11

Multiple regression

This chapter discusses the case of regression analysis with multiple predictors. There is not really much new here since model specification and output do not differ a lot from what has been described for regression analysis and analysis of variance. The news is mainly the model search aspect, namely among a set of potential descriptive variables to look for a subset that describes the response sufficiently well.

The basic model for multiple regression analysis is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

where $x_1, \dots x_k$ are explanatory variables (also called predictors) and the parameters $\beta_1, \dots, \beta_k$ can be estimated using the method of least squares (see Section 6.1). A closed-form expression for the estimates can be derived using matrix calculus, but we do not go into the details of that here.

11.1 Plotting multivariate data

As an example in this chapter, we use a study concerning lung function in patients with cystic fibrosis in Altman (1991, p. 338). The data are in the `cystfibr` data frame in the `ISwR` package.

P. Dalgaard, *Introductory Statistics with R*,
DOI: 10.1007/978-0-387-79054-1_11, © Springer Science+Business Media, LLC 2008

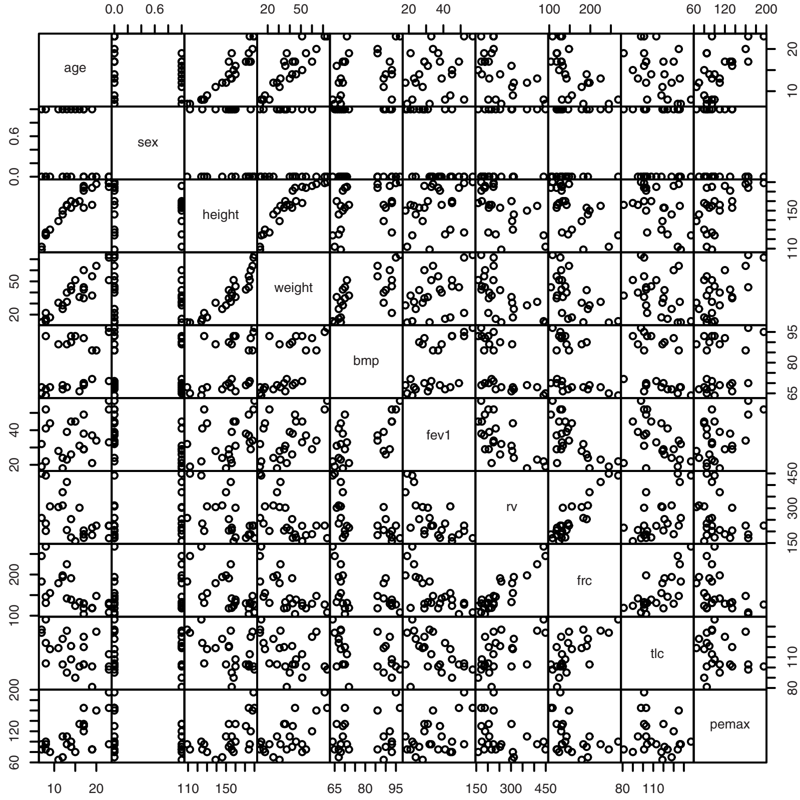

Figure 11.1. Pairwise plots for cystic fibrosis data.

You can obtain pairwise scatterplots between all the variables in the data set. This is done using the function `pairs`. To get Figure 11.1, you simply write

```
> par(mex=0.5)
> pairs(cystfibr, gap=0, cex.labels=0.9)
```

The arguments `gap` and `cex.labels` control the visual appearance by removing the space between subplots and decreasing the font size. The `mex` graphics parameter reduces the interline distance in the margins.

A similar plot is obtained by simply saying `plot(cystfibr)` since the `plot` function is generic and behaves differently depending on the class of its arguments (see Section 2.3.2). Here the argument is a data frame and a `pairs` plot is a fairly reasonable thing to get when asking for a plot of an entire data frame (although you might equally reasonably have expected a histogram or a barchart of each variable instead).

The individual plots do get rather small, probably not suitable for direct publication, but such plots are quite an effective way of obtaining an overview of multidimensional issues. For example, the close relations among age, height, and weight appear clearly on the plot.

In order to be able to refer directly to the variables in `cystfibr`, we add it to the search path (a harmless warning about masking of `tlc` ensues at this point):

```
> attach(cystfibr)
```

Because this data set contains common variable names such as `age`, `height`, and `weight`, it is a good idea to ensure that you do not have identically named variables in the workspace at this point. In particular, such names were used in the introductory session.

11.2 Model specification and output

Specification of a multiple regression analysis is done by setting up a model formula with + between the explanatory variables:

```
lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
```

which is meant to be read as "`pemax` is described using a model that is additive in `age`, `sex`, and so forth." (`pemax` is the maximal expiratory pressure. See Appendix B for a description of the other variables in `cystfibr`.)

As usual, there is not much output from `lm` itself, but with the aid of `summary` you can obtain some more interesting output:

```
> summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))

Call:
lm(formula = pemax ~ age + sex + height + weight + bmp + fev1 +
    rv + frc + tlc)

Residuals:
    Min      1Q  Median      3Q     Max
-37.338 -11.532   1.081  13.386  33.405

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
```

```
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711

Residual standard error: 25.47 on 15 degrees of freedom
Multiple R-squared: 0.6373,     Adjusted R-squared: 0.4197
F-statistic: 2.929 on 9 and 15 DF, p-value: 0.03195
```

The layout should be well known by now. Notice that there is not one single significant t value, but the joint F test is nevertheless significant, so there must be an effect somewhere. The reason is that the t tests only say something about what happens if you remove one variable and leave in all the others. You cannot see whether a variable would be statistically significant in a reduced model; all you can see is that no variable *must* be included.

Note further that there is quite a large difference between the unadjusted and the adjusted R^2, which is due to the large number of variables relative to the number of degrees of freedom for the variance. Recall that the former is the change in residual sum of squares relative to an empty model, whereas the latter is the similar change in residual *variance*:

```
> 1-25.5^2/var(pemax)
[1] 0.4183949
```

The 25.5 comes from "residual standard error" in the `summary` output.

The ANOVA table for a multiple regression analysis is obtained using `anova` and gives a rather different picture:

```
> anova(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))
Analysis of Variance Table

Response: pemax
          Df  Sum Sq Mean Sq F value   Pr(>F)
age        1 10098.5 10098.5 15.5661 0.001296 **
sex        1   955.4   955.4  1.4727 0.243680
height     1   155.0   155.0  0.2389 0.632089
weight     1   632.3   632.3  0.9747 0.339170
bmp        1  2862.2  2862.2  4.4119 0.053010 .
fev1       1  1549.1  1549.1  2.3878 0.143120
rv         1   561.9   561.9  0.8662 0.366757
frc        1   194.6   194.6  0.2999 0.592007
tlc        1    92.4    92.4  0.1424 0.711160
Residuals 15  9731.2   648.7
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that, except for the very last line ("tlc"), there is practically no correspondence between these *F* tests and the *t* tests from summary. In particular, the effect of age is now significant. That is because these tests are successive; they correspond to (reading upward from the bottom) a stepwise removal of terms from the model until finally only age is left. During the process, bmp came close to the magical 5% limit, but in view of the number of tests, this is hardly noteworthy.

The probability that one out of eight independent tests gives a *p*-value of 0.053 or below is actually just over 35%! The tests in the ANOVA table are not completely independent, but the approximation should be good.

The ANOVA table indicates that there is no significant improvement of the model once age is included. It is possible to perform a joint test for whether *all* the other variables can be removed by adding up the sums of squares contributions and using the sum for an *F* test; that is,

```
> 955.4+155.0+632.3+2862.2+1549.1+561.9+194.6+92.4
[1] 7002.9
> 7002.9/8
[1] 875.3625
> 875.36/648.7
[1] 1.349407
> 1-pf(1.349407,8,15)
[1] 0.2935148
```

This corresponds to collapsing the eight lines of the table so that it would look like this:

```
          Df   Sum Sq  Mean Sq       F   Pr(>F)
age        1  10098.5  10098.5  15.566  0.00130
others     8   7002.9    875.4   1.349  0.29351
Residual  15   9731.2    648.7
```

(Note that this is "cheat output", in which we have manually inserted the numbers computed above.)

A procedure leading directly to the result is

```
> m1<-lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
> m2<-lm(pemax~age)
> anova(m1,m2)
Analysis of Variance Table

Model 1: pemax ~ age + sex + height + weight + bmp + fev1 + rv +
    frc + tlc
Model 2: pemax ~ age
```

```
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     15  9731.2
2     23 16734.2 -8   -7002.9 1.3493 0.2936
```

which gives the appropriate *F* test with no manual computation.

Notice, however, that you need to be careful to ensure that the two models are actually nested. R does not check this, although it does verify that the number of response observations is the same to safeguard against the more obvious mistakes. (When there are missing values in the descriptive variables, it's easy for the smaller model to contain more data points.)

From the ANOVA table, we can thus see that it is allowable to remove all variables except `age`. However, that this particular variable is left in the model is primarily due to the fact that it was mentioned first in the model specification, as we see below.

11.3 Model search

R has the `step()` function for performing model searches by the Akaike information criterion. Since that is well beyond the scope of this book, we use simple manual variants of backwards elimination.

In the following, we go through a practical model reduction for the example data. Notice that the output has been slightly edited to take up less space.

```
> summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711
...
```

One advantage of doing model reductions by hand is that you may impose some logical structure on the process. In the present case, it may, for instance, be natural to try to remove other lung function indicators first.

```
> summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc))
```

```
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 221.8055   185.4350   1.196   0.2491
age          -3.1346     4.4144  -0.710   0.4879
sex          -4.6933    14.8363  -0.316   0.7558
height       -0.5428     0.8428  -0.644   0.5286
weight        3.3157     1.7672   1.876   0.0790 .
bmp          -1.9403     1.0047  -1.931   0.0714 .
fev1          1.0183     1.0392   0.980   0.3417
rv            0.1857     0.1887   0.984   0.3396
frc          -0.2605     0.4628  -0.563   0.5813
...
> summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv))
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.71822  154.31294   1.080   0.2951
age          -1.81783    3.66773  -0.496   0.6265
sex           0.10239   11.89990   0.009   0.9932
height       -0.40981    0.79257  -0.517   0.6118
weight        2.87386    1.55120   1.853   0.0814 .
bmp          -1.94971    0.98415  -1.981   0.0640 .
fev1          1.41526    0.74788   1.892   0.0756 .
rv            0.09567    0.09798   0.976   0.3425
...
> summary(lm(pemax~age+sex+height+weight+bmp+fev1))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 260.6313   120.5215   2.163   0.0443 *
age          -2.9062     3.4898  -0.833   0.4159
sex          -1.2115    11.8083  -0.103   0.9194
height       -0.6067     0.7655  -0.793   0.4384
weight        3.3463     1.4719   2.273   0.0355 *
bmp          -2.3042     0.9136  -2.522   0.0213 *
fev1          1.0274     0.6329   1.623   0.1219
...
> summary(lm(pemax~age+sex+height+weight+bmp))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 280.4482   124.9556   2.244   0.0369 *
age          -3.0750     3.6352  -0.846   0.4081
sex         -11.5281    10.3720  -1.111   0.2802
height       -0.6853     0.7962  -0.861   0.4001
weight        3.5546     1.5281   2.326   0.0312 *
bmp          -1.9613     0.9263  -2.117   0.0476 *
...
```

As is seen, there was no obstacle to removing the four lung function variables. Next we try to reduce among the variables that describe the patient's state of physical development or size. Initially, we avoid removing `weight` and `bmp` since they appear to be close to the 5% significance limit.

```
> summary(lm(pemax~age+height+weight+bmp))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 274.5307   125.5745   2.186   0.0409 *
age          -3.0832     3.6566  -0.843   0.4091
height       -0.6985     0.8008  -0.872   0.3934
weight        3.6338     1.5354   2.367   0.0282 *
bmp          -1.9621     0.9317  -2.106   0.0480 *
...
> summary(lm(pemax~height+weight+bmp))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 245.3936   119.8927   2.047   0.0534 .
height       -0.8264     0.7808  -1.058   0.3019
weight        2.7717     1.1377   2.436   0.0238 *
bmp          -1.4876     0.7375  -2.017   0.0566 .
...
> summary(lm(pemax~weight+bmp))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 124.8297    37.4786   3.331 0.003033 **
weight        1.6403     0.3900   4.206 0.000365 ***
bmp          -1.0054     0.5814  -1.729 0.097797 .
...
> summary(lm(pemax~weight))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.5456    12.7016   5.003 4.63e-05 ***
weight        1.1867     0.3009   3.944 0.000646 ***
...
```

Notice that, once `age` and `height` were removed, `bmp` was no longer significant. In the original reference (Altman, 1991), `weight`, `fev1`, and `bmp` all ended up with *p*-values below 5%. However, far from all elimination procedures lead to that result.

It is also a good idea to pay close attention to the `age`, `weight`, and `height` variables, which are heavily correlated since we are dealing with children and adolescents.

```
> summary(lm(pemax~age+weight+height))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.65555   82.40935   0.785    0.441
age          1.56755    3.14363   0.499    0.623
weight       0.86949    0.85922   1.012    0.323
height      -0.07608    0.80278  -0.095    0.925
...
> summary(lm(pemax~age+height))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.8600    68.2493   0.262    0.796
```

```
age           2.7178     2.9325   0.927    0.364
height        0.3397     0.6900   0.492    0.627
...
> summary(lm(pemax~age))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.408     16.657   3.026  0.00601 **
age            4.055      1.088   3.726  0.00111 **
...
> summary(lm(pemax~height))
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -33.2757    40.0445  -0.831  0.41453
height        0.9319     0.2596   3.590  0.00155 **
...
```

As it turns out, there is really no reason to prefer one of the three variables over the two others. The fact that an elimination method ends up with a model containing only `weight` is essentially a coincidence. You can easily be misled by model search procedures that end up with one highly significant variable — it is far from certain that the same variable would be chosen if you were to repeat the analysis on a new, similar data set.

What you may reasonably conclude is that there is probably a connection with the patient's physical development or size, which may be described in terms of age, height, or weight. Which description to use is arbitrary. If you want to choose one over the others, a decision cannot be based on the data, although possibly on theoretical considerations and/or results from previous investigations.

11.4 Exercises

11.1 The `secher` data are best analyzed after log-transforming birth weight as well as the abdominal and biparietal diameters. Fit a prediction equation for birth weight. How much is gained by using both diameters in a prediction equation? The sum of the two regression coefficients is almost exactly 3 — can this be given a nice interpretation?

11.2 The `tlc` data set contains a variable also called `tlc`. This is not in general a good idea; explain why. Describe `tlc` using the other variables in the data set and discuss the validity of the model.

11.3 The analyses of `cystfibr` involve `sex`, which is a binary variable. How would you interpret the results for this variable?

11.4 Consider the `juul2` data set and select the group of those over 25 years old. Perform a regression analysis of $\sqrt{\texttt{igf1}}$ on `age`, and extend the model by including `height` and `weight`. Generate the analysis of variance table for the extended model. What is the surprise, and why does it happen?

11.5 Analyze and interpret the effect of explanatory variables on the milk intake in the `kfm` data set using a multiple regression model. Notice that `sex` is a factor here; what does that imply for the analyses?