Research Methods in English Linguistics
Multiple Regression: Model reduction

Hyunah Ahn

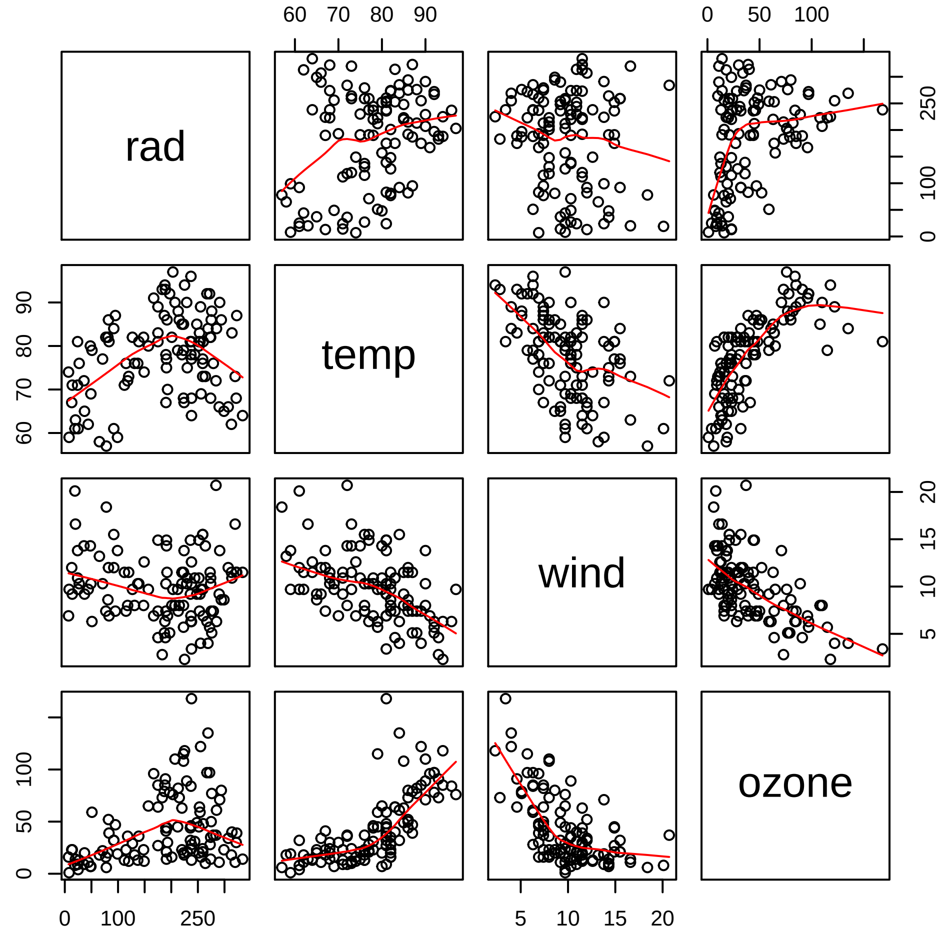
21 November 2019

```
setwd("C:/Users/hyuna/OneDrive/Documents/01TeachingResources/01SNU/03GraduateSeminar/crawley")
```

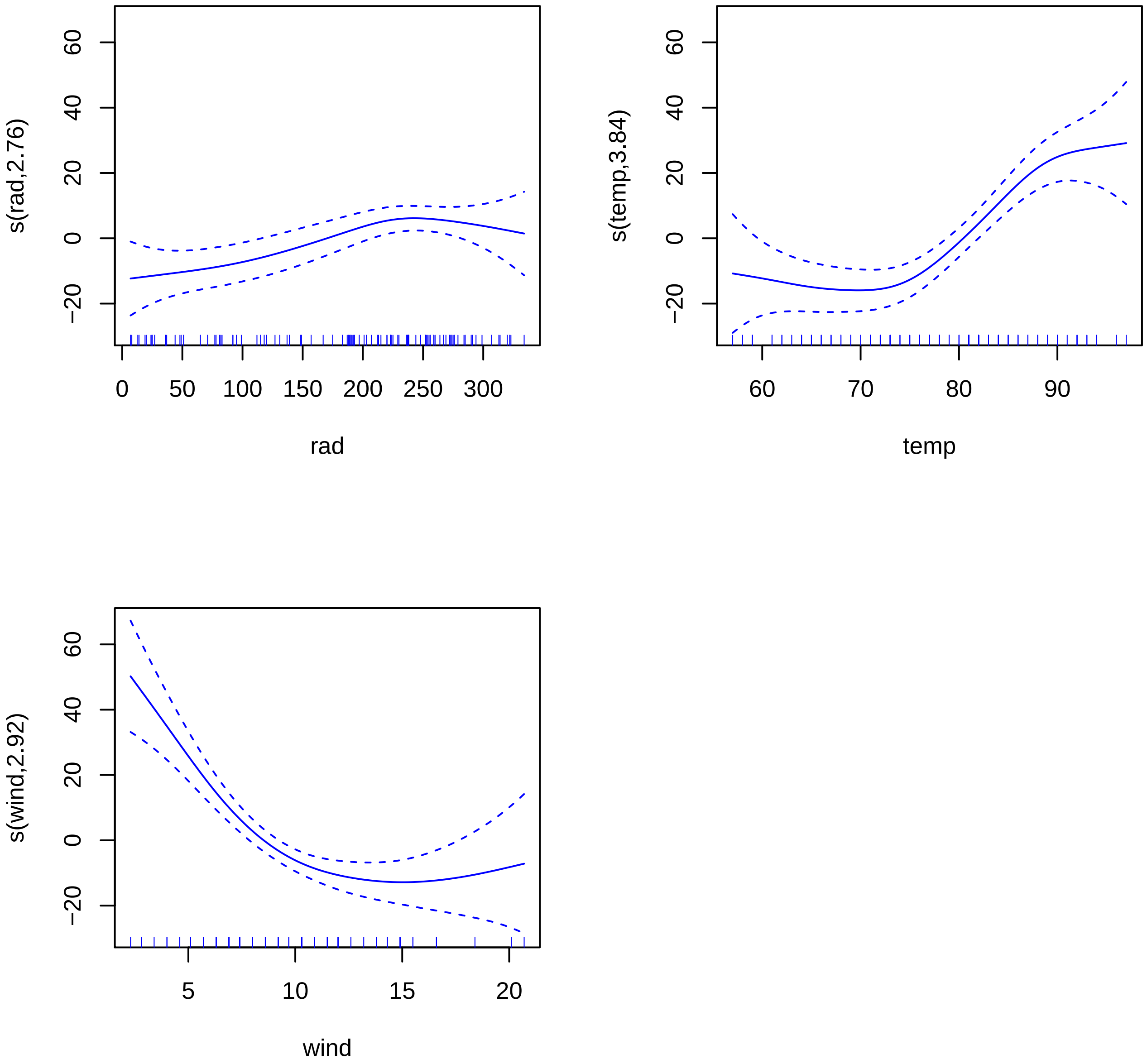
```
library(mgcv)
library(tree)
library(plot3D)
```

```
ozone.pollution←read.csv("data/ozone.data.csv")
```

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



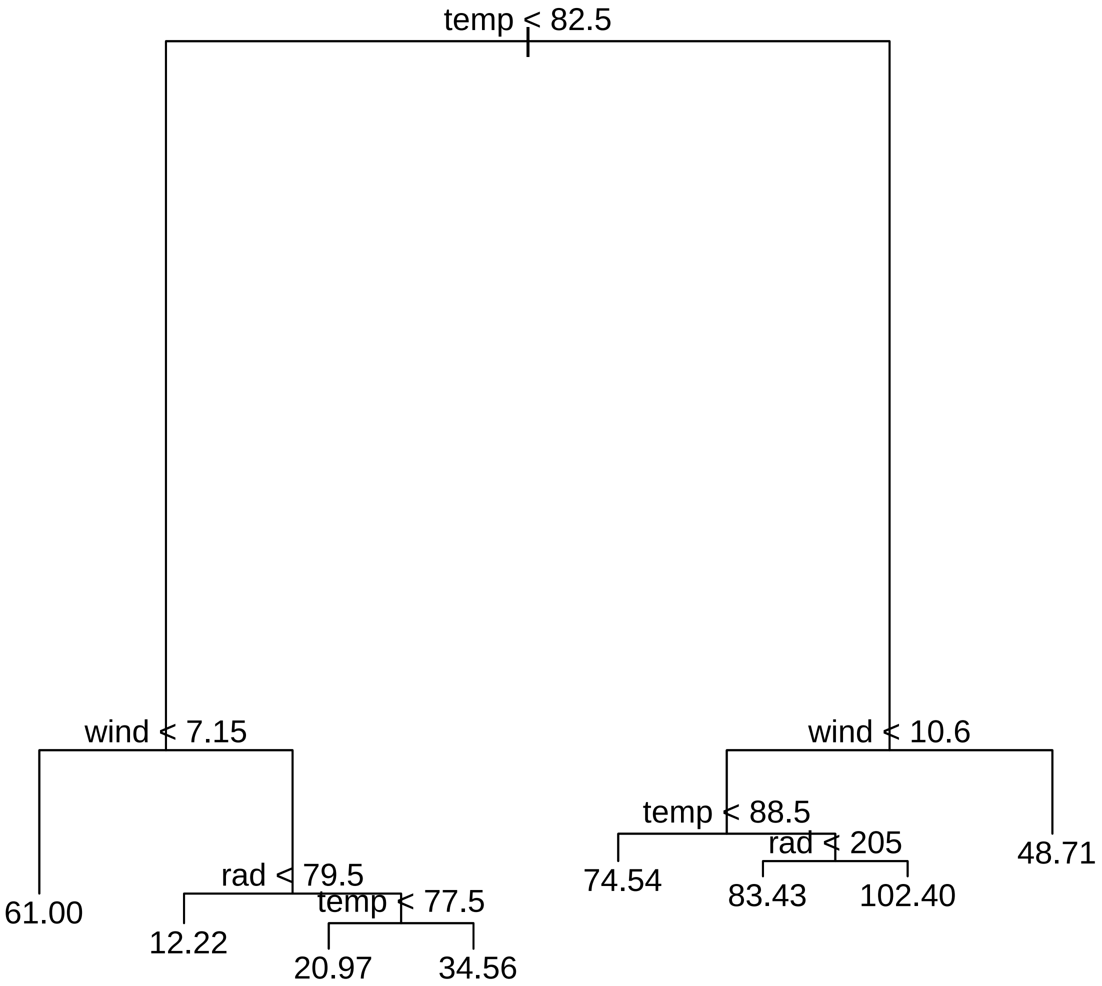
```
par(mfrow=c(2,2))
model.gam←gam(ozone∼s(rad)+s(temp)+s(wind))
plot(model.gam, col = ”blue”)
```



Useful rules to remember when deciding on a model:

- All models are wrong.
- Some models are better than others.
- The correct model can never be known with certainty.
- The simple the model the better it is.

```
par(mfrow=c(1,1))
model.tree←tree(ozone~., data=ozone.pollution)
plot(model.tree)
text(model.tree)
```



Rules of parsimony: We prefer...

- A model with n-1 parameters to a with n parameters
- A model with k-1 explanatory variables to a model with k explanatory variables
- A linear model to a curved one
- A model without a hump than one with a hump
- A model without interactions that one with interactions

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

```

Call:
lm(formula = ozone ~ temp * wind * rad + I(rad^2) + I(temp^2) +
    I(wind^2))

Residuals:
    Min      1Q  Median      3Q     Max
-38.894 -11.205  -2.736   8.809  70.551

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       8.403e-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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.82 on 100 degrees of freedom
Multiple R^2:  0.7394,    Adjusted R^2:  0.7133
F-statistic: 28.37 on 10 and 100 DF,  p-value: < 2.2e-16
```

```
model2←update(model1, ~. −temp:wind:rad)
summary(model2)
```

```

Call:
lm(formula = ozone ~ temp + wind + rad + I(rad^2) + I(temp^2) +
    I(wind^2) + temp:wind + temp:rad + wind:rad)

Residuals:
    Min      1Q  Median      3Q     Max
-39.611 -11.455  -2.901   8.548  70.325

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.245e+02  1.957e+02   2.680   0.0086 **
temp        -1.021e+01  4.209e+00  -2.427   0.0170 *
wind        -2.802e+01  9.645e+00  -2.906   0.0045 **
rad          2.628e-02  2.142e-01   0.123   0.9026
I(rad^2)    -3.388e-04  2.541e-04  -1.333   0.1855
I(temp^2)    5.953e-02  2.382e-02   2.499   0.0141 *
I(wind^2)    6.173e-01  1.461e-01   4.225 5.25e-05 ***
temp:wind    1.734e-01  9.497e-02   1.825   0.0709 .
temp:rad     3.750e-03  2.459e-03   1.525   0.1303
wind:rad    -1.127e-02  6.277e-03  -1.795   0.0756 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.77 on 101 degrees of freedom
Multiple R^2:  0.7383,    Adjusted R^2:  0.715
F-statistic: 31.66 on 9 and 101 DF,  p-value: < 2.2e-16
```

```
model3←update(model2, ~. −wind:rad)
summary(model3)
```

```

Call:
lm(formula = ozone ~ temp + wind + rad + I(rad^2) + I(temp^2) +
    I(wind^2) + temp:wind + temp:rad)

Residuals:
    Min      1Q  Median      3Q     Max
-43.174 -11.020  -4.077   7.316  74.787

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.832e+02  1.964e+02   2.460 0.015592 *
temp        -9.069e+00  4.205e+00  -2.157 0.033391 *
wind        -2.472e+01  9.570e+00  -2.583 0.011223 *
rad         -1.812e-01  1.823e-01  -0.994 0.322483
I(rad^2)    -3.438e-04  2.569e-04  -1.338 0.183762
I(temp^2)    5.461e-02  2.392e-02   2.283 0.024507 *
I(wind^2)    5.809e-01  1.463e-01   3.972 0.000133 ***
temp:wind    1.137e-01  8.993e-02   1.264 0.208995
temp:rad     4.925e-03  2.396e-03   2.055 0.042402 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.96 on 102 degrees of freedom
Multiple R^2:  0.7299,    Adjusted R^2:  0.7087
F-statistic: 34.46 on 8 and 102 DF,  p-value: < 2.2e-16
```

```
model4←update(model3, ~. −temp:wind)
summary(model4)
```

```

Call:
lm(formula = ozone ~ temp + wind + rad + I(rad^2) + I(temp^2) +
    I(wind^2) + temp:rad)

Residuals:
    Min      1Q  Median      3Q     Max
-44.258 -11.174  -3.325   9.562  78.416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.699e+02  1.010e+02   2.673  0.00874 **
temp        -5.090e+00  2.797e+00  -1.820  0.07173 .
wind        -1.296e+01  2.276e+00  -5.695 1.17e-07 ***
rad         -1.902e-01  1.827e-01  -1.041  0.30013
I(rad^2)    -2.994e-04  2.552e-04  -1.173  0.24348
I(temp^2)    3.650e-02  1.921e-02   1.900  0.06027 .
I(wind^2)    4.454e-01  9.979e-02   4.463 2.07e-05 ***
temp:rad     4.857e-03  2.403e-03   2.022  0.04578 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.01 on 103 degrees of freedom
Multiple R^2:  0.7257,    Adjusted R^2:  0.707
F-statistic: 38.93 on 7 and 103 DF,  p-value: < 2.2e-16
```

```
model5←update(model4, ~. −I(rad^2))
summary(model5)
```

```

Call:
lm(formula = ozone ~ temp + wind + rad + I(temp^2) + I(wind^2) +
    temp:rad)

Residuals:
    Min      1Q  Median      3Q     Max
-43.764 -11.157  -3.327   8.499  78.851

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 262.651900 100.961024   2.602   0.0106 *
temp         -4.902890   2.797877  -1.752   0.0827 .
wind        -13.048559   2.278668  -5.726 1.00e-07 ***
rad          -0.253116   0.174922  -1.447   0.1509
I(temp^2)     0.036480   0.019248   1.895   0.0608 .
I(wind^2)     0.446673   0.099963   4.468 2.01e-05 ***
temp:rad      0.004343   0.002366   1.835   0.0693 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 18.04 on 104 degrees of freedom
Multiple R^2:  0.722,    Adjusted R^2:  0.706
F-statistic: 45.02 on 6 and 104 DF,  p-value: < 2.2e-16
```

```
model6←update(model5, ~. −temp:rad)
summary(model6)
```

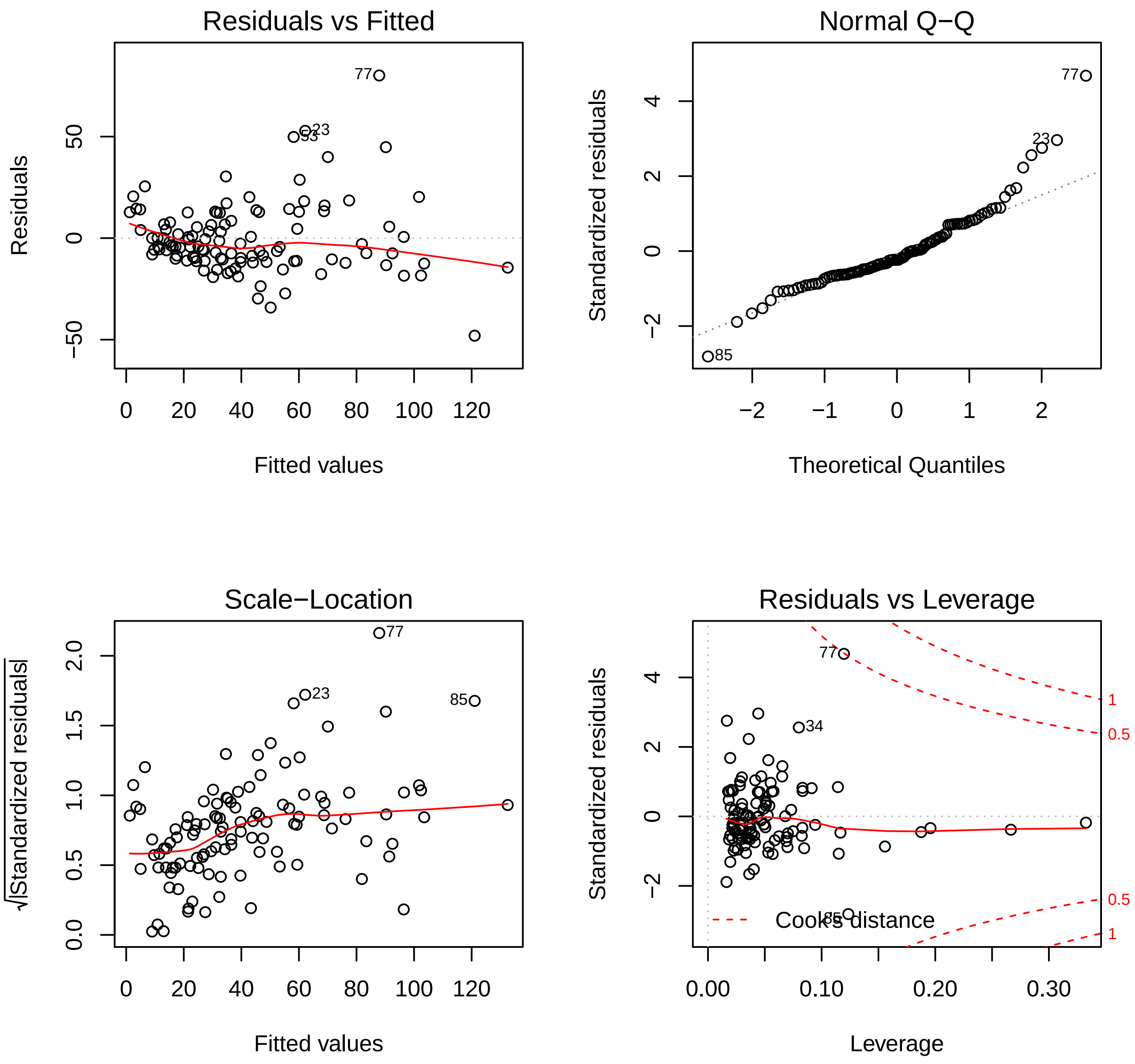
```

Call:
lm(formula = ozone ~ temp + wind + rad + I(temp^2) + I(wind^2))

Residuals:
    Min      1Q  Median      3Q     Max
-48.044 -10.796  -4.138   8.131  80.098

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

Residual standard error: 18.25 on 105 degrees of freedom
Multiple R^2:  0.713,    Adjusted R^2:  0.6994
F-statistic: 52.18 on 5 and 105 DF,  p-value: < 2.2e-16
```

The order of removal in model simplification: Remove in the following order

- Non-significant interaction terms
- Non-significant quadratic or other non-linear terms
- Non-significant explanatory variables
- group together factor levels that do not differ from one another
- in ANCOVA, set non-significant slopes of continuous explanatory variables to zero

Model reduction requires that such simplication does not result in significant reductions in explanatory power.

```
par(mfrow=c(2,2))
plot(model6)
```



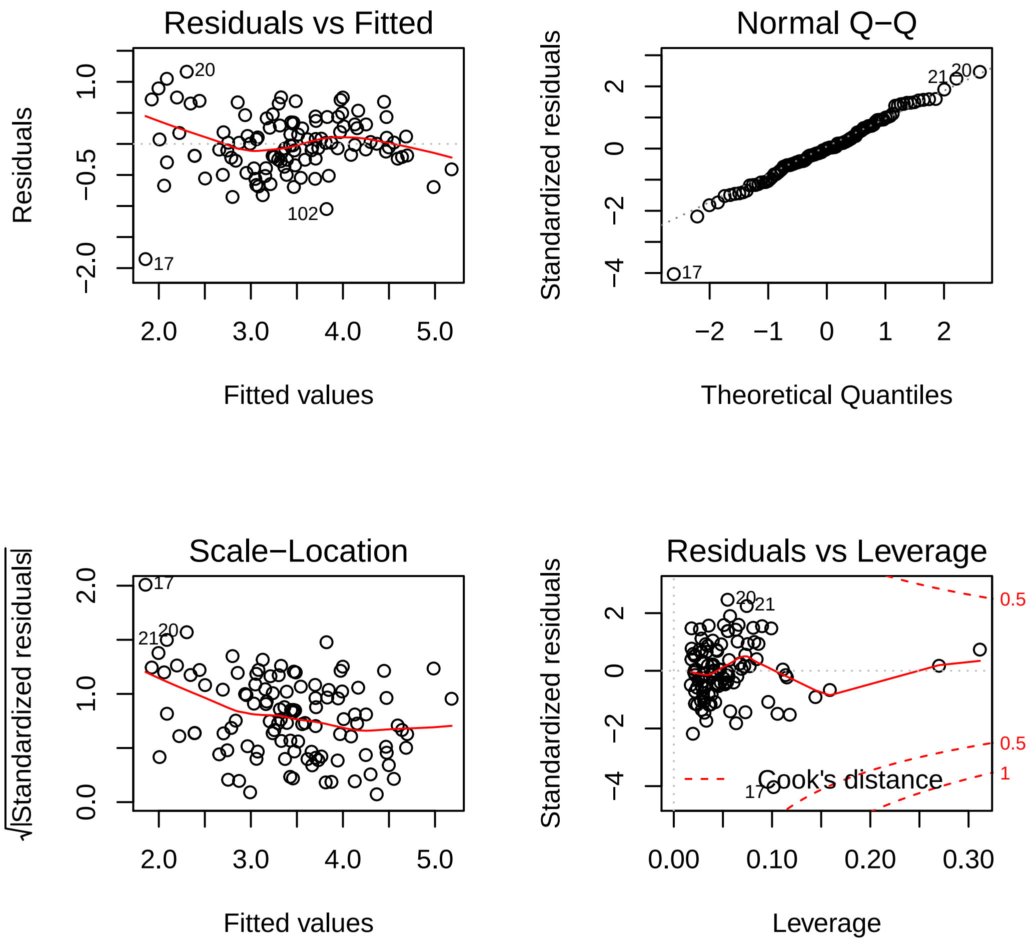
```
model7←lm(log(ozone)~temp*wind*rad+I(rad^2)+I(temp^2)+I(wind^2))
model8←step(model7)
```

```
Start:  AIC=-148.98
log(ozone) ~ temp * wind * rad + I(rad^2) + I(temp^2) + I(wind^2)

                Df Sum of Sq    RSS     AIC
- I(temp^2)      1   0.20130 23.988 -150.05
<none>                       23.787 -148.98
- temp:wind:rad  1   0.46883 24.256 -148.82
- I(rad^2)       1   1.06316 24.850 -146.13
- I(wind^2)      1   1.12186 24.909 -145.87

Step:  AIC=-150.05
log(ozone) ~ temp + wind + rad + I(rad^2) + I(wind^2) + temp:wind +
    temp:rad + wind:rad + temp:wind:rad

                Df Sum of Sq    RSS     AIC
- temp:wind:rad  1   0.42563 24.414 -150.10
<none>                       23.988 -150.05
- I(wind^2)      1   0.92801 24.916 -147.84
- I(rad^2)       1   1.00480 24.993 -147.49

Step:  AIC=-150.1
log(ozone) ~ temp + wind + rad + I(rad^2) + I(wind^2) + temp:wind +
    temp:rad + wind:rad

            Df Sum of Sq    RSS     AIC
- temp:wind  1   0.01438 24.428 -152.03
- temp:rad   1   0.09359 24.508 -151.67
- wind:rad   1   0.11815 24.532 -151.56
<none>                   24.414 -150.10
- I(wind^2)  1   0.87300 25.287 -148.20
- I(rad^2)   1   1.22558 25.639 -146.66

Step:  AIC=-152.03
log(ozone) ~ temp + wind + rad + I(rad^2) + I(wind^2) + temp:rad +
    wind:rad

            Df Sum of Sq    RSS     AIC
- temp:rad   1   0.08429 24.512 -153.65
- wind:rad   1   0.10377 24.532 -153.56
<none>                   24.428 -152.03
- I(rad^2)   1   1.21142 25.640 -148.66
- I(wind^2)  1   1.40005 25.828 -147.84

Step:  AIC=-153.65
log(ozone) ~ temp + wind + rad + I(rad^2) + I(wind^2) + wind:rad

            Df Sum of Sq    RSS     AIC
- wind:rad   1    0.1942 24.707 -154.77
<none>                   24.513 -153.65
- I(rad^2)   1    1.1311 25.644 -150.64
- I(wind^2)  1    1.5001 26.013 -149.06
- temp       1   10.7274 35.240 -115.36

Step:  AIC=-154.77
log(ozone) ~ temp + wind + rad + I(rad^2) + I(wind^2)

            Df Sum of Sq    RSS     AIC
<none>                   24.707 -154.77
- I(rad^2)   1    1.1216 25.828 -151.84
- I(wind^2)  1    1.9234 26.630 -148.45
- rad        1    2.4314 27.138 -146.35
- wind       1    3.3350 28.042 -142.72
- temp       1   10.6366 35.343 -117.03
```

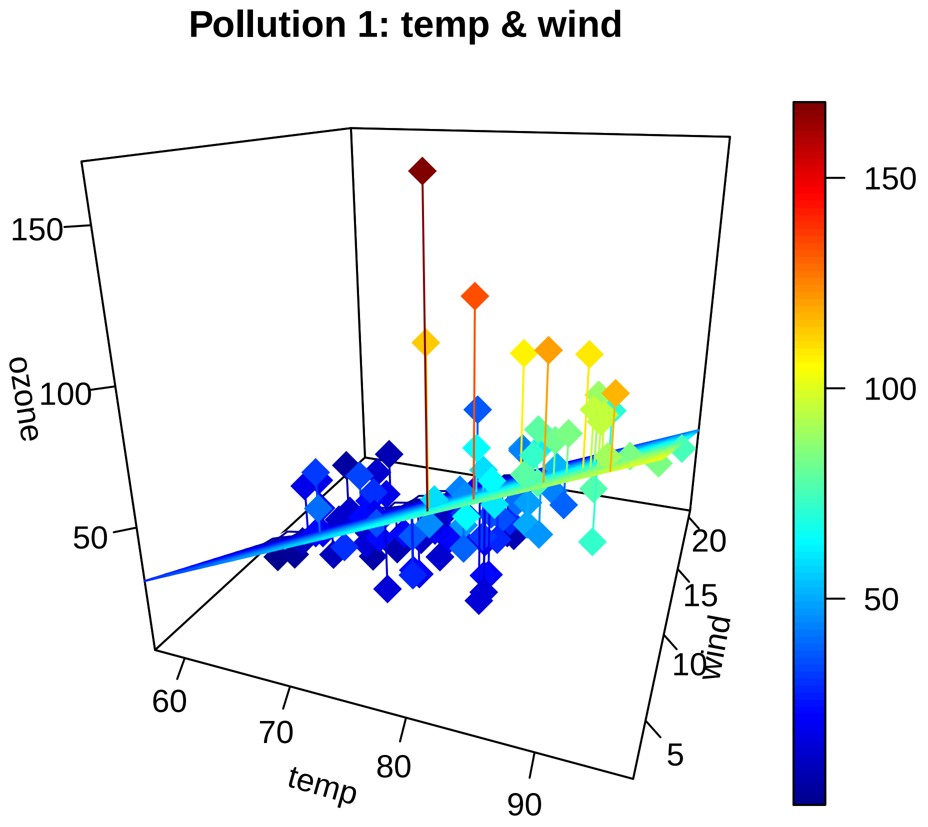
```
summary(model8)
```

```
Call:
lm(formula = log(ozone) ~ temp + wind + rad + I(rad^2) + I(wind^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.85551 -0.25578  0.00248  0.31349  1.16251 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.724e-01  6.350e-01   1.216 0.226543    
temp         4.193e-02  6.237e-03   6.723 9.52e-10 ***
wind        -2.211e-01  5.874e-02  -3.765 0.000275 ***
rad          7.466e-03  2.323e-03   3.215 0.001736 ** 
I(rad^2)    -1.470e-05  6.734e-06  -2.183 0.031246 *  
I(wind^2)    7.390e-03  2.585e-03   2.859 0.005126 ** 
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.4851 on 105 degrees of freedom
Multiple R^2:  0.7004,    Adjusted R^2:  0.6861 
F-statistic:  49.1 on 5 and 105 DF,  p-value: < 2.2e-16
```

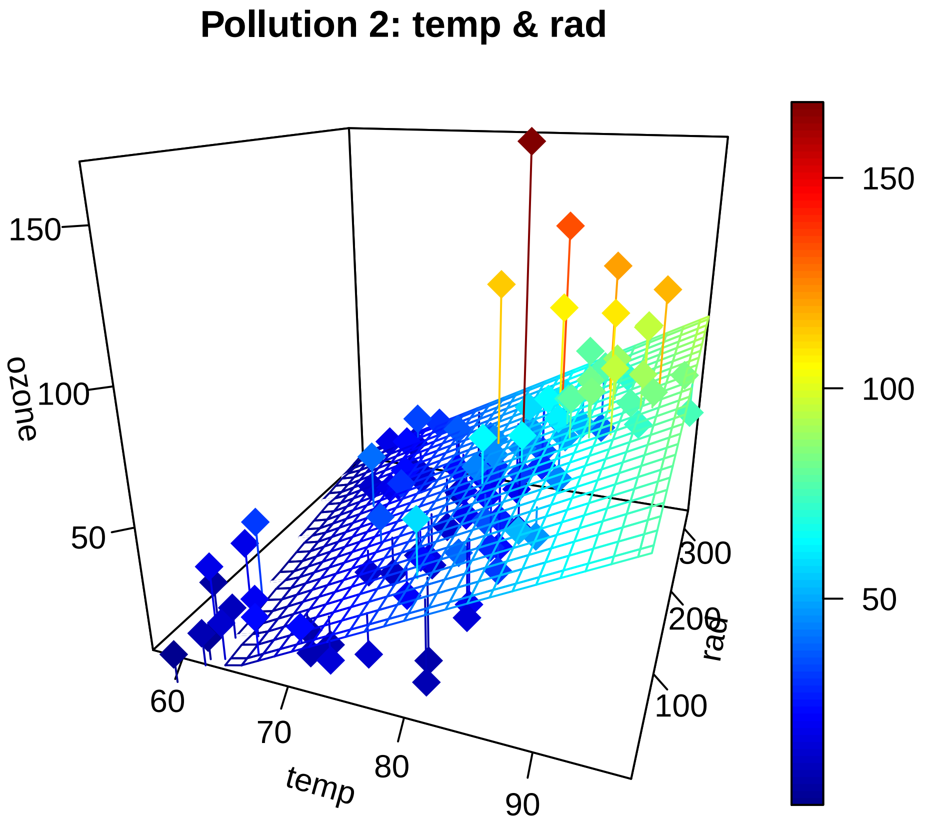
```
par(mfrow=c(2,2))
plot(model8)
```



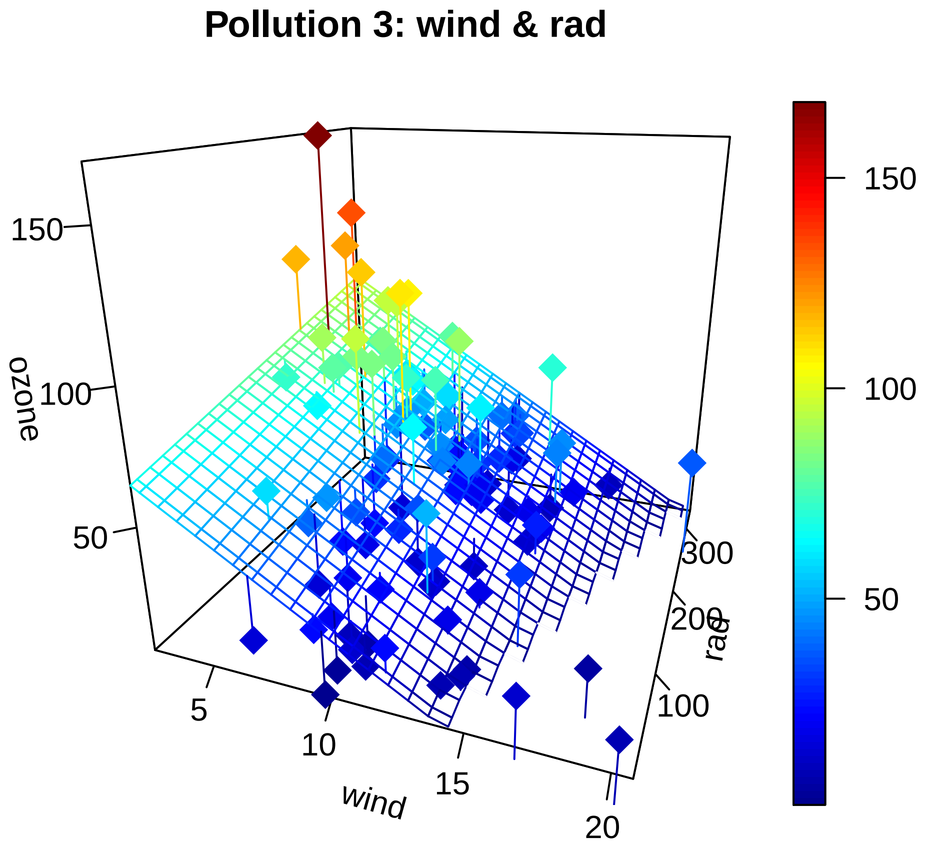
```
# x, y, z variables
x ← temp
y ← wind
z ← ozone
# Compute the linear regression (z = ax + by + d)
fit ← lm(z ∼ x + y)
# predict values on regular xy grid
grid.lines = 26
x.pred ← seq(min(x), max(x), length.out = grid.lines)
y.pred ← seq(min(y), max(y), length.out = grid.lines)
xy ← expand.grid( x = x.pred, y = y.pred)
z.pred ← matrix(predict(fit, newdata = xy),
                 nrow = grid.lines, ncol = grid.lines)
# fitted points for droplines to surface
fitpoints ← predict(fit)
# scatter plot with regression plane
scatter3D(x, y, z, pch = 18, cex = 2,
    theta = 20, phi = 20, ticktype = ”detailed”,
    xlab = ”temp”, ylab = ”wind”, zlab = ”ozone”,
    surf = list(x = x.pred, y = y.pred, z = z.pred,
    facets = NA, fit = fitpoints), main = ”Pollution 1: temp & wind”)
```



```
# x, y, z variables
x ← temp
y ← rad
z ← ozone
# Compute the linear regression (z = ax + by + d)
fit ← lm(z ∼ x + y)
# predict values on regular xy grid
grid.lines = 26
x.pred ← seq(min(x), max(x), length.out = grid.lines)
y.pred ← seq(min(y), max(y), length.out = grid.lines)
xy ← expand.grid( x = x.pred, y = y.pred)
z.pred ← matrix(predict(fit, newdata = xy),
                 nrow = grid.lines, ncol = grid.lines)
# fitted points for droplines to surface
fitpoints ← predict(fit)
# scatter plot with regression plane
scatter3D(x, y, z, pch = 18, cex = 2,
    theta = 20, phi = 20, ticktype = ”detailed”,
    xlab = ”temp”, ylab = ”rad”, zlab = ”ozone”,
    surf = list(x = x.pred, y = y.pred, z = z.pred,
    facets = NA, fit = fitpoints), main = ”Pollution 2: temp & rad”)
```



```
# x, y, z variables
x ← wind
y ← rad
z ← ozone
# Compute the linear regression (z = ax + by + d)
fit ← lm(z ∼ x + y)
# predict values on regular xy grid
grid.lines = 26
x.pred ← seq(min(x), max(x), length.out = grid.lines)
y.pred ← seq(min(y), max(y), length.out = grid.lines)
xy ← expand.grid( x = x.pred, y = y.pred)
z.pred ← matrix(predict(fit, newdata = xy),
                nrow = grid.lines, ncol = grid.lines)
# fitted points for droplines to surface
fitpoints ← predict(fit)
# scatter plot with regression plane
scatter3D(x, y, z, pch = 18, cex = 2,
    theta = 20, phi = 20, ticktype = "detailed",
    xlab = "wind", ylab = "rad", zlab = "ozone",
    surf = list(x = x.pred, y = y.pred, z = z.pred,
    facets = NA, fit = fitpoints), main = "Pollution 3: wind & rad")
```


Pollution 3: wind & rad
ozone
wind
rad