals the logarithm to the base x of y.'

Suppose $y=x^n$ and $z=x^m$ then $m=\log_y(z)$ and it follows that

$$\log_x(y\times z)=n+m=\log_x(y)+\log_x(z) \tag{1.19}$$

Thus, when multiplying two numbers we add their logarithms. In a similar way if we take the logarithm of a ratio of two numbers we obtain

$$\log_x(y/z)=n-m=\log_x(y)-\log_x(z) \tag{1.20}$$

Thus, when dividing two numbers we subtract their logarithms.

Choosing the base

The two most common bases for logarithms are 10 and the quantity $e=2.718281\ldots$, where the dots indicate that the decimals go on indefinitely. This base has the useful property that the slope of the curve $y=e^x$ at any point (x, y) is just y itself, whereas for all other numbers the slope is proportional to y but not equal to it. The formula $y=e^x$ is often written $\exp(x)$ and we will make much use of this form in later chapters. Logarithms to the base e are often denoted '$\log_e$' or more briefly 'log' or 'ln'. We use the form 'log' throughout this book.

The log transformation and the geometric mean

In a study, conducted by Chong, Tay, Subramaniam, *et al*. (2009) of long-term residents of a psychiatric hospital, the variation in the degree of cholinergic medication (*Chol*) received by those in different diagnostic groups was recorded. As shown in Figure 1.15(a), the distribution of the dose of cholinergic medication turns out to have a rather skewed distribution with a median of 9.0 units. However, transforming the data by taking the logarithms of the individual values, so that $LChol=\log(Chol)$, results in the more symmetric distribution of Figure 1.15(b).

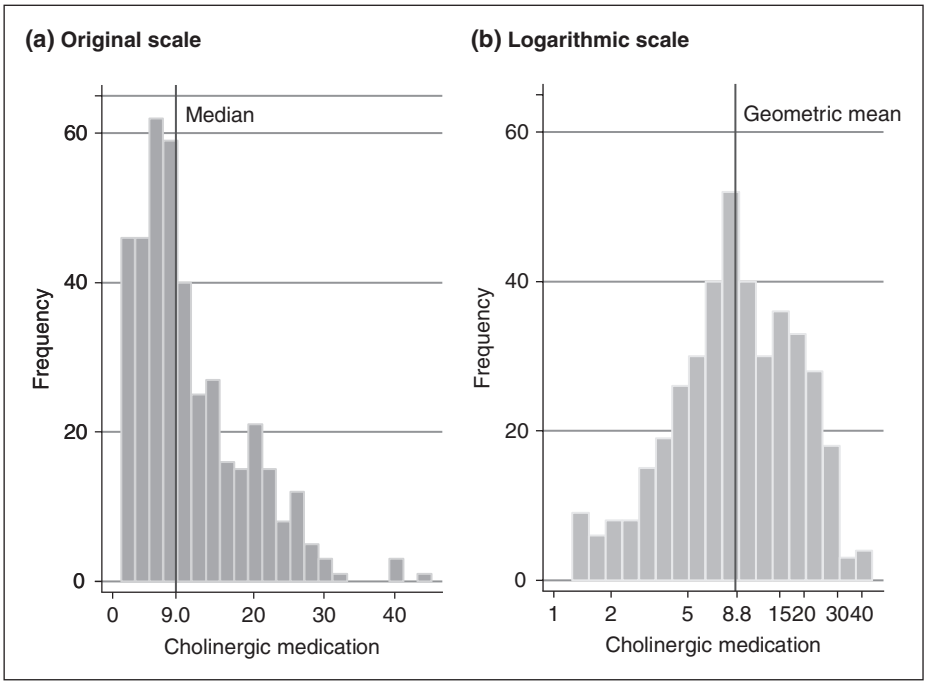


Figure 1.15 Distribution of doses of cholinergic medication given to those resident patients receiving the medication in a psychiatric hospital (a) original scale and (b) after logarithmic transformation (part data from Chong, Tay, Subramaniam, *et al.*, 2009)

It is useful to note that if the logarithm of a variable which takes positive values, $x_1, x_2, \ldots, x_n$ is taken, then the mean of $\log(x_1), \log(x_2), \ldots, \log(x_n)$ is $\overline{\log x}$, then the antilogarithm of this quantity is the geometric mean, $GM = [x_1 \times x_2 \times \ldots \times x_n]^{1/n}$. The geometric mean obtained from the transformed data of Figure 1.15(b) is 8.8 units, which is close to the median of 9.0 units on the untransformed scale.