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?

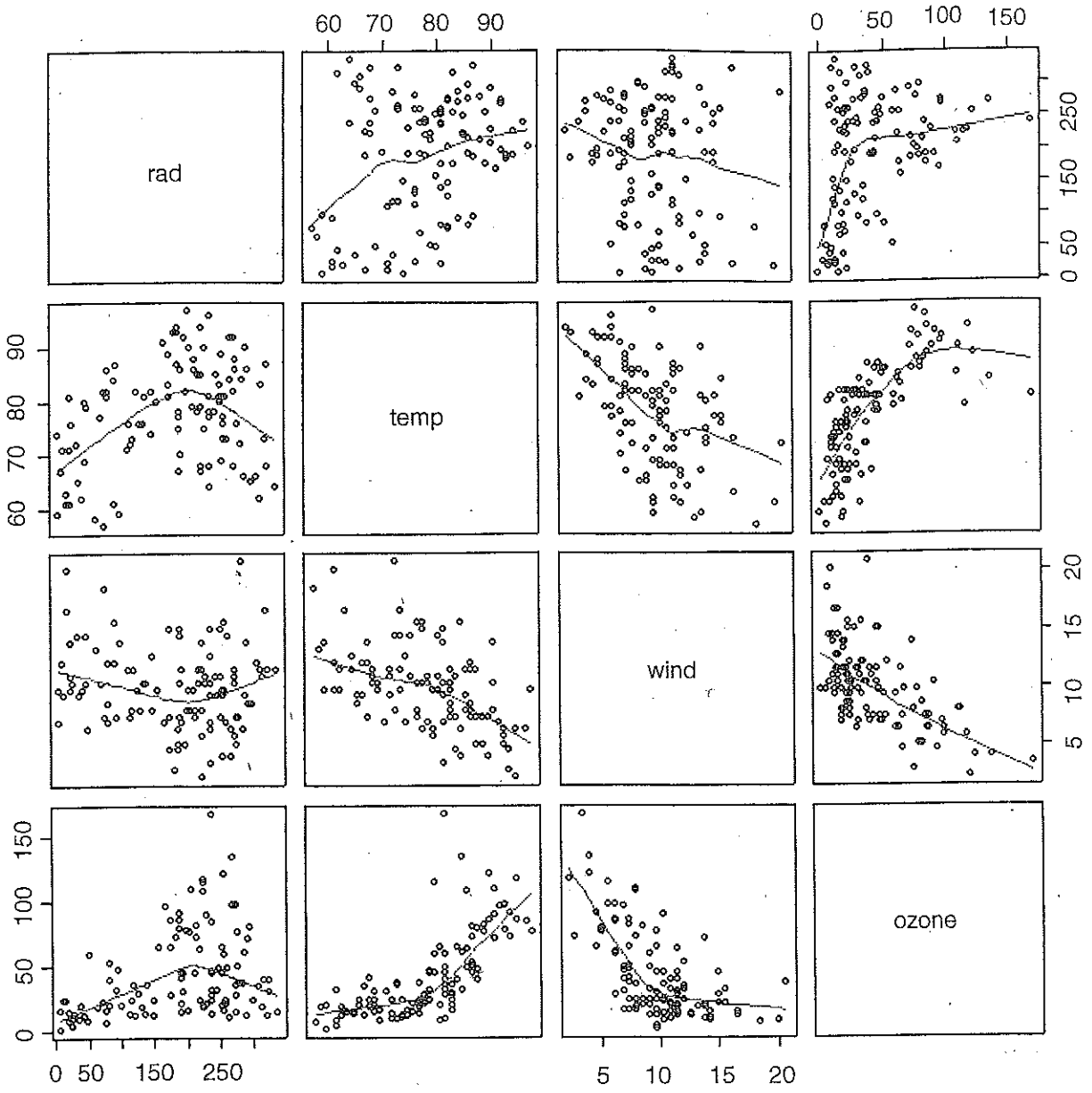
```
ozone.pollution<-read.table("c:\\temp\\ozone.data.txt",header=T)
attach(ozone.pollution)
names(ozone.pollution)
```

```
[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
 ISBNs: 0-470-02298-1 (PBK); 0-470-02297-3 (PPC)

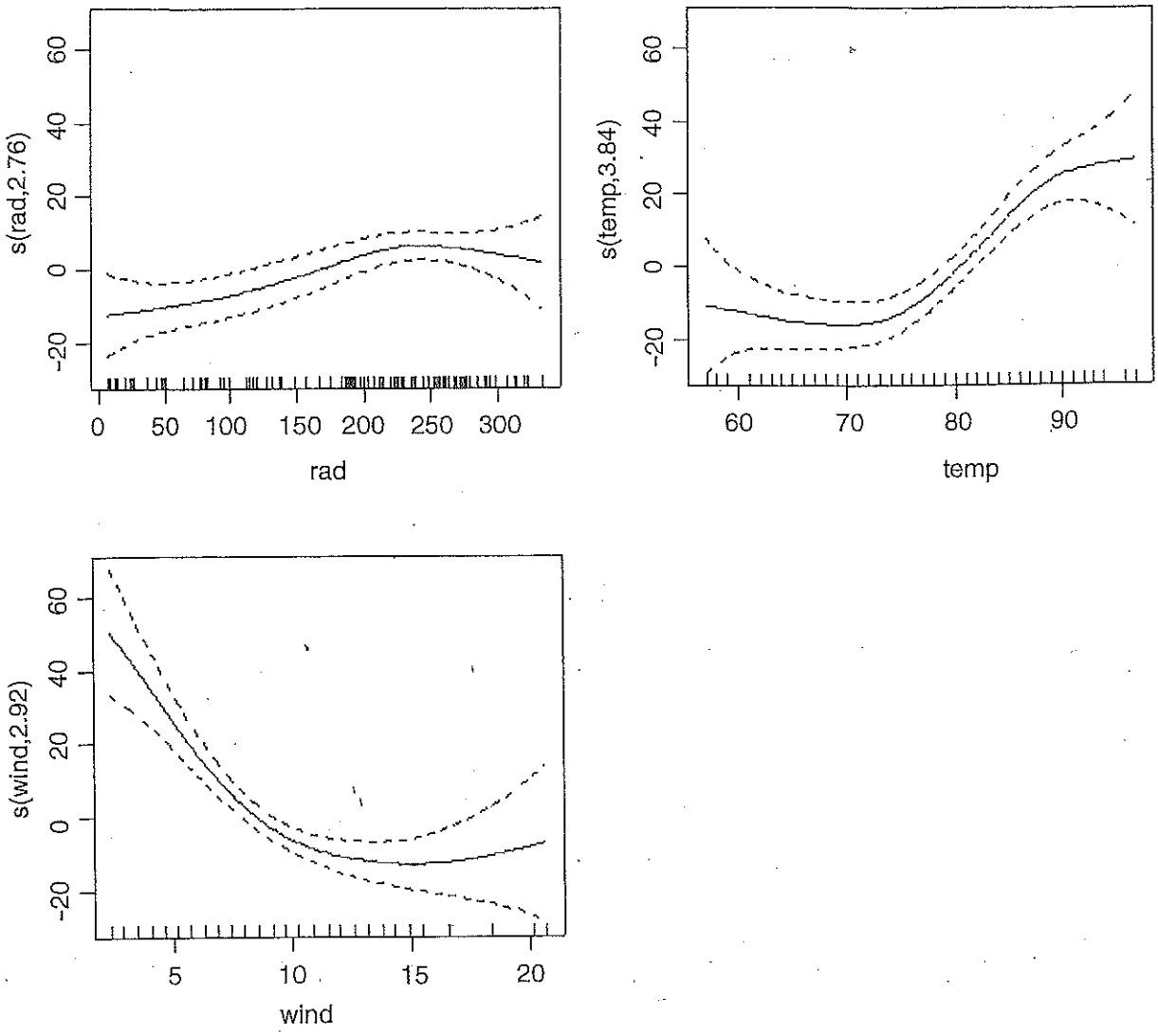


The response variable, ozone concentration, is shown on the *y* axis of the bottom row of panels: there is a strong negative relationship with wind speed, a positive correlation with temperature and a rather unclear, but possibly humped relationship with radiation.

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))
```

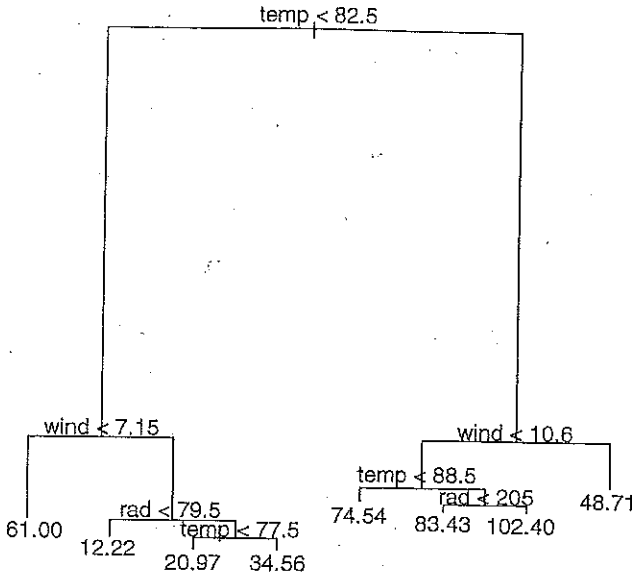
The confidence intervals are sufficiently narrow to suggest that the curvature in the relationship between ozone and temperature is real, but the curvature of the relationship with wind is questionable, and a linear model may well be all that is required for solar



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

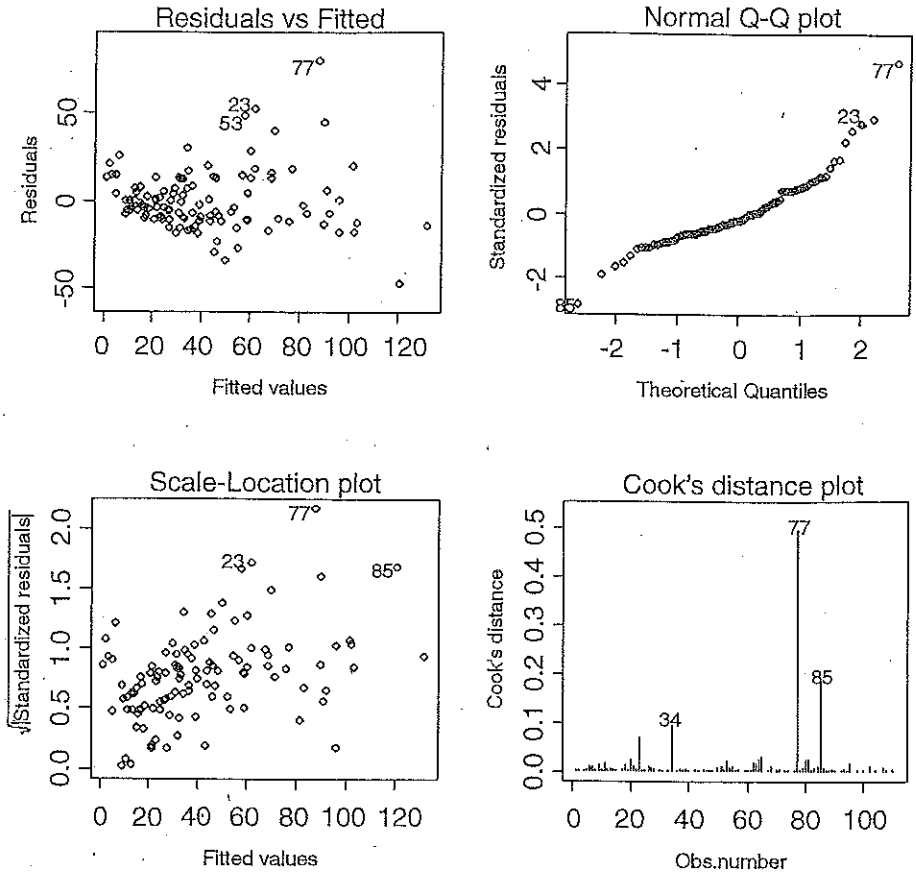
```
model6<-update(model5,~.-temp:rad)
summary(model6)
```

```
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
```

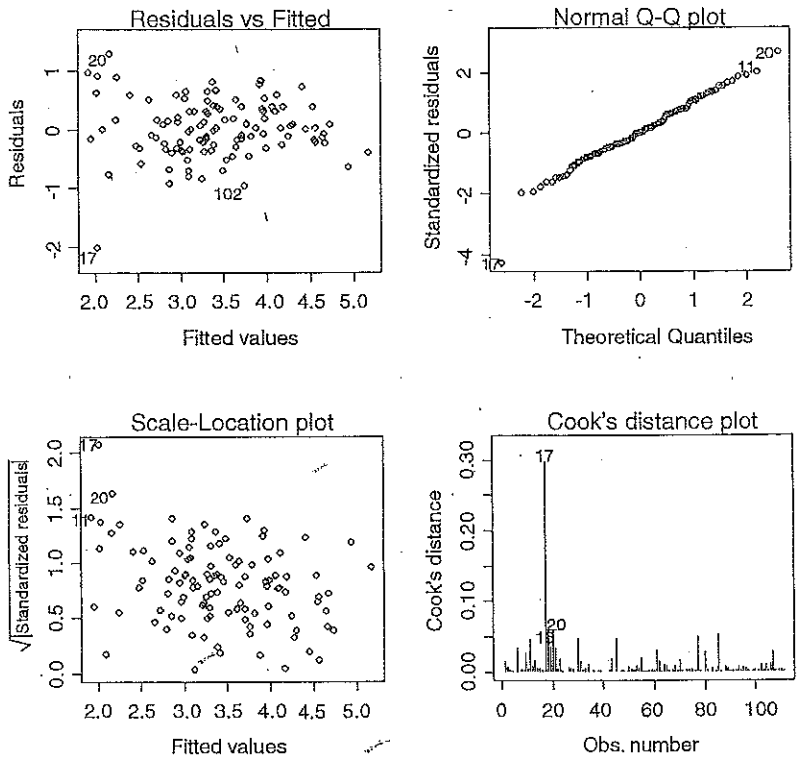
On the log(ozone) scale, there is no evidence for a quadratic term in temperature, so let's remove that:

```
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).

Update in Model Simplification

In the update function used during model simplification, the dot '.' is used to specify 'what is there already' on either side of the tilde. So if your original model said

```
model<-lm(y~A*B)
```

then the update function to remove the interaction term A:B could be written like this:

```
model2<-update(model,~.-A:B)
```

Note that there is no need to repeat the name of the response variable, and the punctuation 'tilde dot' means take model as it is, and remove from it ('minus') the interaction term A:B.

Examples of R Model Formulae

Model	Model formula	Comments
Null	y~1	1 is the intercept in regression models, but here it is the overall mean *y*
Regression	y~x	*x* is a continuous explanatory variable
One-way Anova	y~gender	Gender is a two-level categorical variable
Two-way Anova	y~gender + genotype	Genotype is a four-level categorical variable
Factorial Anova	y~N * P * K	N, P and K are two-level factors to be fit along with all their interactions
Three-way Anova	y~N*P*K-N:P:K	As above, but don't fit the three-way interaction