11

Multiple Regression

In multiple regression we have a continuous response variable and two or more continuous explanatory variables (i.e. no categorical explanatory variables). There are several important issues involved in carrying out a multiple regression:

- which explanatory variables to include,
- curvature in the response to the explanatory variables,
- interactions between explanatory variables,
- correlation between explanatory variables,
- the risk of over-parameterization.

The approach recommended here is that before you begin modelling in earnest you do two things:

- use tree models to investigate whether there are complicated interactions, and
- use generalized additive models (gam's) to investigate curvature.

A Simple Example

Let's begin with an example from air pollution studies. How is ozone concentration related to wind speed, air temperature and the intensity of solar radiation?

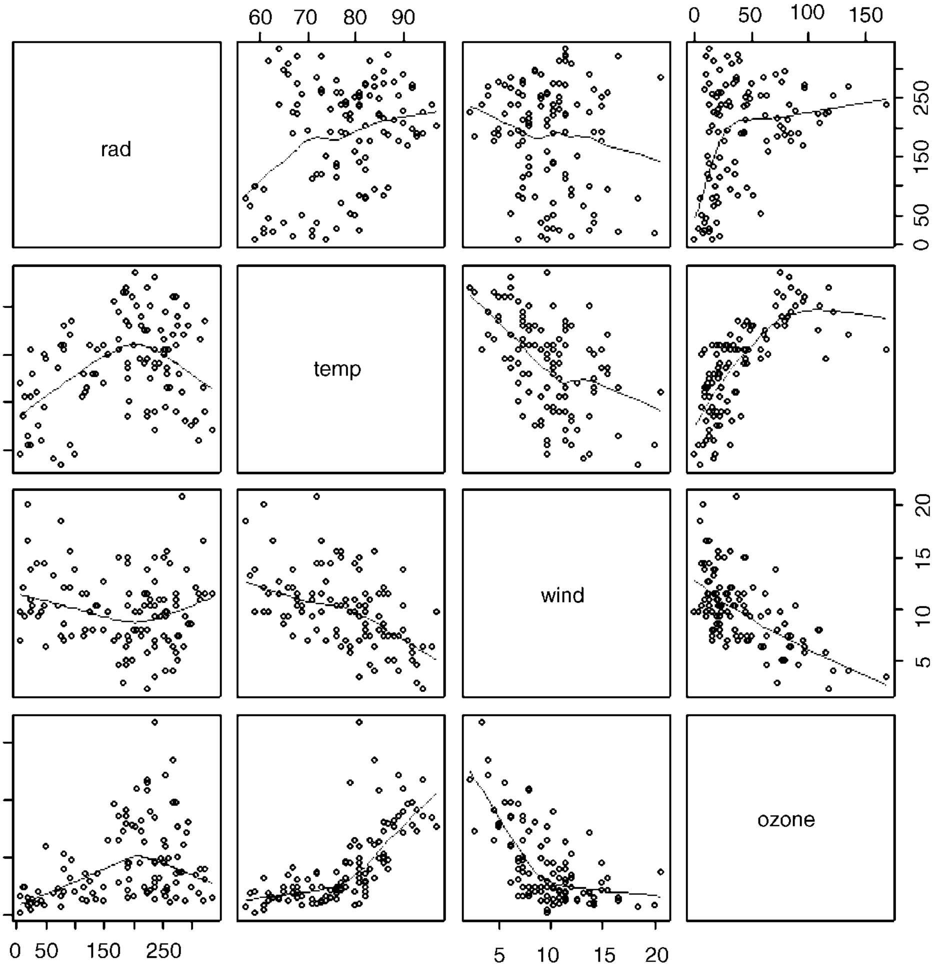
```
ozone.pollution<-read.table("c:\\temp\\ozone.data.txt",header=T)
attach(ozone.pollution)
names(ozone.pollution)
```

```
[ 1] "rad" "temp" "wind" "ozone"
```

In multiple regression, it is always a good idea to use pairs to look at all the correlations:

```
pairs(ozone.pollution,panel=panel.smooth)
```

Statistics: An Introduction using R M. J. Crawley
© 2005 John Wiley & Sons, Ltd ISBNs: 0-470-02298-1 (PBK); 0-470-02297-3 (PPC)

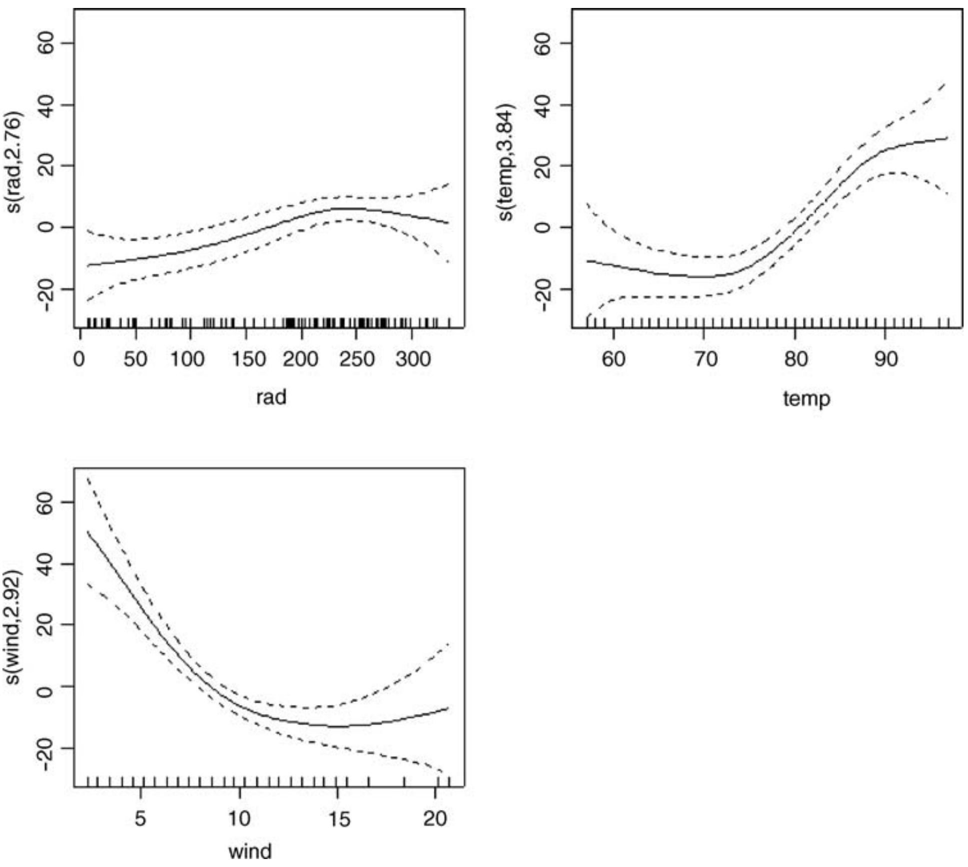


The response variable, ozone concentration, is shown on the *y* axis of the bottom row of panels: there is a strong negative relationship with wind speed, a positive correlation with temperature and a rather unclear, but possibly humped relationship with radiation.

A good way to start a multiple regression problem is using non-parametric smoothers in a generalized additive model (gam) like this:

```
library(mgcv)
par(mfrow=c(2,2))
model<-gam(ozone~s(rad)+s(temp)+s(wind))
plot(model)
par(mfrow=c(1,1))
```

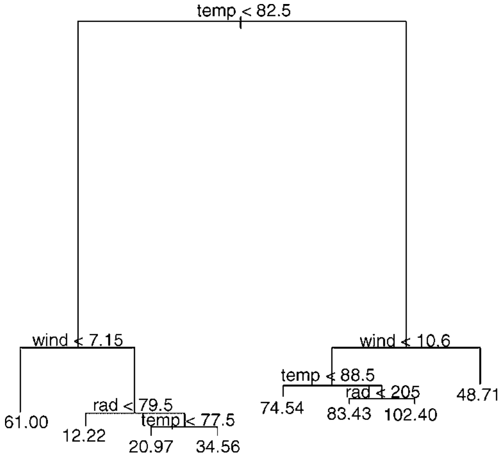
The confidence intervals are sufficiently narrow to suggest that the curvature in the relationship between ozone and temperature is real, but the curvature of the relationship with wind is questionable, and a linear model may well be all that is required for solar



radiation. The next step might be to fit a tree model to see whether complex interactions between the explanatory variables are indicated:

```
library(tree)
model<-tree(ozone~.,data=ozone.pollution)
plot(model)
text(model)
```

This shows that temperature is far and away the most important factor affecting ozone concentration (the longer the branches in the tree, the greater the deviance explained). Wind speed is important at both high and low temperatures, with still air being associated with higher mean ozone levels (the figures at the ends of the branches are mean ozone concentrations). Radiation shows an interesting, but subtle effect. At low temperatures, radiation matters at relatively high wind speeds (>7.15), whereas at high temperatures, radiation matters at relatively low wind speeds (<10.6); in both cases, however, higher radiation is associated with higher mean ozone concentration. The tree model therefore indicates that the interaction structure of the data is not particularly complex (a reassuring finding).



Armed with this background information (likely curvature of the temperature response and an uncomplicated interaction structure) we can begin the linear modelling. We start with the most complicated model: this includes interactions between all three explanatory variables plus quadratic terms to test for curvature in response to each of the three explanatory variables:

```
model1 < -lm(ozone ~ temp*wind*rad + I(rad^2) + I(temp^2) + I(wind^2))
summary(model1)
```

```
Coefficients:
                  Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)      5.683e+02   2.073e+02    2.741    0.00725   **
temp            -1.076e+01   4.303e+00   -2.501    0.01401    *
wind            -3.237e+01   1.173e+01   -2.760    0.00687   **
rad             -3.117e-01   5.585e-01   -0.558    0.57799
I(rad^2)        -3.619e-04   2.573e-04   -1.407    0.16265
I(temp^2)        5.833e-02   2.396e-02    2.435    0.01668    *
I(wind^2)        6.106e-01   1.469e-01    4.157   6.81e-05  ***
temp:wind        2.377e-01   1.367e-01    1.739   0.08519.
temp:rad         8.402e-03   7.512e-03    1.119    0.26602
wind:rad         2.054e-02   4.892e-02    0.420    0.67552
temp:wind:rad   -4.324e-04   6.595e-04   -0.656    0.51358

Residual standard error: 17.82 on 100 degrees of freedom
Multiple R-Squared: 0.7394,     Adjusted R-squared: 0.7133
F-statistic: 28.37 on 10 and 100 DF, p-value:        0
```

The three-way interaction is clearly not significant, so we remove it to begin the process of model simplification:

```
model2 < -update(model1, ~ . – temp:wind:rad)
summary(model2)
```

Next, we remove the least significant two-way interaction term – in this case wind:rad

```
model3 < -update(model2, ~. – wind:rad)
summary(model3)
```

then try removing the temperature by wind interaction:

```
model4 < -update(model3, ~. – temp:wind)
summary(model4)
```

We shall retain the marginally significant interaction between temp and rad ($p = 0.04578$) but leave out all other interactions. In model 4, the least significant quadratic term is for rad, so we delete this:

```
model5 < -update(model4, ~. – I(rad^2))
summary(model5)
```

This deletion has rendered the temp:rad interaction insignificant, and caused the main effect of radiation to become insignificant. We should try removing the temp:rad interaction

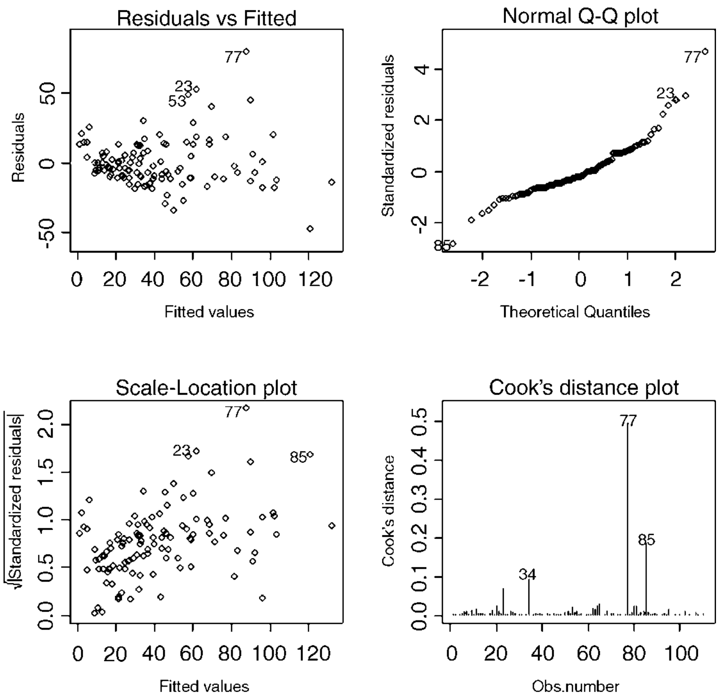
```
model6 < -update(model5, ~. – temp:rad)
summary(model6)
```

```
Coefficients:
                  Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)      291.16758   100.87723    2.886    0.00473   **
temp              -6.33955     2.71627   -2.334    0.02150    *
wind             -13.39674     2.29623   -5.834   6.05e-08  ***
rad                0.06586     0.02005    3.285    0.00139   **
I(temp^2)          0.05102     0.01774    2.876    0.00488   **
I(wind^2)          0.46464     0.10060    4.619   1.10e-05  ***

Residual standard error: 18.25 on 105 degrees of freedom
Multiple R-Squared: 0.713,       Adjusted R-squared: 0.6994
F-statistic: 52.18 on 5 and 105 DF, p-value:          0
```

Now we are making progress. All the terms in model 6 are significant. At this stage, we should check the assumptions, using plot(model6):

There is a clear pattern of variance increasing with the mean of the fitted values. This is bad news (heteroscedasticity). Also, the normality plot is distinctly curved; again, this is bad news. Let's try transformation of the response variable. There are no zeros in the response, so a log transformation is worth trying:



```
model7 < -lm(log(ozone)~temp + wind + rad + I(temp^2) + I(wind^2))
summary(model7)
```

```
Coefficients:
                Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)    2.5538486   2.7359735    0.933    0.35274
temp          -0.0041416   0.0736703   -0.056    0.95528
wind          -0.2087025   0.0622778   -3.351    0.00112  **
rad            0.0025617   0.0005437    4.711   7.58e-06 ***
I(temp^2)      0.0003313   0.0004811    0.689    0.49255
I(wind^2)      0.0067378   0.0027284    2.469    0.01514   *

Residual standard error: 0.4949 on 105 degrees of freedom
Multiple R-Squared: 0.6882,      Adjusted R-squared: 0.6734
F-statistic: 46.36 on 5 and 105 DF, p-value:         0
```

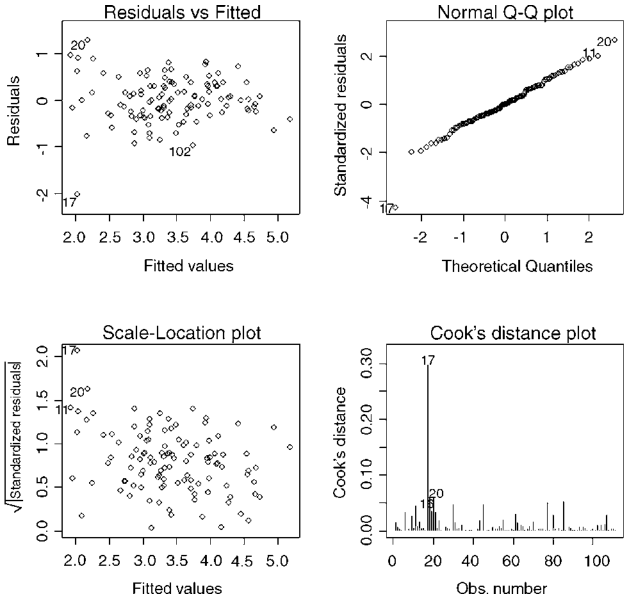
On the log(ozone) scale, there is no evidence for a quadratic term in temperature, so let's remove that:

```
model8 < -update(model7, ~. – I(temp^2))
summary(model8)
```

```
Coefficients:
                Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)     0.7231644   0.6457316   1.120     0.26528
temp            0.0464240   0.0059918   7.748    5.94e-12 ***
wind           -0.2203843   0.0597744  -3.687     0.00036 ***
rad             0.0025295   0.0005404   4.681    8.49e-06 ***
I(wind^2)       0.0072233   0.0026292   2.747     0.00706  **

Residual standard error: 0.4936 on 106 degrees of freedom
Multiple R-Squared: 0.6868,     Adjusted R-squared: 0.675
F-statistic: 58.11 on 4 and 106 DF, p-value:         0
```

```
plot(model8)
```



The heteroscedasticity and the non-normality have been cured, but there is now a highly influential data point (number 17 on the Cook's plot). We should refit the model with this point left out, to see if the parameter estimates or their standard errors are greatly affected:

```
model9<-lm(log(ozone)~temp + wind + rad + I(wind^2),subset=(1:length
(ozone)!=17))
summary(model9)
```

```
Coefficients:
                 Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)     1.1932358   0.5990022    1.992   0.048963    *
temp            0.0419157   0.0055635    7.534   1.81e-11  ***
wind           -0.2208189   0.0546589   -4.040   0.000102  ***
rad             0.0022097   0.0004989    4.429   2.33e-05  ***
I(wind^2)       0.0068982   0.0024052    2.868   0.004993   **
Residual standard error: 0.4514 on 105 degrees of freedom
Multiple R-Squared: 0.6974,     Adjusted R-squared: 0.6859
F-statistic: 60.5 on 4 and 105 DF, p-value:          0
```

Finally, plot(model9) shows that the variance and normality are well behaved, so we can stop at this point. We have found the minimal adequate model. It is on a scale of log(ozone concentration), all the main effects are significant, but there are no interactions, and there is a single quadratic term for wind speed (five parameters in all, with 105 d.f. for error).

A More Complex Example

In the next example we introduce two new difficulties: more explanatory variables and fewer data points. It is another air pollution dataframe, but the response variable in this case is sulphur dioxide concentration. There are six continuous explanatory variables:

```
pollute<-read.table("c:\\temp\\sulphur.dioxide.txt",header=T)
attach(pollute)
names(pollute)
```

```
[ 1] "Pollution"    "Temp"  "Industry"   "Population"   "Wind"
[ 6] "Rain"         "Wet.days"
```

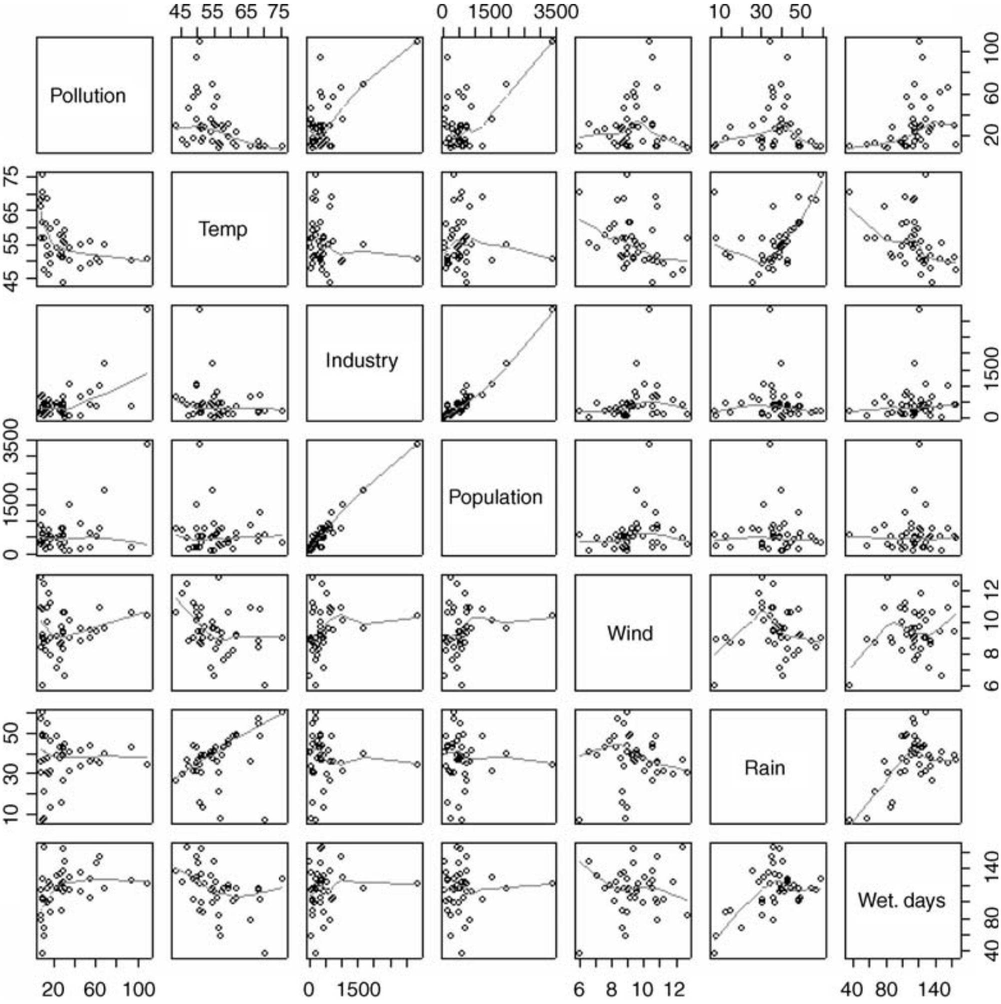
Here are the 36 scatter plots:

```
pairs(pollute,panel=panel.smooth)
```

This time, let's begin with the tree model rather than the generalized additive model. A look at the pairs plots suggests that interactions may be more important than non-linearity in this case.

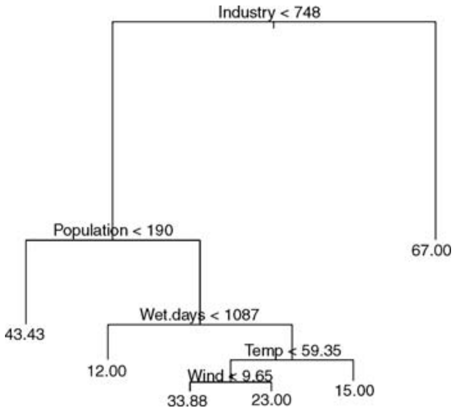
```
library(tree)
model<-tree(Pollution~.,data=pollute)
plot(model)
text(model)
```

This is interpreted as follows. The most important explanatory variable is Industry, and the threshold value separating low and high values of industry is 748. The right-hand branch of the tree indicates the mean value of air pollution for high levels of industry (67.00). The fact that this limb is unbranched means that no other variables explain a



significant amount of the variation in pollution levels for high values of industry. The left-hand limb does not show the mean values of pollution for low values of industry, because there are other significant explanatory variables. Mean values of pollution are only shown at the extreme ends of branches. For low values of industry, the tree shows us that population has a significant impact on air pollution. At low values of population (<190) the mean level of air pollution was 43.43. For high values of population, the number of wet days is significant. Low numbers of wet days (<108) have mean pollution levels of 12.00 while temperature has a significant impact on pollution for places where the number of wet days is large. At high temperatures (>59.35 °F) the mean pollution level was 15.00 while at lower temperatures the run of wind is important. For still air (wind < 9.65) pollution was higher (33.88) than for higher wind speeds (23.00). The virtues of tree-based models are numerous:

- they are easy to appreciate and to describe to other people,
- the most important variables stand out,
- interactions are clearly displayed,
- non-linear effects are captured effectively, and
- the complexity of the behaviour of the explanatory variables is plain to see.



We conclude that the interaction structure is highly complex. We shall need to carry out the linear modelling with considerable care.

Start with some elementary calculations. With six explanatory variables, how many interactions might we fit? Well, there are $5+4+3+2+1=15$ two-way interactions for a start. Plus 20 three-way, 15 four-way and six five-way interactions, plus one six-way interaction for good luck. Then there are quadratic terms for each of the six explanatory variables. So we are looking at about 70 parameters that might be estimated from the data, but how many data points have we got?

```
length(Pollution)
```

```
[ 1] 41
```

Oh dear. We are planning to estimate almost twice as many parameters as there are data points. That's taking over-parameterization to new heights. We already know that you cannot estimate more parameter values than there are data points (i.e. a maximum of 41 parameters); but we also know that when we fit a saturated model to the data, it has no explanatory power (there are no degrees of freedom, so the model, by explaining everything, ends up explaining nothing at all). There is a useful rule of thumb: **don't try to estimate more than n/3 parameters during a multiple regression**. In the present case $n=41$ so the rule of thumb is suggesting that we restrict ourselves to estimating about $41/3 \approx 13$ parameters at any one time. We know from the tree model that the interaction structure is going to be complicated so we shall concentrate on that. We begin, therefore, by looking for curvature, to see if we can eliminate it as a major cause of variation:

```
model1 <-
lm(Pollution ~ Temp + I(Temp^2) + Industry + I(Industry^2) + Population +
  I(Population^2) + Wind + I(Wind^2) + Rain + I(Rain^2) + Wet.days + I(Wet.days^2))
summary(model1)
```

```
Coefficients:
                 Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)     -6.641e+01   2.234e+02  -0.297    0.76844
Temp             5.814e-01   6.295e+00   0.092    0.92708
I(Temp^2)       -1.297e-02   5.188e-02  -0.250    0.80445
Industry         8.123e-02   2.868e-02   2.832    0.00847   **
I(Industry^2)   -1.969e-05   1.899e-05  -1.037    0.30862
Population      -7.844e-02   3.573e-02  -2.195    0.03662    *
I(Population^2) 2.551e-05    2.158e-05   1.182    0.24714
Wind             3.172e+01   2.067e+01   1.535    0.13606
I(Wind^2)       -1.784e+00   1.078e+00  -1.655    0.10912
Rain             1.155e+00   1.636e+00   0.706    0.48575
I(Rain^2)       -9.714e-03   2.538e-02  -0.383    0.70476
Wet.days        -1.048e+00   1.049e+00  -0.999    0.32615
I(Wet.days^2)    4.555e-03   3.996e-03   1.140    0.26398

Residual standard error: 14.98 on 28 degrees of freedom
Multiple R-Squared: 0.7148,       Adjusted R-squared: 0.5925
F-statistic: 5.848 on 12 and 28 DF,  p-value: 5.868e-005
```

So that's our first bit of good news. There is no evidence of curvature for any of the six explanatory variables. Only the main effects of industry and population are significant in this (over-parameterized) model. Now we need to consider the interaction terms. We do not fit interaction terms without both the component main effects, so we cannot fit all the two-way interaction terms at the same time (that would be $15 + 6 = 21$ parameters; well above the rule of thumb value of 13). One approach is to fit the interaction terms in randomly selected sets. With all six main effects, we can afford to assess $13 - 6 = 7$ interaction terms at a time, so we'll try this. Make a vector containing the names of the 15 two-way interactions:

```
interactions <-c("ti","tp","tw","tr","td","ip","iw","ir","id","pw","pr","pd","wr","wd", "rd")
```

Now shuffle the interactions into random order using sample without replacement:

```
sample(interactions)
```

```
[1] "wr" "wd" "id" "ir" "rd" "pr" "tp" "pw" "ti" "iw" "tw" "pd" "tr" "td" "ip"
```

It would be pragmatic to test the two-way interactions in three models each containing five two-way interaction terms:

```
model2 < -
lm(Pollution ~ Temp + Industry + Population + Wind + Rain + Wet.days + Wind:Rain +
  Wind: Wet.days + Industry:Wet.days + Industry:Rain + Rain:Wet.days)
model3 < -
lm(Pollution ~ Temp + Industry + Population + Wind + Rain + Wet.days + Population:
  Rain  + Temp:Population + Population:Wind + Temp:Industry + Industry:Wind)
model4 < -
lm(Pollution ~ Temp + Industry + Population + Wind + Rain + Wet.days + Temp:Wind +
  Population:Wet.days + Temp:Rain + Temp:Wet.days + Industry:Population)
```

Extracting only the interaction terms from the three models, we see:

```
Industry:Rain        -1.616e-04   9.207e-04   -0.176   0.861891
Industry:Wet.days     2.311e-04   3.680e-04    0.628   0.534949
Wind:Rain             9.049e-01   2.383e-01    3.798   0.000690 ***
Wind:Wet.days        -1.662e-01   5.991e-02   -2.774   0.009593  **
Rain:Wet.days         1.814e-02   1.293e-02    1.403   0.171318

Temp:Industry        -1.643e-04   3.208e-03   -0.051     0.9595
Temp:Population       1.125e-03   2.382e-03    0.472     0.6402
Industry:Wind         2.668e-02   1.697e-02    1.572     0.1267
Population:Wind      -2.753e-02   1.333e-02   -2.066     0.0479   *
Population:Rain       6.898e-04   1.063e-03    0.649     0.5214

Temp:Wind             1.261e-01   2.848e-01    0.443    0.66117
Temp:Rain            -7.819e-02   4.126e-02   -1.895   0.06811.
Temp:Wet.days         1.934e-02   2.522e-02    0.767    0.44949
Industry:Population   1.441e-06   4.178e-06    0.345    0.73277
Population:Wet.days   1.979e-05   4.674e-04    0.042    0.96652
```

The next step might be to put all of the significant or close-to-significant interactions into the same model, and see which survive:

```
model5 < -
lm(Pollution ~ Temp + Industry + Population + Wind + Rain + Wet.days + Wind:Rain +
  Wind: Wet.days + Population:Wind + Temp:Rain)
summary(model5)
```

```
Coefficients:
                       Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)          323.054546  151.458618    2.133  0.041226   *
Temp                  -2.792238    1.481312   -1.885  0.069153   .
Industry               0.073744    0.013646    5.404  7.44e-06 ***
Population             0.008314    0.056406    0.147  0.883810
Wind                 -19.447031    8.670820   -2.243  0.032450   *
Rain                  -9.162020    3.381100   -2.710  0.011022   *
Wet.days               1.290201    0.561599    2.297  0.028750   *
Temp:Rain              0.017644    0.027311    0.646  0.523171
Population:Wind       -0.005684    0.005845   -0.972  0.338660
Wind:Rain              0.997374    0.258447    3.859  0.000562 ***
Wind:Wet.days         -0.140606    0.053582   -2.624  0.013530   *
```

We certainly do not need Temp:Rain

```
model6 < -update(model5, ~. –Temp:Rain)
```

or Population:Wind

```
model7 < -update(model6, ~. –Population:Wind)
```

All the terms in model 7 are significant. Time for a check on the behaviour of the model:

```
plot(model7)
```

That's not bad at all, but what about the higher-order interactions? One way to proceed is to specify the interaction level using ^3 in the model formula, but if you do this, you will find that we run out of degrees of freedom straight away. A sensible option is to fit three-way terms for the variables that already appear in two-way interactions – in our case, that is just one term: Wind:Rain:Wet.days

```
model8 < -update(model7, ~. + Wind:Rain:Wet.days)
summary(model8)
```

```
Coefficients:
                        Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)           278.464474   68.041497    4.093   0.000282 ***
Temp                   -2.710981    0.618472   -4.383   0.000125 ***
Industry                0.064988    0.012264    5.299    9.1e-06 ***
Population             -0.039430    0.011976   -3.293   0.002485  **
Wind                   -7.519344    8.151943   -0.922   0.363444
Rain                   -6.760530    1.792173   -3.772   0.000685 ***
Wet.days                1.266742    0.517850    2.446   0.020311   *
Wind:Rain               0.631457    0.243866    2.589   0.014516   *
Wind:Wet.days          -0.230452    0.069843   -3.300   0.002440  **
Wind:Rain:Wet.days      0.002497    0.001214    2.056   0.048247   *

Residual standard error: 11.2 on 31 degrees of freedom
Multiple R-Squared: 0.8236,       Adjusted R-squared: 0.7724
F-statistic: 16.09 on 9 and 31 DF,  p-value:  2.231e-009
```

That's enough for now. You are probably getting the idea. Multiple regression is difficult, time consuming and always vulnerable to subjective decisions about what to include and what to leave out. The linear modelling confirms the early impression from the tree model: for low levels of industry, the SO_2 level depends in a simple way on population (people tend to want to live where the air is clean) and in a complicated way on daily weather (the three-way interaction between wind, total rainfall and the number of wet days (i.e. on rainfall intensity).

Automating the Process of Model Simplification Using step

In a model with many interaction terms or a large number of explanatory variables, the procedure of model simplification can be very time-consuming. Help is at hand, however, in the form of the step function. The complex model1 is automatically simplified to model2 like this:

```
model2<-step(model1)
```

You can control whether the procedure steps “backwards”, “forwards” or “both”, and you can fix the most complex (“upper”) and most simple (“lower”) models between which simplification is carried out. The criterion used for dropping terms from the model is AIC; the smaller the AIC, the better the fit (see below).

Typically, step is generous in the sense that it leaves close-to-significant terms in the model. Therefore, the simplified model2 needs to be subjected to manual model simplification using update in order to arrive at a minimal adequate model which contains nothing but significant terms. You should not use step to simplify complex contingency table models without specifying “lower”, because step could eliminate nuisance variables that need to be retained in the model to constrain the marginal totals.

AIC (Akaike’s Information Criterion)

As you add parameters to a model you inevitably improve the fit. In the limit, you would have a parameter for every data point, and the fit of the model to the data would be perfect (see p. 153). There is always a trade-off between model simplicity and fit, and the ideal model is typically a compromise between these two. One way of determining whether extra parameters are justified is to use AIC. In the jargon, this is a penalized log-likelihood. It is like a deviance (−2*log-likelihood), but with a ‘penalty’ of 2 added to the score for every extra parameter in the model:

$$\text{AIC} = -2^{*}\text{log-likelihood} + 2\text{p}$$

where p represents the number of parameters in the fitted model. It is useful in model simplification because a model with lower AIC is preferred to one with a higher AIC, and there are built-in tests for assessing the significance of the difference between two AICs. Unless an additional parameter causes a reduction in deviance of at least 2.0 then AIC will not decrease, and the additional parameter will not we warranted.