Chapter 2 **Multiple linear regression**

Summary

When we wish to model a continuous outcome variable, then an appropriate analysis is often *multiple linear regression*. For simple linear regression we have one continuous input variable.[1] In multiple regression we generalise the method to more than one input variable and we will allow them to be continuous or categorical. We will discuss the use of *dummy* or *indicator variables* to model categories and investigate the sensitivity of models to individual data points using concepts such as *leverage* and *influence*. Multiple regression is a generalisation of the *analysis of variance* and *analysis of covariance.* The modelling techniques used here will be useful for the subsequent chapters.

2.1 The model

In multiple regression the basic model is the following:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ip} + \varepsilon_i. \qquad (2.1)$$

We assume that the error term ε_i is Normally distributed, with mean 0 and standard deviation σ.

In terms of the model structure described in Chapter 1, the link is a linear one and the error term is Normal.

Here y_i is the output for unit or subject i and there are k input variables X_{i1}, $X_{i2}, \ldots, X_{ip}$. Often y_i is termed the *dependent* variable and the input variables $X_{i1}, X_{i2}, \ldots, X_{ip}$ are termed the *independent variables*. The latter can be continuous or nominal. However the term “independent” is a misnomer since the X’s need not be independent of each other. Sometimes they are called the *explanatory* or *predictor* variables. Each of the input variables is associated with a *regression coefficient* $\beta_1, \beta_2, \ldots, \beta_p$. There is also an additive constant term β_0. These are the *model parameters.*

We can write the first section on the right-hand side of equation (2.1) as

$$LP_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ip}$$

where LP_i is known as the *linear predictor* and is the value of y_i predicted by the input variables. The difference $y_i - LP_i = \varepsilon_i$ is the *error* term.

The models are fitted by choosing estimates $b_0, b_1, \ldots, b_p$, which minimise the sum of squares (SS) of the predicted error. These estimates are termed *ordinary least squares* estimates. Using these estimates we can calculate the fitted values y_i^{fit}, and the observed residuals $e_i = y_i - y_i^{\text{fit}}$ as discussed in Chapter 1. Here it is clear that the residuals estimate the error term. Further details are given in Draper and Smith[2].

2.2 Uses of multiple regression

1 To adjust the effects of an input variable on a continuous output variable for the effects of confounders. For example, to investigate the effect of diet on weight allowing for smoking habits. Here the dependent variable is the outcome from a clinical trial. The independent variables could be the two treatment groups (as a 0/1 binary variable), smoking (as a continuous variable in numbers of packs per week) and baseline weight. The multiple regression model allows one to compare the outcome between groups, having adjusted for differences in baseline weight and smoking habit. This is also known as *analysis of covariance*.

2 To analyse the simultaneous effects of a number of categorical variables on an output variable. An alternative technique is the *analysis of variance* but the same results can be achieved using multiple regression.

3 To predict a value of an outcome, for given inputs. For example, an investigator might wish to predict the forced expiratory volume (FEV_1) of a subject given age and height, so as to be able to calculate the observed FEV_1 as a percentage of predicted, and to decide if the observed FEV_1 is below, say, 80% of the predicted one.

2.3 Two independent variables

We will start off by considering two independent variables, which can be either continuous or binary. There are three possibilities: both variables continuous, both binary (0/1), or one continuous and one binary. We will anchor the examples in some real data.

Example

Consider the data given on the pulmonary anatomical deadspace and height in 15 children given in Swinscow and Campbell.[1] Suppose that of the 15 children, 8 had asthma and 4 bronchitis. The data are given in Table 2.1.

2.3.1 One continuous and one binary independent variable

In Swinscow and Campbell,[1] the question posed was whether there is a relationship between deadspace and height. Here we might ask, is there a different relationship between deadspace and height for asthmatics than for non-asthmatics?

Suppose the two independent variables are height and asthma status. There are a number of possible models:

1 *The slope and the intercept are the same for the two groups even though the means are different.*

The model is

$$\text{Deadspace} = \beta_0 + \beta_{\text{Height}} \times \text{Height}. \tag{2.2}$$

This is illustrated in Figure 2.1. This is the simple linear regression model described in Swinscow and Campbell.[1]

2 *The slopes are the same, but the intercepts are different.*

The model is

$$\text{Deadspace} = \beta_0 + \beta_{\text{Height}} \times \text{Height} + \beta_{\text{Asthma}} \times \text{Asthma}. \tag{2.3}$$

Table 2.1 Lung function data on 15 children

Child Number	Deadspace (ml)	Height (cm)	Asthma (0 = no, 1 = yes)	Age (years)	Bronchitis (0 = no, 1 = yes)
1	44	110	1	5	0
2	31	116	0	5	1
3	43	124	1	6	0
4	45	129	1	7	0
5	56	131	1	7	0
6	79	138	0	6	0
7	57	142	1	6	0
8	56	150	1	8	0
9	58	153	1	8	0
10	92	155	0	9	1
11	78	156	0	7	1
12	64	159	1	8	0
13	88	164	0	10	1
14	112	168	0	11	0
15	101	174	0	14	0

This is illustrated in Figure 2.2.

It can be seen from model (2.3) that the interpretation of the coefficient β_{Asthma} is the difference in the intercepts of the two parallel lines which have slope β_{Height}. It is the difference in deadspace between asthmatics and

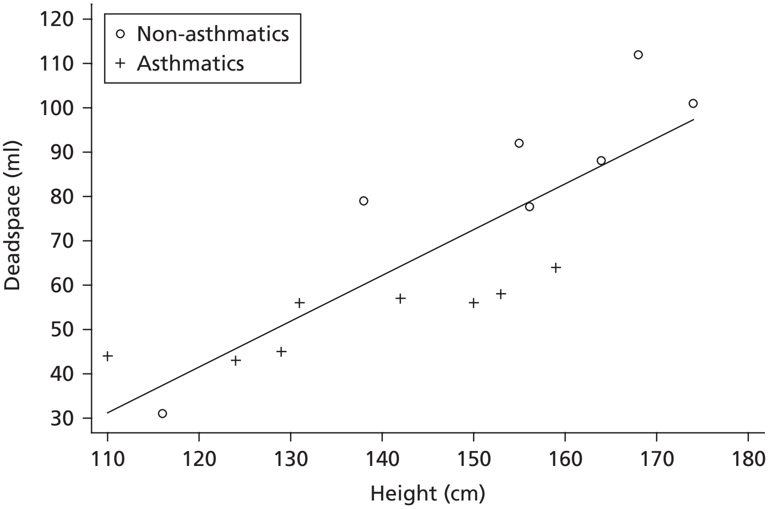

Figure 2.1 Deadspace vs height ignoring asthma status.

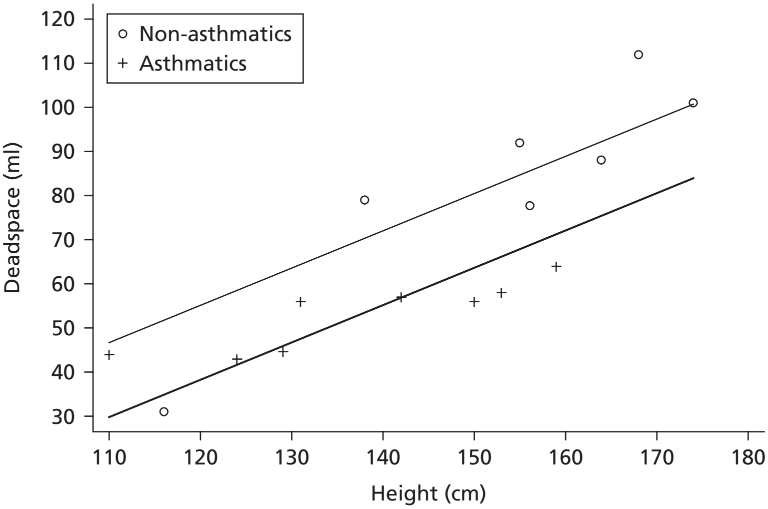

Figure 2.2 Parallel slopes for asthmatics and non-asthmatics.

non-asthmatics for any value of height, or in other words, it is the difference *allowing for* height. Thus if we thought that the only reason that asthmatics and non-asthmatics in our sample differed in the deadspace was because of a difference in height, and this is the sort of model we would fit. This type of model is termed an *analysis of covariance*. It is very common in the medical literature. An important assumption is that the slope is the same for the two groups.

We shall see later that, although they have the same symbol, we will get different estimates of β_{Height} when we fit equations (2.2) and (2.3).

3 *The slopes and the intercepts are different in each group.*
To model this we form a third variable $x_3 = \text{Height} \times \text{Asthma}$. Thus x_3 is the same as height when the subject is asthmatic and is 0 otherwise. The variable x_3 measures the *interaction* between asthma status and height. It measures by how much the slope between deadspace and height is affected by being an asthmatic.

The model is

$$\begin{aligned}\text{Deadspace} = \beta_0 + \beta_{\text{Height}} \times \text{Height} + \beta_{\text{Asthma}} \times \text{Asthma}\\ + \beta_3 \times \text{Height} \times \text{Asthma}. \qquad (2.4)\end{aligned}$$

This is illustrated in Figure 2.3, in which we have separate slopes for non-asthmatics and asthmatics.

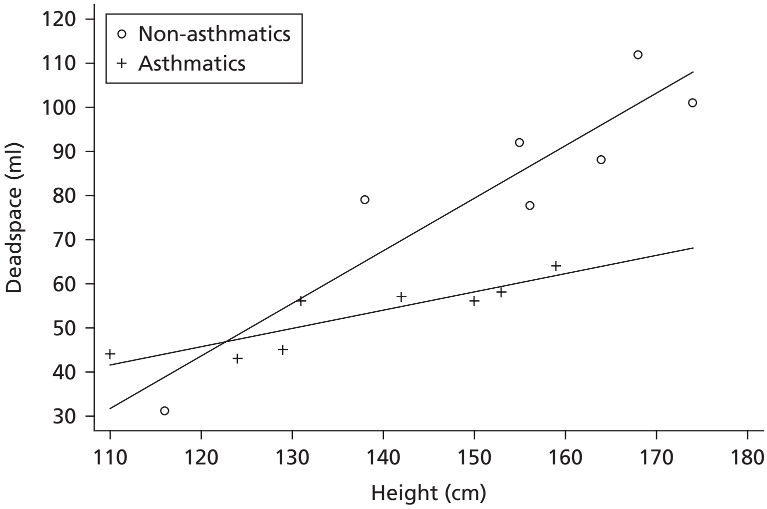

Figure 2.3 Separate lines for asthmatic and non-asthmatics.

The two lines are:

- *Non-asthmatics*
  Group = 0:

$$\text{Deadspace} = \beta_0 + \beta_{\text{Height}} \times \text{Height}$$

- *Asthmatics*
  Group = 1:

$$\text{Deadspace} = (\beta_0 + \beta_{\text{Asthma}}) + (\beta_{\text{Height}} + \beta_3) \times \text{Height}$$

In this model the interpretation of β_{Height} has changed from model (2.3). It is now the slope of the expected line for non-asthmatics. The slope of the line for asthmatics is $\beta_{\text{Height}} + \beta_3$. We then get the difference in slopes between asthmatics and non-asthmatics, which is given by β_3.

2.3.2 Two continuous independent variables

As an example of a situation where both independent variables are continuous, consider the data given in Table 2.1, but suppose we were interested in whether height and age together were important in the prediction of deadspace.

The equation is

$$\text{Deadspace} = \beta_0 + \beta_{\text{Height}} \times \text{Height} + \beta_{\text{Age}} \times \text{Age}. \tag{2.5}$$

The interpretation of this model is trickier than the earlier one and the graphical visualisation is more difficult. We have to imagine that we have a whole variety of subjects all of the same age, but of different heights. Then we expect the deadspace to go up by β_{Height} (ml) for each centimetre in height, irrespective of the age of the subjects. We also have to imagine a group of subjects, all of the same height, but different ages. Then we expect the deadspace to go up by β_{Age} (ml) for each year of age, irrespective of the heights of the subjects. The nice feature of this model is that we can estimate these coefficients reasonably even if none of the subjects has exactly the same age or height.

If age and height were independent then we can reasonably expect the β_{Height} in equation (2.2) to be close to the β_{Height} in equation (2.5), but clearly in this case they are not.

This model is commonly used in prediction as described in Section 2.2.

2.3.3 Categorical independent variables

In Table 2.1 the way that asthmatic status was coded is known as a *dummy* or *indicator* variable. There are two levels, asthmatic and non-asthmatic, and just one dummy variable, the coefficient of which measures the difference in the *y* variable between asthmatics and normals. For inference it does not matter

Table 2.2 One method of coding a three category variable

Status	x_1	x_2	x_3
Asthmatic	1	0	0
Bronchitic	0	1	0
Normal	0	0	1

if we code 1 for asthmatics and 0 for normals or vice versa. The only effect is to change the sign of the coefficient; the *P*-value will remain the same. However, Table 2.2 describes three categories: asthmatic, bronchitic and neither (taken as normal!), and these categories are mutually exclusive (i.e. there are no children with both asthma and bronchitis). Table 2.2 gives possible dummy variables for a group of three subjects.

We now have three possible contrasts: asthmatics vs bronchitics, asthmatics vs normals and bronchitics vs normals, but they are not all independent. Knowing two of the contrasts we can deduce the third (if you are not asthmatic or bronchitic, then you *must* be normal!). Thus we need to choose two of the three contrasts to include in the regression and thus two dummy variables to include in the regression. If we included all three variables, most regression programs would inform us politely that x_1, x_2 and x_3 were *aliased* (i.e. mutually dependent) and omit one of the variables from the equation. The dummy variable that is omitted from the regression is the one that the coefficients for the other variables are contrasted with, and is known as the *baseline* variable. Thus if x_3 is omitted in the regression that includes x_1 and x_2 in Table 2.2, then the coefficient attached to x_1 is the difference between deadspace for asthmatics and normals. Another way of looking at it is that the coefficient associated with the baseline is constrained to be 0.

2.4 Interpreting a computer output

We now describe how to interpret a computer output for linear regression. Most statistical packages produce an output similar to this one. The models are fitted using the *principle of least squares*, as explained in Appendix 2, and is equivalent to maximum likelihood when the error distribution is Normal. The estimate of the standard error (SE) is more sensitive to the Normality assumption than the estimate of the coefficients. There are two options available which do not require this assumption; these are the *bootstrap* and the *robust standard error.* Many computer packages have options for using these procedures. They are described in Appendix 3.

2.4.1 One continuous variable

The results of fitting model (2.2) to the data are shown in Table 2.3.

The computer program gives two sections of output. The first part refers to the fit of the overall model. The $F(1,13) = 32.81$ is what is known as an F-statistic (after the statistician Fisher), which depends on two numbers known as the *degrees of freedom*. The first, k, *is* the number of parameters in the model (excluding the constant term β_0) which in this case is 1 and the second is $n - p - 1$, where n is the number of observations and in this case is $15 - 1 - 1 = 13$. The Prob $>F$ is the probability that the variability associated with the model could have occurred by chance, on the assumption that the true model has only a constant term and no explanatory variables; in other words the overall significance of the model. This is given as 0.0001. An important statistic is the value R^2, which is the proportion of variance of the original data explained by the model and in this model it is 0.7162. It is the ratio of the sum of squares (SS) due to the model (5607) and the total SS (7828). For models with only one independent variable, as in this case, it is simply the square of the correlation coefficient described in Swinscow and Campbell.[1] However, one can always obtain an arbitrarily good fit by fitting as many parameters as there are observations. To allow for this, we calculate the R^2 *adjusted for degrees of freedom*, which is $R_a^2 = 1 - (1 - R^2)(n - 1)/(n - p - 1)$ and in this case is given by 0.6944. The Root MSE means the Residual Mean Square Error and has the value 13.072. It is an estimate of σ in equation (2.1), and can be deduced as the square root of the residual MS (mean square) on the left-hand side of the table. Thus $\sqrt{170.8847} = 13.072$.

The second part examines the coefficients in the model. The slope $\beta_{\text{Height}} = 1.0333$ and suggests that if one person was 1 cm taller than another we would expect their deadspace to be about 1 ml greater (perhaps easier to think

Table 2.3 Output from computer program fitting height to deadspace for data from Table 2.1

```
   Source |       SS       df       MS              Number of obs  =      15
          |                                                                
   -------+--------------------                     F( 1, 13)      =   32.81
    Model |  5607.43156     1   5607.43156          Prob > F       =  0.0001
 Residual |  2221.50178    13   170.884752          R-squared      =  0.7162
   -------+--------------------                     Adj R-squared  =  0.6944
    Total |  7828.93333    14   559.209524          Root MSE       =  13.072
   -------+------------------------------------------------------------------
Deadspace |     Coef.    Std. Err.      t     P>|t|       [95% Conf. Interval]
   -------+------------------------------------------------------------------
   Height |  1.033323    .1803872     5.73    0.000        .6436202   1.423026
    _cons |  -82.4852    26.30147    -3.14    0.008       -139.3061  -25.66433
------------------------------------------------------------------------------
```

if one person were 10 cm taller their deadspace is expected to be 10 ml greater). It is the slope of the line in Figure 2.1. The intercept $\beta_0 = -82.4582$. This is the value when the line cuts the x-axis when $x = 0$ (not the axis on the figure which is at $x = 110$). This is the predicted value of deadspace for someone with no height and is clearly a nonsense value. However, the parameter is necessary for correct interpretation of the model. Note these values are derived directly in Swinscow and Campbell (Chapter 11 in *Statistics at Square One*, 10th edn).

2.4.2 One continuous variable and one binary independent variable

We must first create a new variable Asthma = 1 for asthmatics and Asthma = 0 for non-asthmatics. This gives model (2.2), and the results of fitting this model are shown in Table 2.4.

In the top part of the output, the F-statistic now has 2 and 12 d.f., because we are fitting two independent variables. The P-value is given as 0.0000, which we interpret as <0.0001. It means that fitting both variables *simultaneously* gives a highly significant fit. It does *not* tell us about individual variables. One can see that the adjusted R^2 is greater and the Root MSE is smaller than that in Table 2.3, indicating a better fitting model than model (2.2).

In the bottom part of the output the coefficient associated with height is $\beta_{\text{Height}} = 0.845$, which is less than the same coefficient in Table 2.3. It is the slope of each of the parallel lines in Figure 2.2. It can be seen that because non-asthmatics have a higher deadspace forcing a single line through the data gives a greater slope. The vertical distance between the two lines is the coefficient associated with asthma, $\beta_{\text{Asthma}} = -16.81$. As we coded asthma as 1

Table 2.4 Output from computer program fitting height and asthma to deadspace from Table 2.1

```
   Source |      SS       df       MS              Number of obs =      15
----------+-------------------------------         F( 2, 12)     =   28.74
    Model | 6476.91571     2   3238.45785          Prob > F      =  0.0000
 Residual | 1352.01763    12   112.668136          R-squared     =  0.8273
----------+-------------------------------         Adj R-squared =  0.7985
    Total | 7828.93333    14   559.209524          Root MSE      =  10.615
          |
----------+-----------------------------------------------------------------
Deadspace |      Coef.   Std. Err.     t     P>|t|     [95% Conf.   Interval]
----------+-----------------------------------------------------------------
   Height |   .8450468    .1613921    5.24   0.000     .4934035      1.19669
   Asthma |  -16.81551    6.053128   -2.78   0.017    -30.00414    -3.626881
    _cons |  -46.29216    25.01679   -1.85   0.089    -100.7991     8.214733
----------------------------------------------------------------------------
```

and non-asthma as 0, the negative sign indicates asthmatics have a lower deadspace for a given height.

2.4.3 One continuous variable and one binary independent variable with their interaction

We now create a new variable AsthmaHt = Asthma × Height for the interaction of asthma and height. Some packages can do both of these automatically if one declares asthma as a "factor" or as "categorical", and fits a term such as "Asthma*Height" to give model (2.4).

The results of fitting these variables using a computer program are given in Table 2.5.

We fit three independent variables: Height, Asthma and AsthmaHt on Deadspace. This is equivalent to model (2.4), and is shown in Figure 2.3. Now $F(3,11) = 37.08$ and $R^2 = 0.91$, the R^2 adjusted for d.f. is given by 0.89 which is an improvement on model (2.3). The Root MSE has the value 8.0031, which again indicates an improvement on the earlier model.

In the second part of the output we see that the interaction term between height and asthma status is significant ($P = 0.009$). The *difference* in the slopes is −0.778 units (95% CI −1.317 to −0.240). There are no terms to drop from the model. Note, even if one of the main terms, asthma or height was not significant, we would *not* drop it from the model if the interaction was significant, since the interaction cannot be interpreted in the absence of the main effects, which in this case are asthma and height.

The two lines of best fit are:

Non-asthmatics:

$$\text{Deadspace} = -99.46 + 1.193 \times \text{Height}$$

Table 2.5 Output from computer program fitting height and asthma status and their interaction to deadspace from Table 2.1

```
 Source |  SS          df      MS              Number of obs   =      15
--------+-------------------                     F( 3, 11)     =   37.08
   Model|  7124.3865    3    2374.7955            Prob > F     =  0.0000
Residual|  704.546834   11   64.0497122           R-squared    =  0.9100
--------+-------------------                     Adj R-squared =  0.8855
   Total|  7828.93333   14   559.209524           Root MSE     =  8.0031
        |
--------+-----------------------------------------------------------------
Deadspace|     Coef.    Std. Err.     t       P>|t|    [95% Conf.   Interval]
--------+-----------------------------------------------------------------
   Height|   1.192565  .1635673    7.291    0.000    .8325555     1.552574
   Asthma|   95.47263  35.61056    2.681    0.021    17.09433     173.8509
 AsthmaHt|  -.7782494  .2447751   -3.179    0.009   -1.316996    -.239503
    _cons|  -99.46241  25.20795   -3.946    0.002   -154.9447    -43.98009
--------------------------------------------------------------------------
```

Asthmatics:

$$\begin{aligned}\text{Deadspace} &= (-99.46 + 95.47) + (1.193 - 0.778) \times \text{Height}\\ &= -3.99 + 0.415 \times \text{Height}\end{aligned}$$

Thus the deadspace in asthmatics appears to grow more slowly with height than that of non-asthmatics.

This is the best fit model for the data. Using model (2.2) or (2.3) for prediction, say, would result in a greater error. It is important, when considering which is the best model to look at the R^2 adjusted as well as the *P*-values. Sometimes a term can be added that gives a significant *P*-value, but only a marginal improvement in R^2 adjusted, and for the sake of simplicity may not be included as the best model.

2.4.4 Two independent variables: both continuous

Here we were interested in whether height or age or both were important in the prediction of deadspace. The analysis is given in Table 2.6.

The equation is

$$\text{Deadspace} = -59.05 + 0.707 \times \text{Height} + 3.045 \times \text{Age}.$$

The interpretation of this model is described in Section 2.3.2. Note a peculiar feature of this output. Although the overall model is significant ($P = 0.0003$) neither of the coefficients associated with height and age are significant ($P = 0.063$ and 0.291, respectively)! This occurs because age and height are strongly correlated, and highlights the importance of looking at the overall fit of a model. Dropping either will leave the other as a significant predictor in the model. Note that if we drop age, the adjusted R^2 is not greatly affected ($R^2 = 0.6944$ for height alone compared to 0.6995 for age and height) suggesting that height is a better predictor.

Table 2.6 Output from computer program fitting age and height to deadspace from Table 2.1

```
   Source |      SS       df       MS              Number of obs =      15
----------+---------------------------             F( 2, 12)     =   17.29
    Model | 5812.17397     2   2906.08698          Prob > F      =  0.0003
 Residual | 2016.75936    12    168.06328          R-squared     =  0.7424
----------+---------------------------             Adj R-squared =  0.6995
    Total | 7828.93333    14   559.209524          Root MSE      =  12.964
          |
----------+----------------------------------------------------------------
Deadspace |    Coef.    Std. Err.      t      P>|t|      [95% Conf. Interval]
----------+----------------------------------------------------------------
   Height |  .7070318   .3455362     2.046    0.063     -.0458268     1.45989
      Age |  3.044691   2.758517     1.104    0.291     -2.965602    9.054984
    _cons | -59.05205   33.63162    -1.756    0.105      -132.329    14.22495
---------------------------------------------------------------------------
```

2.4.5 Categorical independent variables

It will help the interpretation in this section to know that the mean values (ml) for deadspace for the three groups are normals 97.33, asthmatics 52.88 and bronchitics 72.25.

The analysis is given in the first half of Table 2.7. Here the two independent variables are x_1 and x_2 (refer to Table 2.2). As we noted before an important point to check is that, in general, one should see that the overall model is significant, before looking at the individual contrasts. Here we have Prob $>$ $F = 0.0063$, which means that the overall model is highly significant. If we look at the individual contrasts we see that the coefficient associated with asthma -44.46 is the difference in means between normals and asthmatics. This has a SE of 11.33 and so is highly significant. The coefficient associated with bronchitics is -25.08, is the contrast between bronchitics and normals and is not significant, implying that the mean deadspace is not significantly different in bronchitics and normals.

If we wished to contrast asthmatics and bronchitics, we need to make one of them the baseline. Thus we make x_1 and x_3 the independent variables to make bronchitics the baseline and the output is shown in the second half of Table 2.7. As would be expected the Prob $>$ F and the R^2 value are the same as the earlier

Table 2.7 Output from computer program fitting two categorical variables to deadspace from Table 2.2

Asthma and bronchitis as independent variables

```
Number of obs = 15, F(2,12) = 7.97,Prob > F = 0.0063
R-squared     = 0.5705 Adj R-squared = 0.4990
----------------------------------------------------------------
    y |  Coef.    Std. Err.    t     P>|t|   [95% Conf. Interval]
-----|----------------------------------------------------------
Asthma| -44.45833  11.33229  -3.923  0.002   -69.14928  -19.76739
Bronch| -25.08333  12.78455  -1.962  0.073   -52.93848   2.771809
 _cons|  97.33333  9.664212  10.072  0.000    76.27683   118.3898
----------------------------------------------------------------
```

Asthma and Normal as independent variables

```
Number of obs = 15, F(2, 12) = 7.97, Prob > F = 0.0063
R-squared  =  0.5705, Adj R-squared  =  0.4990
----------------------------------------------------------------
     y|  Coef.    Std. Err.    t      P>|t|   [95% Conf. Interval]
-----|----------------------------------------------------------
Asthma|   -19.375  10.25044  -1.890   0.083    -41.7088     2.9588
Normal| 25.08333  12.78455   1.962   0.073   -2.771809   52.93848
 _cons|    72.25  8.369453   8.633   0.000    54.01453   90.48547
----------------------------------------------------------------
```

model because these refer to the overall model which differs from the earlier one only in the formulation of the parameters. However, now the coefficients refer to the contrast with bronchitis, and we can see that the difference between asthmatics and bronchitics has a difference −19.38 with SE 10.25, which is not significant.

Thus the only significant difference is between asthmatics and normals.

This method of analysis is also known as *one-way analysis of variance*. It is a generalisation of the *t*-test referred to in Swinscow and Campbell.[1] One could ask what is the difference between this and simply carrying out two *t*-tests: asthmatics vs normals and bronchitics vs normals. In fact, the analysis of variance accomplishes two extra refinements. Firstly, the overall *P*-value controls for the problem of multiple testing referred to in Swinscow and Campbell.[1] By doing a number of tests against the baseline we are increasing the chances of a Type I error. The overall *P*-value in the *F*-test allows for this and since it is significant, we know that some of the contrasts must be significant. The second improvement is that in order to calculate a *t*-test we must find the pooled SE. In the *t*-test this is done from two groups, whereas in the analysis of variance it is calculated from all three, which is based on more subjects and so is more precise.

2.5 Multiple regression in action

2.5.1 Analysis of covariance

We mentioned that model (2.3) is very commonly seen in the literature. To see its application in a clinical trial consider the results of Llewellyn-Jones *et al.*,[3] part of which are given in Table 2.8. This study was a randomised-controlled trial of the effectiveness of a shared care intervention for depression in 220 subjects over the age of 65 years. Depression was measured using the Geriatric Depression Scale, taken at baseline and after 9.5 months of blinded follow-up. The figure that helps the interpretation is Figure 2.2. Here y is the depression scale after 9.5 months of treatment (continuous), x_1 is the value of the same scale at baseline and x_2 is the group variable, taking the value 1 for intervention and 0 for control.

Table 2.8 Factors affecting Geriatric Depression Scale score at follow-up

Variable	Regression coefficient (95% CI)	Standardised Regression Coefficient	*P*-value
Baseline score	0.73 (0.56 to 0.91)	0.56	<0.0001
Treatment Group	−1.87 (−2.97 to −0.76)	−0.22	0.0011

The *standardised regression coefficient* is not universally defined, but in this case is obtained when the *x* variable is replaced by *x* divided by its standard deviation. Thus the interpretation of the standardised regression coefficient is the amount the *y* changes for 1 standard deviation increase in *x*. One can see that the baseline values are highly correlated with the follow-up values of the score. The intervention resulted, on average, in patients with a score 1.87 units (95% CI 0.76 to 2.97) lower than those in the control group, throughout the range of the baseline values.

This analysis assumes that the treatment effect is the same for all subjects and is not related to values of their baseline scores. This possibility could be checked by the methods discussed earlier. When two groups are balanced with respect to the baseline value, one might assume that including the baseline value in the analysis will not affect the comparison of treatment groups. However, it is often worthwhile including because it can improve the precision of the estimate of the treatment effect; that is, the SEs of the treatment effects may be smaller when the baseline covariate is included.

2.5.2 Two continuous independent variables

Sorensen *et al.*[4] describe a cohort study of 4300 men, aged between 18 and 26, who had their body mass index (BMI) measured. The investigators wished to relate adult BMI to the men's birth weight and body length at birth. Potential confounding factors included gestational age, birth order, mother's marital status, age and occupation. In a multiple linear regression they found an association between birth weight (coded in units of 250 g) and BMI (allowing for confounders), regression coefficient 0.82, and SE 0.17, but not between birth length (cm) and BMI, regression coefficient 1.51, SE 3.87. Thus, for every increase in birth weight of 250 g, the BMI increases on average by 0.82 kg/m^2. The authors suggest that *in utero* factors that affect birth weight continue to have an affect even into adulthood, even allowing for factors, such as gestational age.

2.6 Assumptions underlying the models

There are a number of assumptions implicit in the choice of the model. The most fundamental assumption is that the model is *linear*. This means that each increase by one unit of an *x* variable is associated by a fixed increase in the *y* variable, irrespective of the starting value of the *x* variable.

There are a number of ways of checking this when *x* is continuous:

- For single continuous independent variables the simplest check is a visual one from a scatter plot of *y* vs *x*.

- Try transformations of the x variables ($\log x$, x^2 and $1/x$ are the commonest). There is not a simple significance test for one transformation against another, but a good guide would be if the R^2 value gets larger.
- Include a quadratic term (x^2) as well as the linear term (x) in the model. This model is the one where we fit two continuous variables, x and x^2. A significant coefficient for x^2 indicates a lack of linearity.
- Divide x into a number groups such as by quintiles. Fit separate dummy variables for the four largest quintile groups and examine the coefficients. For a linear relationship, the coefficients themselves will increase linearly.

Another fundamental assumption is that the error terms are independent of each other. An example of where this is unlikely is when the data form a time series. A simple check for sequential data for independent errors is whether the residuals are correlated, and a test known as the *Durbin–Watson* test is available in many packages. Further details are given in Chapter 6, on time series analysis. A further example of lack of independence is where the main unit of measurement is the individual, but several observations are made on each individual, and these are treated as if they came from different individuals. This is the problem of *repeated measures*. A similar type of problem occurs when groups of patients are randomised, rather than individual patients. These are discussed in Chapter 5, on repeated measures.

The model also assumes that the error terms are independent of the x variables and variance of the error term is constant (the latter goes under the more complicated term of *heteroscedascity*). A common alternative is when the error increases as one of the x variables increases, so one way of checking this assumption would be to plot the residuals, e_i, against each of the independent variables and also against the fitted values. If the model were correct one would expect to see the scatter of residuals evenly spread about the horizontal axis and not showing any pattern. A common departure from this is when the residuals fan out; that is, the scatter gets larger as the x variable gets larger. This is often also associated with nonlinearity as well, and so attempts at transforming the x variable may resolve this issue.

The final assumption is that the error term is Normally distributed. One could check this by plotting a histogram of the residuals, although the method of fitting will mean that the observed residuals e_i are likely to be closer to a Normal distribution than the true ones ε_i. The assumption of Normality is important mainly so that we can use normal theory to estimate confidence intervals (CIs) around the coefficients, but luckily with reasonably large sample sizes, the estimation method is robust to departures from Normality. Thus moderate departures from Normality are allowable. If one was concerned, then one could also use bootstrap methods and the robust standard error described in Appendix 3.

It is important to remember that the main purpose of this analysis is to assess a relationship, *not* test assumptions, so often we can come to a useful conclusion *even when the assumptions are not perfectly satisfied.*

2.7 Model sensitivity

Model sensitivity refers to how estimates are affected by subgroups of the data. Suppose we had fitted a simple regression (model (2.2)), and we were told that the estimates b_0 and b_1 altered dramatically if you delete a subset of the data, or even a single individual. This is important, because we like to think that the model applies generally, and we do not wish to find that we should have different models for different subgroups of patients.

2.7.1 Residuals, leverage and influence

There are three main issues in identifying model sensitivity to individual observations: *residuals*, *leverage* and *influence*. The residuals are the difference between the observed and fitted data: $e_i = y_i^{\text{obs}} - y_i^{\text{fit}}$. A point with a large residual is called an outlier. In general, we are interested in outliers because they may influence the estimates, but it is possible to have a large outlier which is not influential.

Another way that a point can be an outlier is if the values of the x_i are a long way from the mass of x. For a single variable, this means if x_i is a long way from $\bar{x}$. Imagine a scatter plot of y against x, with a mass of points in the bottom-left-hand corner and a single point in the top right. It is possible that this individual has unique characteristics that relate to both the x and y variables. A regression line fitted to the data will go close, or even through the isolated point. This isolated point will not have a large residual, yet if this point was deleted the regression coefficient might change dramatically. Such a point is said to have high *leverage* and this can be measured by a number, often denoted h_i; large values of h_i indicate a high leverage.

An influential point is one that has a large effect on an estimate. Effectively one fits the model with and without that point and finds the effect of the regression coefficient. One might look for points that have a large effect on b_0, or on b_1 or on other estimates such as $\mathrm{SE}(b_1)$. The usual output is the difference in the regression coefficient for a particular variable when the point is included or excluded, scaled by the estimated SE of the coefficient. The problem is that different parameters may have different influential points. Most computer packages now produce residuals, leverages and influential points as a matter of routine. It is the task for an analyst to examine these and to identify important cases. However, just because a point is influential or has a large residual it does not follow that it should be deleted, although the data should be examined carefully for possible measurement or transcription errors. A proper analysis of such data would report such sensitivities to individual points.

2.7.2 Computer analysis: model checking and sensitivity

We will illustrate model checking and sensitivity using the deadspace, age and height data in Table 2.1.

Figure 2.1 gives us reassurance that the relationship between deadspace and height is plausibly linear. We could plot a similar graph for deadspace and age. The standard diagnostic plot is a plot of the residuals against the fitted values, and for the model fitted in Table 2.6 it is shown in Figure 2.4. There is no apparent pattern, which gives us reassurance about the error term being relatively constant and further reassurance about the linearity of the model.

The diagnostic statistics are shown in Table 2.9 where the *influence* statistics are *inf_age* associated with age and *inf_ht* associated with height. As one might expect the children with the highest leverages are the youngest (who is also the shortest) and the oldest (who is also the tallest). Note that the largest residuals are associated with small leverages. This is because points with large leverage will tend to force the line close to them.

The child with the most influence on the age coefficient is also the oldest, and removal of that child would change the standardised regression coefficient by 0.79 units. The child with the most influence on height is the shortest child.

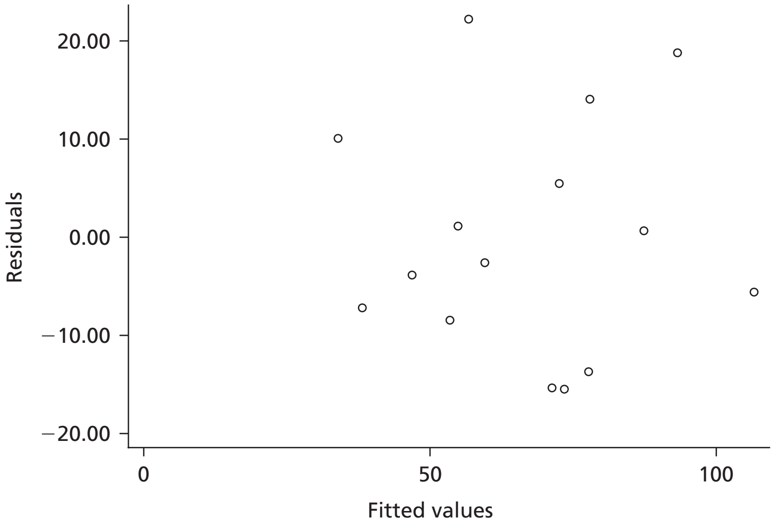

Figure 2.4 Graph of residuals against fitted values for regression model in Table 2.4 with age and height as the independent variables.

Table 2.9 Diagnostics from model fitted in Table 2.4 (output from computer program)

	Height	Age	resids	leverage	inf_age	inf_ht
1	110	5	10.06	0.33	0.22	−0.48
2	116	5	−7.19	0.23	−0.04	0.18
3	124	6	−3.89	0.15	−0.03	0.08
4	129	7	−8.47	0.15	−0.14	0.20
5	131	7	1.12	0.12	0.01	−0.02
6	138	6	22.21	0.13	−0.52	0.34
7	142	6	−2.61	0.17	0.08	−0.06
8	150	8	−15.36	0.08	0.11	−0.14
9	153	8	−15.48	0.10	0.20	−0.26
10	155	9	14.06	0.09	0.02	0.07
11	156	7	5.44	0.28	−0.24	0.25
12	159	8	−13.72	0.19	0.38	−0.46
13	164	10	0.65	0.14	0.00	0.01
14	168	11	18.78	0.19	0.29	0.08
15	174	14	−5.60	0.65	−0.79	0.42

However, neither child should be removed without strong reason. (A strong reason may be if it was discovered the child had some relevant disease, such as cystic fibrosis.)

2.8 Stepwise regression

When one has a large number of independent variables, a natural question to ask is what is the best combination of these variables to predict the *y* variable? To answer this, one may use *stepwise* regression that is available in a number of packages. *Step-down* or *backwards* regression starts by fitting all available variables and then discarding sequentially those that are not significant. *Step-up* or *forwards* regression starts by fitting an overall mean, and then selecting variables to add to the model according to their significance. *Stepwise* regression is a mixture of the two, where one can specify a *P*-value for a variable to be entered into the model, and then a *P*-value for a variable to be discarded. Usually one chooses a larger *P*-value for entry (say, 0.1) than for exclusion (say, 0.05), since variables can jointly be predictive, and separately they are not. This also favours *step-down* regression. As an example consider an outcome variable being the amount a person limps. The length of the left or right legs is not predictive, but the difference in lengths is highly predictive. Stepwise regression is best used in the *exploratory* phase of an analysis (see Chapter 1), to identify a few predictors in a mass of data, the association of which can be verified by further data collection.

There are a few problems with stepwise regression:

- The *P*-values are invalid since they do not take account of the vast number of tests that have been carried out; different methods, such as step-up and step-down, are likely to produce different models and experience shows that the same model rarely emerges when a second data set is analysed. One way of trying to counter this is to split a large data set into two, and run the stepwise procedure on both separately. Choose the variables that are common to both data sets, and fit these to the combined data set as the final model.
- Many large data sets contain missing values. With stepwise regression, usually only the subjects who have no missing values on *any* of the variables under consideration are chosen. The final model may contain only a few variables, but if one refits the model, the parameters change because now the model is being fitted to those subjects who have no missing values on only the few chosen variables, which may be a considerably larger data set than the original.
- If a categorical variable is coded as a number of dummies, some of these may be lost in the fitting process, and this changes the interpretation of the others. Thus, if we fitted x_1 and x_2 from Table 2.2, and then we lost x_2, the interpretation of x_1 is of a contrast between asthmatics with bronchitics and normals *combined.*

Thus stepwise regression is useful in the *exploratory* phase of an analysis, but not the *confirmatory* one.

2.9 Reporting the results of a multiple regression

- As a minimum, report the regression coefficients and SEs or CIs for the main independent variables, together with the adjusted R^2 for the whole model.
- If there is one main dependent variable, show a scatter plot of each independent variable vs dependent variable with the best-fit line.
- Report how the assumptions underlying the model were tested and verified. In particular is linearity plausible?
- Report any sensitivity analysis carried out.
- Report *all* the variables included in the model. For a stepwise regression, report *all* the variables that could have entered the model.
- Note that if an interaction term is included in a model, the main effects *must* be included.

2.10 Reading the results of a multiple regression

In addition to the points in Section 1.11:

- Note the value of R^2. With a large study, the coefficients in the model can be highly significant, but only explain a low proportion of the variability of the outcome variable. Thus they may be of no use for prediction.

- Are the models plausibly linear? Are there any boundaries, which may cause the slope to flatten?
- Were outliers and influential points identified, and how were they treated?
- An analysis of covariance *assumes* that the slopes are the same in each group. Is this plausible and has it been tested?

FREQUENTLY ASKED QUESTIONS

1 *Does it matter how a dummy variable is coded?*

If you have only one binary variable, then coding the dummy variable 0 and 1 is the most convenient. Coding it 1 and 2 is commonly the method in questionnaires. It will make no difference to the coefficient estimate or P-value. However it will change the value of the intercept, because now the value in the group assigned 1 will be $a + b$ and the value in the group assigned 2 will be $a + 2b$. Thus in Figure 2.2 when "asthma" is coded 0 or 1 the regression coefficient is -16.8 and the intercept is -46.3. If we had coded the variable 1 or 2 we would find the regression coefficient is still -16.8 but the intercept would be $(-46.3 - 16.8) = -63.1$. Coding the dummy variable to -1 and $+1$, (e.g. as is done in the package SAS) does not change the P-value but the coefficient is halved.

If you have a categorical variable with, say, three groups, then this will be coded with two dummy variables. As shown earlier, the overall F-statistic will be unchanged no matter which two groups are chosen to be represented by dummies, but the coefficient of group 2, say, will be dependent on whether group 1 or 3 is the omitted variable.

2 *How do I treat an ordinal independent variable?*

Most packages assume that the predictor variable, X, in a regression model is either continuous or binary. Thus one has a number of options:

(i) Treat the predictor as if it were continuous. This incorporates into the model the fact that the categories are ordered, but also assumes that equal changes in X mean equal changes in y.

(ii) Treat the predictor as if it were categorical, by fitting dummy variables to all but one of the categories. This loses the fact that the predictor is ordinal, but makes no assumption about linearity.

(iii) Dichotomise the X variable, by recoding it as binary, say 1 if X is in a particular category or above, and 0 otherwise. The cut-point should be chosen on external grounds and not because it gives the best fit to the data.

Which of these options you choose depends on a number of factors. With a large amount of data, the loss of information by ignoring the ordinality in option

(*continued*)

(ii) is not critical and especially if the *X* variable is a confounder and not of prime interest. For example, if *X* is age grouped in 10-year intervals, it might be better to fit dummy variables, than assume a linear relation with the *y*-variable.

3 *Do the assumptions underlying multiple regression matter?*

Often the assumptions underlying multiple regression are not checked, partly because the investigator is confident that they hold true and partly because mild departures are unlikely to invalidate an analysis. However, lack of independence may be obvious on empirical grounds (the data form repeated measures or a time series) and so the analysis should accommodate this from the outset. Linearity is important for inference and so may be checked by fitting transformations of the independent variables. Lack of homogeneity of variance and lack of Normality may affect the SEs and often indicate the need for a transformation of the dependent variable. The most common departure from Normality is when outliers are identified, and these should be carefully checked, particularly those with high leverage.

4 *I have a variable that I believe should be a confounder but it is not significant. Should I include it in the analysis?*

There are certain variables (such as age or sex) for which one might have strong grounds for believing that they could be confounders, but in any particular analysis might emerge as not significant. These should be retained in the analysis because, even if not significantly related to the outcome themselves, they may modify the effect of the prime independent variable.

5 *What happens if I have a dependent variable, which is 0 or 1?*

When the dependent variable is 0 or 1 then the coefficients from a linear regression are proportional to what is known as the *linear discriminant function*. This can be useful for discriminating between groups, even if the assumption about Normality of the residuals is violated. However discrimination is normally carried out now using *logistic regression* (Chapter 3).

6 *Why not analyse the difference between outcome and baseline (change score) rather than use analysis of covariance?*

Analysing change does not properly control for baseline imbalance because of what is known as regression to the mean; baseline values are negatively correlated with change and subjects with low scores at baseline will tend to increase more than those with high values. However, if the correlation between baseline and follow-up measurements is large (say, $r > 0.8$) and randomisation has ensured that baseline values are comparable between groups, then an analysis of change scores may produce lower SEs. Note that if the change score is the dependent variable and baseline is included as an independent variable, then the results will be the same as an analysis of covariance.

EXERCISE

Melchart *et al.*[5] describe a randomised trial of acupuncture in patients with tension-type headache with 2:1 randomisation to either acupuncture for 8 weeks or a waiting list control. Partial results are given in the following table.

Results from Melchart *et al.*[5]

	Acupuncture	Waiting list
Baseline	17.5 (6.9) (n = 132)	17.3 (6.9) (n = 75)
After treatment	9.9 (8.7) (n = 118)	16.3 (7.4) (n = 63)

Values are represented as days with headache during a 28-day period (Mean (SD)).

Difference between groups after treatment: 5.7 days (95% CI 4.2 to 7.2) $P < 0.001$.

Analysis of covariance adjusting for baseline value – Difference between groups after treatment: 5.8 days (95% CI 4.0 to 7.6) $P < 0.001$.

1 Give three assumptions made for the analysis of covariance.

2 What evidence do we have that these may not be satisfied?

3 Contrast the two CIs.

4 What other data might one like to see?

References

1. Swinscow TDV, Campbell MJ. *Statistics at Square One*, 10th edn. London: BMJ Books, 2002.
2. Draper NR, Smith H. *Applied Regression Analysis*, 3rd edn. New York: John Wiley, 1998.
3. Llewellyn-Jones RH, Baikie KA, Smithers H, Cohen J, Snowdon J, Tennant CC. Multifaceted shared care intervention for late life depression in residential care: randomised controlled trial. *Br Med J* 1999; **319:** 676–82.
4. Sorensen HT, Sabroe S, Rothman KJ, Gillman M, Fischer P, Sorensen TIA. Relation between weight and length at birth and body mass index in young adulthood: cohort study. *Br Med J* 1997; **315:** 1137.
5. Melchart D, Streng A, Hoppe A, Brinkhaus B, Witt C, *et al.* Acupuncture in patients with tension-type headache: randomised controlled trial. *Br Med J* 2005; **331:** 376–82.