Chapter 11: More Model Selection

Mortality and Pollution

Does pollution kill people? This is data from any early study (1960s) of 60 metropolitan areas. The response variable of interest is mortality (mortality from all causes per 100,000 people). The explanatory variables include weather variables: precipitation (inches/year), humidity (%), January and July temperature (°F), demographic variables: percent of the population over 65, population per household, education (years), housing quality (% sound), population density (pop/mi$^2$), percent non-white, white collar work (%), percent at poverty level, and pollution potentials: hydrocarbons, nitrogen oxides, and sulphur dioxide.

There are two questions we’ll focus on in this problem. First, is there any relationship between mortality and air pollution after accounting for mortality differences

Before we begin our analysis, we’ll look at some plots. There’s too many to print, so we’ll just do this part in class and discuss...

Basic Data Summary

```r
> mort = read.csv('data/mortality.csv')

> basic.sums = rbind(mean = sapply(mort, mean),
+     SD = sapply(mort, sd), n = sapply(mort, length))
> signif(basic.sums, 3)

    CITY Mortality Precip Humid JanTemp JulyTemp Over65  House  Educ  Sound
mean      NA     940.0  37.40   57.70    34.0    74.60   8.80 11.000  80.90
SD        NA      62.2   9.98   5.37     10.2     4.76   1.46  0.135  0.845
n         60      60.00  60.00  60.00     60.0    60.00  60.00 60.000 60.000

    Density NonWhite WhiteCol  Poor   HC  NOX  SO2
mean    3870    11.90    46.10 14.40 37.9 22.6 53.8
SD      1450     8.92     4.61  4.16 92.0 46.3 63.4
n         60    60.00    60.00 60.00 60.0 60.0 60.0

Are the pollution variables good (significant) indicators of mortality? How would you summarize the significance of these variables?
> ml = lm(Mortality ~ log10(HC) + log10(NOX) + log10(SO2), data = mort)
> drop1(ml, test = 'F')

Single term deletions

Model:
Mortality ~ log10(HC) + log10(NOX) + log10(SO2)

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
<th>F value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>165446</td>
<td>483.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log10(HC)</td>
<td>1</td>
<td>25723.6</td>
<td>191170</td>
<td>489.99</td>
<td>8.7069</td>
</tr>
<tr>
<td>log10(NOX)</td>
<td>1</td>
<td>21247.3</td>
<td>186693</td>
<td>488.57</td>
<td>7.1917</td>
</tr>
<tr>
<td>log10(SO2)</td>
<td>1</td>
<td>7971.9</td>
<td>173418</td>
<td>484.15</td>
<td>2.6983</td>
</tr>
</tbody>
</table>

> summary(ml)

Residual standard error: 54.35 on 56 degrees of freedom
Multiple R-squared: 0.2752, Adjusted R-squared: 0.2364
F-statistic: 7.089 on 3 and 56 DF, p-value: 0.0004035

After adding the weather variables what can we say about the pollution variables?

Should we use tests based on Adj or Seq SS? Why or why not?

> m2 = update(ml, . ~ . + Precip + Humid + JanTemp + JulyTemp)
> drop1(m2, test = 'F')

Single term deletions

Model:
Mortality ~ log10(HC) + log10(NOX) + log10(SO2) + Precip + Humid + JanTemp + JulyTemp

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
<th>F value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>88378</td>
<td>453.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log10(HC)</td>
<td>1</td>
<td>356</td>
<td>88734</td>
<td>451.94</td>
<td>0.2095</td>
</tr>
<tr>
<td>log10(NOX)</td>
<td>1</td>
<td>8454</td>
<td>96832</td>
<td>457.18</td>
<td>4.9742</td>
</tr>
<tr>
<td>log10(SO2)</td>
<td>1</td>
<td>99</td>
<td>88477</td>
<td>451.77</td>
<td>0.0583</td>
</tr>
<tr>
<td>Precip</td>
<td>1</td>
<td>40503</td>
<td>128880</td>
<td>474.34</td>
<td>23.8312</td>
</tr>
<tr>
<td>Humid</td>
<td>1</td>
<td>597</td>
<td>88974</td>
<td>452.11</td>
<td>0.3510</td>
</tr>
<tr>
<td>JanTemp</td>
<td>1</td>
<td>6212</td>
<td>94589</td>
<td>455.78</td>
<td>3.6548</td>
</tr>
<tr>
<td>JulyTemp</td>
<td>1</td>
<td>8580</td>
<td>96958</td>
<td>457.26</td>
<td>5.0486</td>
</tr>
</tbody>
</table>
summary(m2)
Residual standard error: 41.23 on 52 degrees of freedom
Multiple R-squared: 0.6128, Adjusted R-squared: 0.5607
F-statistic: 11.76 on 7 and 52 DF, p-value: 7.332e-09

What has happened to the $R^2$ value when the HC and SO2 variables are dropped? How about the Adj-$R^2$?

Why aren’t these changing in the same way?

> m3 = update(m2, . ~ . - log10(HC) - log10(SO2))
> drop1(m3)

Single term deletions

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
<th>F value</th>
<th>Pr(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>88835</td>
<td>450.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log10(NOX)</td>
<td>1</td>
<td>78081</td>
<td>166915</td>
<td>485.85</td>
<td>47.4629</td>
<td>6.232e-09 ***</td>
</tr>
<tr>
<td>Precip</td>
<td>1</td>
<td>59267</td>
<td>148101</td>
<td>478.68</td>
<td>36.0264</td>
<td>1.692e-07 ***</td>
</tr>
<tr>
<td>Humid</td>
<td>1</td>
<td>427</td>
<td>89261</td>
<td>448.30</td>
<td>0.2594</td>
<td>0.612630</td>
</tr>
<tr>
<td>JanTemp</td>
<td>1</td>
<td>14475</td>
<td>103310</td>
<td>457.07</td>
<td>8.7990</td>
<td>0.004482 **</td>
</tr>
<tr>
<td>JulyTemp</td>
<td>1</td>
<td>10849</td>
<td>99684</td>
<td>454.92</td>
<td>6.5947</td>
<td>0.013028 *</td>
</tr>
</tbody>
</table>

> summary(m3)
Residual standard error: 40.56 on 54 degrees of freedom
Multiple R-squared: 0.6108, Adjusted R-squared: 0.5748
F-statistic: 16.95 on 5 and 54 DF, p-value: 4.785e-10

$R^2$-adjusted versus $R^2$

$$R^2 = \frac{SSY - SSE}{SSY} = \frac{SSF}{SSY}$$

$$R^2_{adj} = \frac{MSY - MSE}{MSE} = 1 - (1 - R^2)\left(\frac{df_Y}{df_E}\right)$$

Looking at the formula what happens if you add terms to a model but $R^2$ doesn’t change much?

How can we use $R^2_{adj}$ to help us find good models?
The effect of extra (non useful) variables on predictions.

Suppose we wanted to predict the mortality for a city with 40 inches of annual precipitation, 60% average humidity, 35°F January temp, 80°F July temp, 3.97 NOX, 12.18 SO2, and 12.18 HC. (Is this going to be an interpolation or extrapolation? √ this).

\[
\begin{align*}
\text{> new.dat} &= \text{data.frame(Precip = 40, Humid = 60, JanTemp = 35,} \\
&\quad \text{ JulyTemp = 80, NOX = 3.97, SO2 = 12.18, HC = 12.18)} \\
\text{> formula(m2)} &= \text{Mortality ~ log10(HC) + log10(NOX) + log10(SO2) + Precip + Humid +} \\
&\quad \text{JanTemp + JulyTemp} \\
\text{> formula(m3)} &= \text{Mortality ~ log10(NOX) + Precip + Humid + JanTemp + JulyTemp} \\
\text{> predict(m2, new.dat, interval = 'prediction')} &= \text{fit lwr upr} \\
&\quad 1 935.8411 845.8713 1025.811 \\
\text{> predict(m3, new.dat, interval = 'prediction')} &= \text{fit lwr upr} \\
&\quad 1 940.4278 856.1633 1024.692
\end{align*}
\]

Recall prediction interval formula for regression:

\[
\hat{y} \pm t_\alpha \frac{s}{\sqrt{n}} \sqrt{\frac{1}{m} + \frac{(x^*-\bar{x})^2}{(n-1)s_x^2}}
\]

Why is the residual SD larger for model 2 compared to model 3 when model 2 has more predictors than model 3?
Automated Model Selection
With so many variables at play, there are lots of p-values to consider. And, every time you change the model (add or remove a variable) there is another whole set of p-values. The number of statistical tests can grow very large. In our example there are 15 predictor variables. There are 32,767 possible models for mortality that can be created with these variables, not counting those that include transformations, interactions, or polynomials!

Use your head
• You might have a lot of data, but you don’t need to include all of it.
  o Don’t include highly correlated predictors
  o Unless you have a particular hypothesis, don’t include nonsense variables
• Look at previous studies.
  o If earlier studies deemed certain variables to be significant, you might consider including it in your model regardless of its p-value. (I.e., no test required).
• Use variables that will continue to be useful in the future.

Bonferroni correction
Decrease $\alpha$ so that your overall risk of making a Type I error does not get too high.

• For example, if you use going to conduct $m$ tests at level $\alpha$, your overall chance of making at least one Type I error is higher than $\alpha$ and at most $m \alpha$. So a conservative thing to do is to carry out each test at a lower $\alpha$ level, say $\alpha/m$, the so-called Bonferroni adjusted $\alpha$.
• From the text, assuming independent tests... if you use $\alpha = 0.01$ and are going to conduct $m$ tests, your overall chance of making at least one Type I error is approximately $1 - (1 - \alpha)^m = 1 - (.99)^m$.

Collect terms to tests many variables with one test
Combine SS

• Put the terms you want to test at the end of your model and add their SeqSS and their df to obtain a single MS (MSG) for the group of variables. The compute $F = \frac{MSG}{MSE}$ and the corresponding p-value.
> anova(lm(Mortality ~ Precip + Humid + JanTemp + JulyTemp +
+     log10(NOX) + log10(SO2) + log10(HC), data = mort))

Analysis of Variance Table

Response: Mortality

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precip</td>
<td>1</td>
<td>59256</td>
<td>59256</td>
<td>34.8652</td>
<td>2.722e-07 ***</td>
</tr>
<tr>
<td>Humid</td>
<td>1</td>
<td>554</td>
<td>554</td>
<td>0.3260</td>
<td>0.5705</td>
</tr>
<tr>
<td>JanTemp</td>
<td>1</td>
<td>1246</td>
<td>1246</td>
<td>0.7330</td>
<td>0.3958</td>
</tr>
<tr>
<td>JulyTemp</td>
<td>1</td>
<td>302</td>
<td>302</td>
<td>0.1779</td>
<td>0.6749</td>
</tr>
<tr>
<td>log10(NOX)</td>
<td>1</td>
<td>78081</td>
<td>78081</td>
<td>45.9414</td>
<td>1.118e-08 ***</td>
</tr>
<tr>
<td>log10(SO2)</td>
<td>1</td>
<td>101</td>
<td>101</td>
<td>0.0595</td>
<td>0.8083</td>
</tr>
<tr>
<td>log10(HC)</td>
<td>1</td>
<td>356</td>
<td>356</td>
<td>0.2095</td>
<td>0.6491</td>
</tr>
<tr>
<td>Residuals</td>
<td>52</td>
<td>88378</td>
<td>1700</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

> anova(m3,m2)

Analysis of Variance Table

Model 1: Mortality ~ log10(NOX) + Precip + Humid + JanTemp + JulyTemp
Model 2: Mortality ~ log10(HC) + log10(NOX) + log10(SO2) + Precip + Humid +
JanTemp + JulyTemp

<table>
<thead>
<tr>
<th></th>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54</td>
<td>88835</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>52</td>
<td>88378</td>
<td>2</td>
<td>457.12</td>
<td>0.1345</td>
<td>0.8745</td>
</tr>
</tbody>
</table>

**Stepwise procedures**

**Backwards Selection**

1. Build the full model that includes all of your predictor variables.
2. Remove the variable with the largest p-value above some threshold, say \( \alpha = 0.10 \) and refit the model with the reduced set of variables. OR, use some other criteria to decide what variables offer the least/most utility to the model (Akaike Information Criterion (AIC) - models with lower AIC values are better)
3. Repeat step 2 until no variables meet the removal criterion.

> formula(m2)
Mortality ~ log10(HC) + log10(NOX) + log10(SO2) + Precip + Humid +
JanTemp + JulyTemp

> m2.backwards = step(m2, direction = 'backward')
Start: AIC=453.7
Mortality ~ log10(HC) + log10(NOX) + log10(SO2) + Precip + Humid +
JanTemp + JulyTemp

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>- log10(SO2)</td>
<td>1</td>
<td>99</td>
<td>88477</td>
<td>451.77</td>
</tr>
<tr>
<td>- log10(HC)</td>
<td>1</td>
<td>356</td>
<td>88734</td>
<td>451.94</td>
</tr>
<tr>
<td>- Humid</td>
<td>1</td>
<td>597</td>
<td>88974</td>
<td>452.11</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td></td>
<td>88378</td>
<td>453.70</td>
</tr>
<tr>
<td>- JanTemp</td>
<td>1</td>
<td>6212</td>
<td>94589</td>
<td>455.78</td>
</tr>
<tr>
<td>- log10(NOX)</td>
<td>1</td>
<td>8454</td>
<td>96832</td>
<td>457.18</td>
</tr>
<tr>
<td>- JulyTemp</td>
<td>1</td>
<td>8580</td>
<td>96958</td>
<td>457.26</td>
</tr>
</tbody>
</table>
- Precip       1     40503 128880 474.34

Step:  AIC=451.77
Mortality ~ log10(HC) + log10(NOX) + Precip + Humid + JanTemp + JulyTemp

Df Sum of Sq    RSS    AIC
- log10(HC)     1       358  88835 450.01
- Humid         1       504  88980 450.11
  <none>                     88477 451.77
- JulyTemp      1       8516  96993 455.28
- JanTemp       1      11290 100939 456.97
- log10(NOX)    1      12462 100939 457.68
- Precip        1      53418 141895 478.11

Step:  AIC=450.01
Mortality ~ log10(NOX) + Precip + Humid + JanTemp + JulyTemp

Df Sum of Sq    RSS    AIC
- Humid         1       427  89261 448.30
  <none>                     88835 450.01
- JulyTemp      1     10849  99684 454.92
- JanTemp       1     14475 103310 457.07
- Precip        1     59267 148101 478.68
- log10(NOX)    1     78081 166915 485.85

Step:  AIC=448.3
Mortality ~ log10(NOX) + Precip + JanTemp + JulyTemp

Df Sum of Sq    RSS    AIC
  <none>                    89261 448.30
- JulyTemp     1     12267 101529 454.03
- JanTemp      1     14360 103622 455.25
- Precip       1     64166 153427 478.80
- log10(NOX)   1     77719 166980 483.88

Call:
  lm(formula = Mortality ~ log10(NOX) + Precip + JanTemp + JulyTemp,     data = mort)

Coefficients:
  (Intercept)       log10(NOX)       Precip      JanTemp      JulyTemp
  476.764         81.086         3.930       -1.712         3.930

> drop1(m2.backwards, test = 'F')
Single term deletions

Model:
Mortality ~ log10(NOX) + Precip + JanTemp + JulyTemp
  Df Sum of Sq    RSS    AIC F value     P   r(F)
  <none>                    89261 448.30
log10(NOX)    1     77719 166980 483.88 47.8878 5.099 e-09 ***
Precip        1     64166 153427 478.80 39.5369 5.495 e-08 ***
JanTemp       1     14360 103622 455.25  8.8484   0.00 4350 **
JulyTemp      1     12267 101529 454.03  7.5587   0.008065 **
Forwards Selection

1. Build all single variable models.
2. Choose the model with the lowest p-value below some threshold, say $\alpha = 0.10$.
3. Build all two variable models that include the variable just chosen in step 2.
4. Choose the model whose second added variable has the lowest p-value below some threshold, say $\alpha = 0.10$.
5. Repeat 3 and 4 for three, four, five...variable models stopping when no additional variable produces a significant term.

```r
> m2.forward = step(lm(Mortality ~ 1, data = mort),
+     scope = ~ log10(HC) + log10(NOX) + log10(SO2) + Precip + Humid +
+     JanTemp + JulyTemp, direction = 'forward')
Start:  AIC=496.64
Mortality ~ 1

Df Sum of Sq    RSS    AIC
+ Precip      1     59256 169017 480.60
+ log10(SO2)  1     37098 191176 488.00
+ log10(NOX)  1     19463 208810 493.29
+ JulyTemp    1     17515 210758 493.85
<none>                    228273 496.64
+ log10(HC)   1      5188 223085 497.26
+ Humid       1      1789 226484 498.17
+ JanTemp     1       206 228067 498.58

Step:  AIC=480.6
Mortality ~ Precip

Df Sum of Sq    RSS    AIC
+ log10(NOX)  1     60759 108259 455.88
+ log10(SO2)  1     50065 118952 461.53
+ log10(HC)   1     42317 126700 465.31
<none>                    108259 455.88
+ JanTemp     1     1367 167650 482.12
+ Humid       1     554 168463 482.41
+ JulyTemp    1     130 168888 482.56

Step:  AIC=455.88
Mortality ~ Precip + log10(NOX)

Df Sum of Sq    RSS    AIC
+ JanTemp     1     6730.2 101529 454.03
+ log10(HC)   1     6262.2 101997 454.30
+ JulyTemp    1     4637.1 103622 455.25
+ log10(SO2)  1     3593.7 104665 455.85
<none>                    108259 455.88
+ Humid       1     2280.9 105978 456.60

Step:  AIC=454.03
Mortality ~ Precip + log10(NOX) + JanTemp

Df Sum of Sq    RSS    AIC
+ JulyTemp    1    12267.3  89261 448.30
+ log10(HC)   1     3919.2  97609 453.66
```
<none> 101529 454.03
+ Humid 1 1845.1 99684 454.92
+ log10(SO2) 1 8.8 101520 456.02

Step: AIC=448.3
Mortality ~ Precip + log10(NOX) + JanTemp + JulyTemp

Df Sum of Sq RSS AIC
<none> 89261 448.30
+ Humid 1 426.68 88835 450.01
+ log10(HC) 1 281.22 88980 450.11
+ log10(SO2) 1 9.46 89252 450.29

Call:
lm(formula = Mortality ~ Precip + log10(NOX) + JanTemp + JulyTemp, data = mort)

Coefficients:
(Intercept) Precip log10(NOX) JanTemp JulyTemp
476.764 3.930 81.086 -1.712 3.930

> drop1(m2.forward, test = 'F')
Single term deletions

Model:
Mortality ~ Precip + log10(NOX) + JanTemp + JulyTemp

Df Sum of Sq RSS AIC F value Pr(F)
<none> 89261 448.30
Precip 1 64166 153427 478.80 39.5369 5.495e-08 ***
log10(NOX) 1 77719 166980 483.88 47.8878 5.099e-09 ***
JanTemp 1 14360 103622 455.25 8.8484 0.004350 **
JulyTemp 1 12267 101529 454.03 7.5587 0.008065 **

Forwards/Backwards Selection
1. Build all single variable models.
2. Choose the model with the lowest p-value below some threshold.
3. Build all two variable models that include the variable chosen in step 2.
4. Choose the model whose second variable has the lowest p-value below some threshold.
5. If the first variable is no longer significant in this new model, drop that variable.
6. Repeat steps 3, 4, and 5 adding variables one at a time stopping when no more variables can be added or removed.

> m2.both.forward = step(m2, direction = 'both')
Start: AIC=453.7
Mortality ~ log10(HC) + log10(NOX) + log10(SO2) + Precip + Humid + JanTemp + JulyTemp

Df Sum of Sq RSS AIC
- log10(SO2) 1 99 88477 451.77
- log10(HC) 1 356 88734 451.94
- Humid 1 597 88974 452.11
<none> 88378 453.70
- JanTemp 1 6212 94589 455.78
- log10(NOX) 1 8454 96832 457.18
Step:  \textbf{AIC=451.77}
Mortality \sim \log_{10}(HC) + \log_{10}(NOX) + \text{Precip} + \text{Humid} + \text{JanTemp} + \text{JulyTemp}

\begin{tabular}{lrrrr}
 \textbf{Df} & \textbf{Sum of Sq} & \textbf{RSS} & \textbf{AIC} \\
- \log_{10}(HC) & 1 & 358 & 88835 & 450.01 \\
- Humid & 1 & 504 & 88980 & 450.11 \\
<none> & & & 88477 & 451.77 \\
+ \log_{10}(SO2) & 1 & 99 & 88378 & 453.70 \\
- JulyTemp & 1 & 8516 & 96993 & 455.28 \\
- JanTemp & 1 & 11290 & 99767 & 456.97 \\
- \log_{10}(NOX) & 1 & 12462 & 100939 & 457.68 \\
- Precip & 1 & 53418 & 141895 & 478.11 \\
\end{tabular}

Step:  \textbf{AIC=450.01}
Mortality \sim \log_{10}(NOX) + \text{Precip} + \text{Humid} + \text{JanTemp} + \text{JulyTemp}

\begin{tabular}{lrrrr}
 \textbf{Df} & \textbf{Sum of Sq} & \textbf{RSS} & \textbf{AIC} \\
- Humid & 1 & 427 & 89261 & 448.30 \\
<none> & & & 88835 & 450.01 \\
+ \log_{10}(HC) & 1 & 358 & 88477 & 451.77 \\
+ \log_{10}(SO2) & 1 & 101 & 88734 & 451.94 \\
- JulyTemp & 1 & 10849 & 99684 & 454.92 \\
- JanTemp & 1 & 14475 & 103310 & 457.07 \\
- Precip & 1 & 59267 & 148101 & 478.68 \\
- \log_{10}(NOX) & 1 & 78081 & 166915 & 485.85 \\
\end{tabular}

Step:  \textbf{AIC=448.3}
Mortality \sim \log_{10}(NOX) + \text{Precip} + \text{JanTemp} + \text{JulyTemp}

\begin{tabular}{lrrrr}
 \textbf{Df} & \textbf{Sum of Sq} & \textbf{RSS} & \textbf{AIC} \\
<none> & & & 89261 & 448.30 \\
+ Humid & 1 & 427 & 88835 & 450.01 \\
+ \log_{10}(HC) & 1 & 281 & 88980 & 450.11 \\
+ \log_{10}(SO2) & 1 & 9 & 89252 & 450.29 \\
- JulyTemp & 1 & 12267 & 101529 & 454.03 \\
- JanTemp & 1 & 14360 & 103622 & 455.25 \\
- Precip & 1 & 64166 & 153427 & 478.80 \\
- \log_{10}(NOX) & 1 & 77719 & 166980 & 483.88 \\
\end{tabular}

\begin{verbatim}
> m2.both.backward = step(lm(Mortality ~ 1, data = mort), 
+ scope = ~ \log_{10}(HC) + \log_{10}(NOX) + \log_{10}(SO2) + \text{Precip} + \text{Humid} + 
+ \text{JanTemp} + \text{JulyTemp}, direction = 'both')
Start:  \textbf{AIC=496.64}
Mortality \sim 1

\begin{tabular}{lrrrr}
 \textbf{Df} & \textbf{Sum of Sq} & \textbf{RSS} & \textbf{AIC} \\
+ \text{Precip} & 1 & 59256 & 169017 & 480.60 \\
+ \log_{10}(SO2) & 1 & 37098 & 191176 & 488.00 \\
+ \log_{10}(NOX) & 1 & 19463 & 208810 & 493.29 \\
+ JulyTemp & 1 & 17515 & 210758 & 493.85 \\
<none> & & & 228273 & 496.64 \\
+ \log_{10}(HC) & 1 & 5188 & 223085 & 497.26 \\
+ Humid & 1 & 1789 & 226484 & 498.17 \\
\end{tabular}
\end{verbatim}
+ JanTemp     1       206 228067 498.58  

Step:  **AIC=480.6**  
Mortality ~ Precip

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ log10(NOX)</td>
<td>1</td>
<td>60759</td>
<td>108259</td>
</tr>
<tr>
<td>+ log10(SO2)</td>
<td>1</td>
<td>50065</td>
<td>118952</td>
</tr>
<tr>
<td>+ log10(HC)</td>
<td>1</td>
<td>42317</td>
<td>126700</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>169017</td>
<td>480.60</td>
</tr>
<tr>
<td>+ JanTemp</td>
<td>1</td>
<td>1367</td>
<td>167650</td>
</tr>
<tr>
<td>+ Humid</td>
<td>1</td>
<td>554</td>
<td>168463</td>
</tr>
<tr>
<td>+ JulyTemp</td>
<td>1</td>
<td>130</td>
<td>168888</td>
</tr>
<tr>
<td>- Precip</td>
<td>1</td>
<td>59256</td>
<td>228273</td>
</tr>
</tbody>
</table>

Step:  **AIC=455.88**  
Mortality ~ Precip + log10(NOX)

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ JanTemp</td>
<td>1</td>
<td>6730</td>
<td>101529</td>
</tr>
<tr>
<td>+ log10(HC)</td>
<td>1</td>
<td>6262</td>
<td>101997</td>
</tr>
<tr>
<td>+ JulyTemp</td>
<td>1</td>
<td>4637</td>
<td>103622</td>
</tr>
<tr>
<td>+ log10(SO2)</td>
<td>1</td>
<td>3594</td>
<td>104665</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>108259</td>
<td>455.88</td>
</tr>
<tr>
<td>+ Humid</td>
<td>1</td>
<td>2281</td>
<td>105978</td>
</tr>
<tr>
<td>- log10(NOX)</td>
<td>1</td>
<td>60759</td>
<td>169017</td>
</tr>
<tr>
<td>- Precip</td>
<td>1</td>
<td>100551</td>
<td>208810</td>
</tr>
</tbody>
</table>

Step:  **AIC=454.03**  
Mortality ~ Precip + log10(NOX) + JanTemp

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ JulyTemp</td>
<td>1</td>
<td>12267</td>
<td>89261</td>
</tr>
<tr>
<td>+ log10(HC)</td>
<td>1</td>
<td>3919</td>
<td>97609</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>101529</td>
<td>454.03</td>
</tr>
<tr>
<td>+ Humid</td>
<td>1</td>
<td>1845</td>
<td>99684</td>
</tr>
<tr>
<td>- JanTemp</td>
<td>1</td>
<td>6730</td>
<td>108259</td>
</tr>
<tr>
<td>+ log10(SO2)</td>
<td>1</td>
<td>9</td>
<td>101520</td>
</tr>
<tr>
<td>- log10(NOX)</td>
<td>1</td>
<td>66122</td>
<td>167650</td>
</tr>
<tr>
<td>- Precip</td>
<td>1</td>
<td>106202</td>
<td>207731</td>
</tr>
</tbody>
</table>

Step:  **AIC=448.3**  
Mortality ~ Precip + log10(NOX) + JanTemp + JulyTemp

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>89261</td>
<td>448.30</td>
</tr>
<tr>
<td>+ Humid</td>
<td>1</td>
<td>427</td>
<td>88835</td>
</tr>
<tr>
<td>+ log10(HC)</td>
<td>1</td>
<td>801</td>
<td>88980</td>
</tr>
<tr>
<td>+ log10(SO2)</td>
<td>1</td>
<td>9</td>
<td>89252</td>
</tr>
<tr>
<td>- JulyTemp</td>
<td>1</td>
<td>12267</td>
<td>101529</td>
</tr>
<tr>
<td>- JanTemp</td>
<td>1</td>
<td>14360</td>
<td>103622</td>
</tr>
<tr>
<td>- Precip</td>
<td>1</td>
<td>64166</td>
<td>153427</td>
</tr>
<tr>
<td>- log10(NOX)</td>
<td>1</td>
<td>77719</td>
<td>166980</td>
</tr>
</tbody>
</table>
Drawbacks to Stepwise procedures
  • Some implementations (minitab) can ignore the distinction between numeric and categorical variables and hierarchy (R is okay).
  • Ignores any insights of the researcher

Best Subsets Regression
  • Computer intensive method of considering all possible models.