NCSU ST 503 HW 9

Probems 10.1 (a - c), 10.4, and 10.5 Faraway, Julian J. Linear Models with R, Second Edition Chapman & Hall / CRC Press.

Bruce Campbell

27 October, 2017

10.1 1 (a - c) Subset Selection with prostate data

For 10.1 (a): Please use Backward Elimination in 3 ways: (i) a 0.05 p-value criterion as the stopping rule, (ii) using AIC as the stopping rule, and (iii) using BIC as the stopping rule.

For 10.1 (b-c): You should be comparing all possible subsets.

Use the prostate data with lpsa as the response and the other variables as predictors. Implement the following variable selection methods to determine the "best" model:

(a) Backward elimination

It was not clear to be that it is possible to use regsubsets with the backward method to perform Backward Elimination based on p-value.

```
rm(list = ls())
data(prostate, package="faraway");
df <- prostate
n <-nrow(df)

lm.fit <- lm(lpsa ~ ., data=prostate)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```
lm.subset1 <- update(lm.fit,. ~ . - gleason)
summary(lm.subset1)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     pgg45, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73117 -0.38137 -0.01728  0.43364  1.63513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.953926   0.829439   1.150  0.25319
## lcavol       0.591615   0.086001   6.879 8.07e-10 ***
## lweight      0.448292   0.167771   2.672  0.00897 **
## age         -0.019336   0.011066  -1.747  0.08402 .
## lbph         0.107671   0.058108   1.853  0.06720 .
## svi          0.757734   0.241282   3.140  0.00229 **
## lcp         -0.104482   0.090478  -1.155  0.25127
## pgg45        0.005318   0.003433   1.549  0.12488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 89 degrees of freedom
## Multiple R-squared:  0.6544, Adjusted R-squared:  0.6273
## F-statistic: 24.08 on 7 and 89 DF,  p-value: < 2.2e-16
```

```r
lm.subset1 <- update(lm.subset1,. ~ . - lcp)
summary(lm.subset1)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + pgg45,
##     data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77711 -0.41708  0.00002  0.40676  1.59681
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.980085   0.830665   1.180  0.24116
## lcavol       0.545770   0.076431   7.141 2.31e-10 ***
## lweight      0.449450   0.168078   2.674  0.00890 **
## age         -0.017470   0.010967  -1.593  0.11469
## lbph         0.105755   0.058191   1.817  0.07249 .
## svi          0.641666   0.219757   2.920  0.00442 **
## pgg45        0.003528   0.003068   1.150  0.25331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7061 on 90 degrees of freedom
## Multiple R-squared:  0.6493, Adjusted R-squared:  0.6259
## F-statistic: 27.77 on 6 and 90 DF,  p-value: < 2.2e-16
```

```r
lm.subset1 <- update(lm.subset1,. ~ . - pgg45)
summary(lm.subset1)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
```

```
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
lm.subset1 <- update(lm.subset1,. ~ . - age)
summary(lm.subset1)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82653 -0.42270  0.04362  0.47041  1.48530
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14554    0.59747   0.244  0.80809
## lcavol       0.54960    0.07406   7.422 5.64e-11 ***
## lweight      0.39088    0.16600   2.355  0.02067 *
## lbph         0.09009    0.05617   1.604  0.11213
## svi          0.71174    0.20996   3.390  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7108 on 92 degrees of freedom
## Multiple R-squared:  0.6366, Adjusted R-squared:  0.6208
## F-statistic: 40.29 on 4 and 92 DF,  p-value: < 2.2e-16
```

```
lm.subset1 <- update(lm.subset1,. ~ . - lbph)
summary(lm.subset1)
```

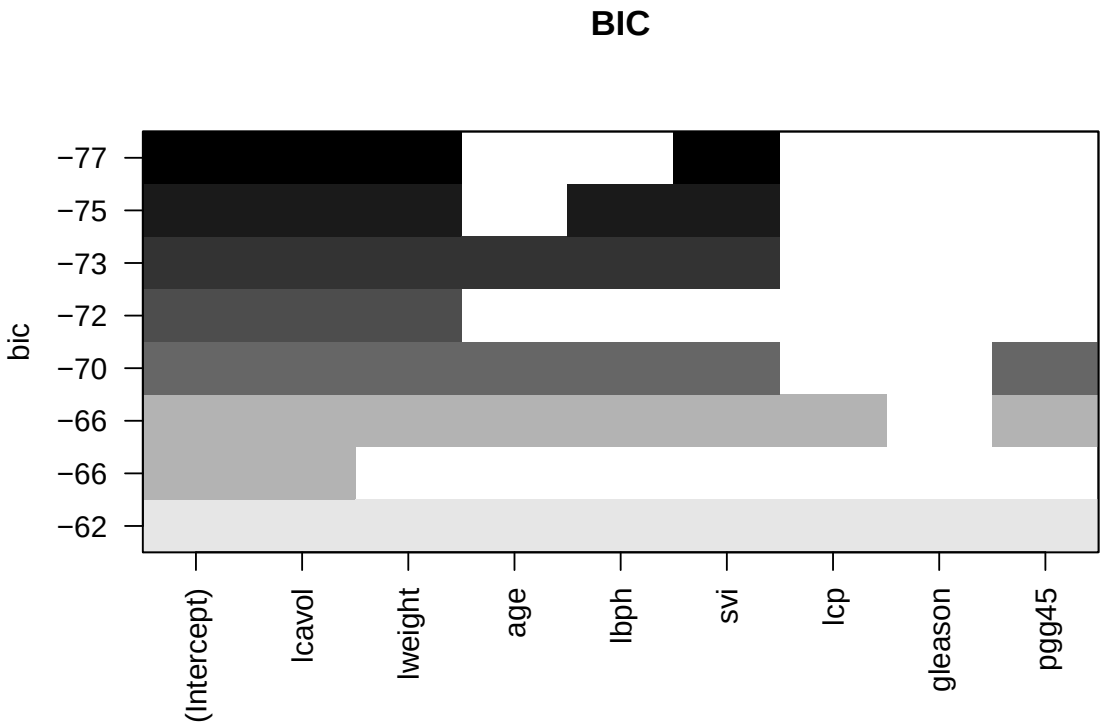
```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi          0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

Backward BIC

We can use regsubsets with the backwards method to find the best model by the BIC criteria. The plot method will show us the top models. Interestingly there does not appear a way to use the plot with the AIC.

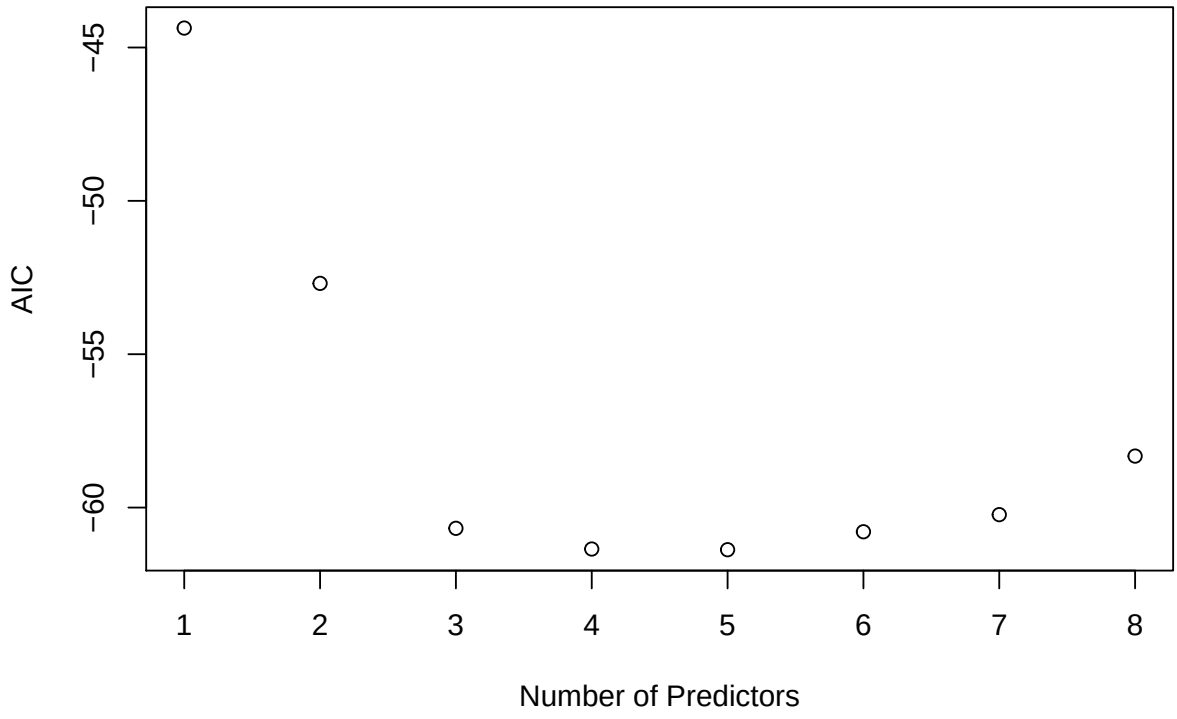
```
##   (Intercept) lcavol lweight   age  lbph   svi   lcp gleason pgg45
## 1        TRUE   TRUE   FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## 2        TRUE   TRUE    TRUE FALSE FALSE FALSE FALSE   FALSE FALSE
## 3        TRUE   TRUE    TRUE FALSE FALSE  TRUE FALSE   FALSE FALSE
## 4        TRUE   TRUE    TRUE FALSE  TRUE  TRUE FALSE   FALSE FALSE
## 5        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE FALSE
## 6        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE  TRUE
## 7        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE   FALSE  TRUE
## 8        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE    TRUE  TRUE
```



There does not appear to be a scale="aci" option for the regsubsets plot. This is interesting to note.
We plot the AIC for the models here and compare to the exhaustive search for reference.

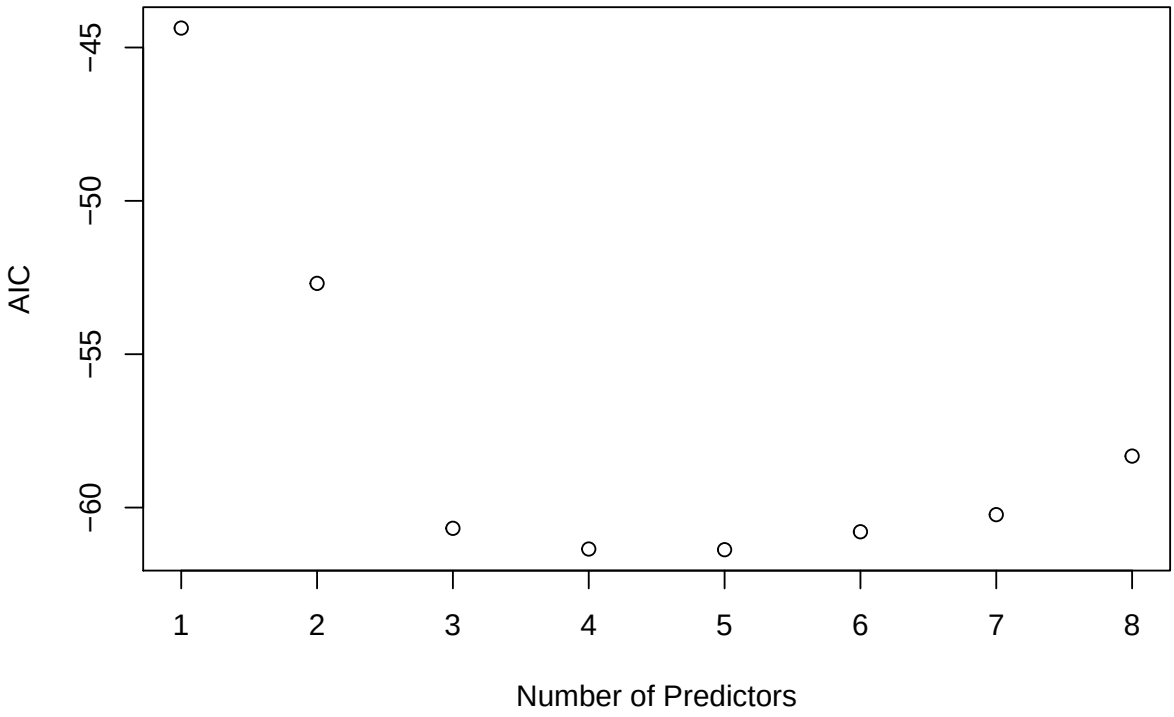
Backward AIC

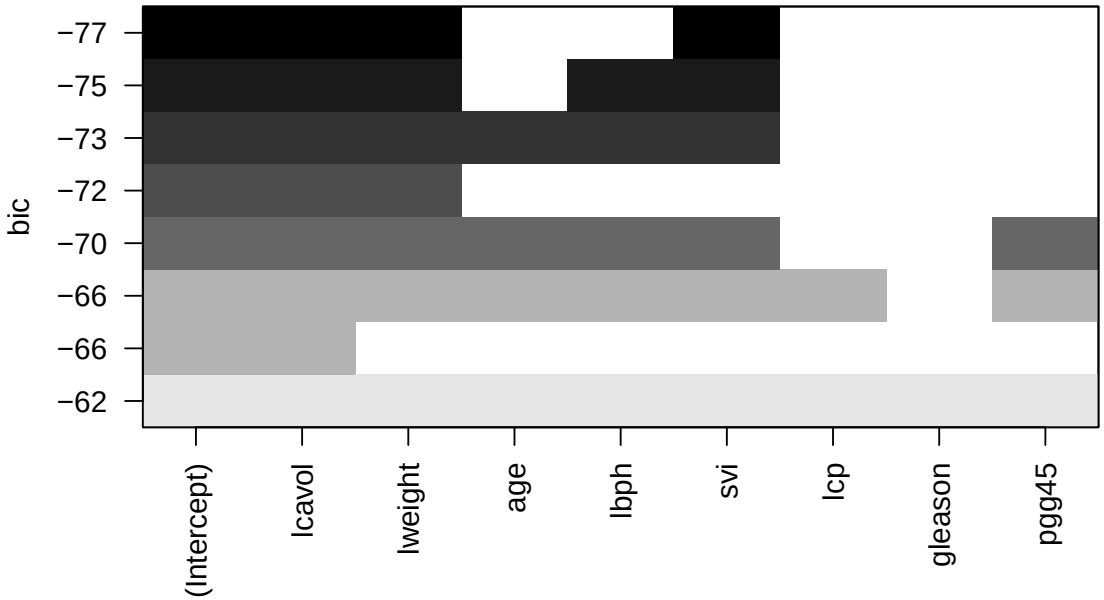
```
##   (Intercept) lcavol lweight   age  lbph   svi   lcp gleason pgg45
## 1        TRUE   TRUE   FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## 2        TRUE   TRUE    TRUE FALSE FALSE FALSE FALSE   FALSE FALSE
## 3        TRUE   TRUE    TRUE FALSE FALSE  TRUE FALSE   FALSE FALSE
## 4        TRUE   TRUE    TRUE FALSE  TRUE  TRUE FALSE   FALSE FALSE
## 5        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE FALSE
## 6        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE  TRUE
## 7        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE   FALSE  TRUE
## 8        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE    TRUE  TRUE
```



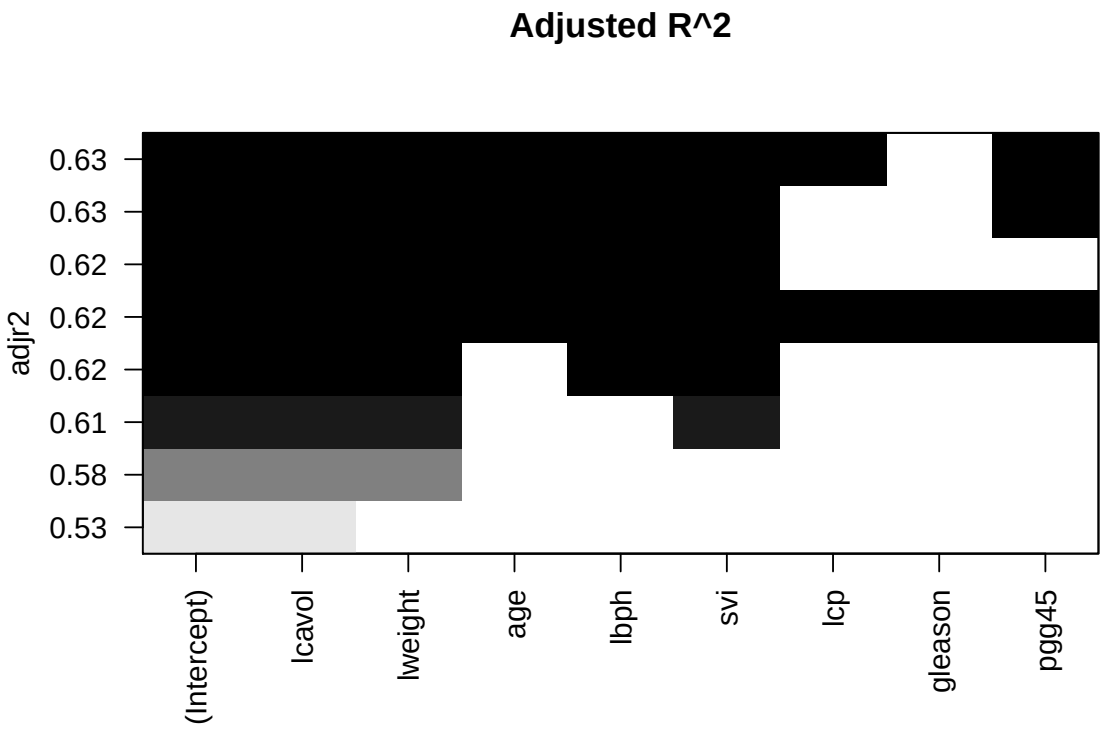
(b) exhaustive AIC

```
##   (Intercept) lcavol lweight   age  lbph   svi   lcp gleason pgg45
## 1        TRUE   TRUE   FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## 2        TRUE   TRUE    TRUE FALSE FALSE FALSE FALSE   FALSE FALSE
## 3        TRUE   TRUE    TRUE FALSE FALSE  TRUE FALSE   FALSE FALSE
## 4        TRUE   TRUE    TRUE FALSE  TRUE  TRUE FALSE   FALSE FALSE
## 5        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE FALSE
## 6        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE  TRUE
## 7        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE   FALSE  TRUE
## 8        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE    TRUE  TRUE
```


AIC
−45
−50
−55
−60
1
2
3
4
5
6
7
8
Number of Predictors


bic
−77
−75
−73
−72
−70
−66
−66
−62
(Intercept)
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45

(c) exhaustive Adjusted R^2



10.4 Simplifying trees model

Using the trees data, fit a model with log(Volume) as the response and a second-order polynomial (including the interaction term) in Girth and Height. Determine whether the model may be reasonably simplified.

```
##
## Call:
## lm(formula = log(Volume) ~ polym(Girth, Height, degree = 2),
##     data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.159718 -0.041905 -0.003371  0.055167  0.133780
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.27472    0.02370 138.163  < 2e-16
```

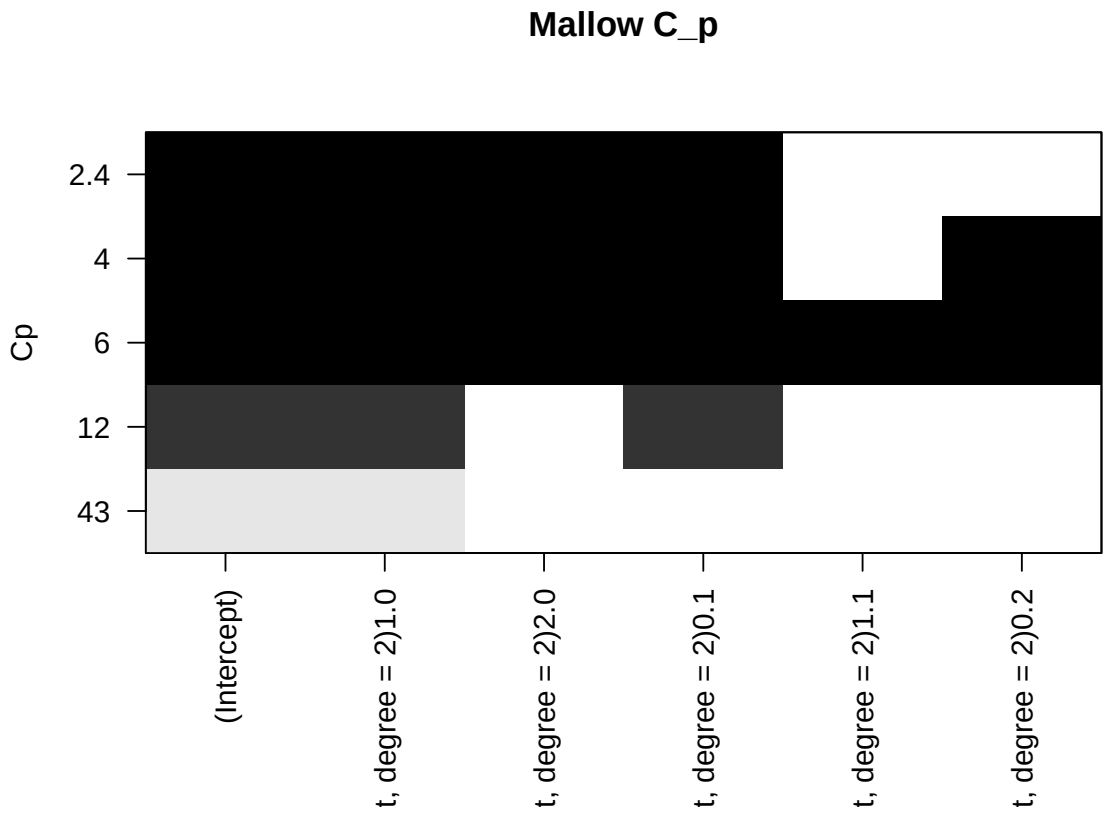
```
## polym(Girth, Height, degree = 2)1.0  2.51882    0.11972  21.039  < 2e-16
## polym(Girth, Height, degree = 2)2.0 -0.24312    0.18449  -1.318    0.200
## polym(Girth, Height, degree = 2)0.1  0.54249    0.11339   4.784 6.52e-05
## polym(Girth, Height, degree = 2)1.1 -0.11845    1.08511  -0.109    0.914
## polym(Girth, Height, degree = 2)0.2 -0.05025    0.10402  -0.483    0.633
##
## (Intercept)                          ***
## polym(Girth, Height, degree = 2)1.0 ***
## polym(Girth, Height, degree = 2)2.0
## polym(Girth, Height, degree = 2)0.1 ***
## polym(Girth, Height, degree = 2)1.1
## polym(Girth, Height, degree = 2)0.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08469 on 25 degrees of freedom
## Multiple R-squared:  0.9784, Adjusted R-squared:  0.9741
## F-statistic: 226.7 on 5 and 25 DF,  p-value: < 2.2e-16
```

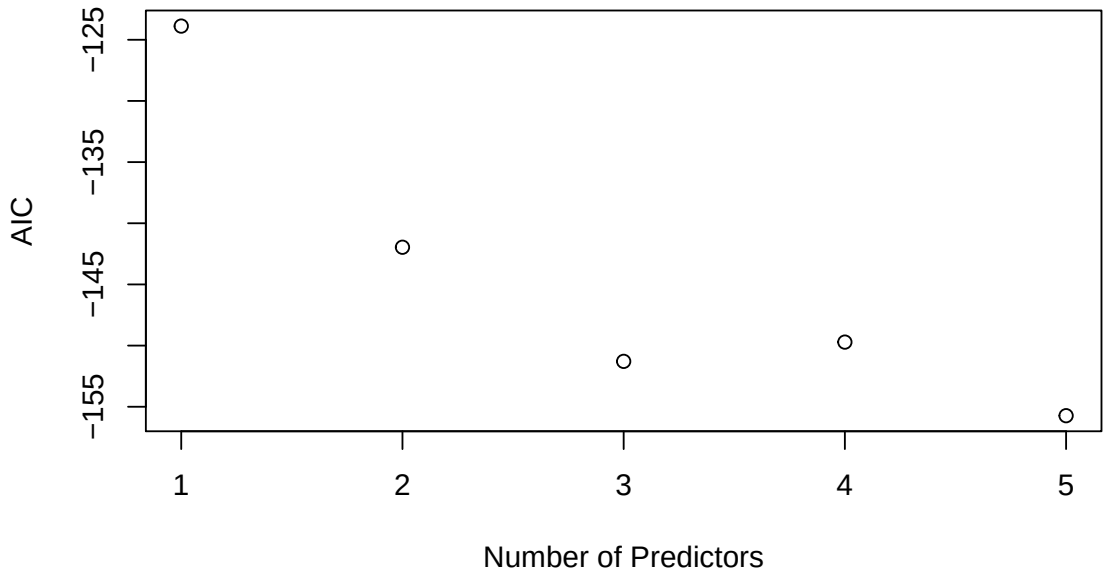
Now we run subset selection. We'll use an exhaustive method since there are not too many predictors. And we'll use the Mallow C_p as our criteria.

```
##                          (Intercept) polym(Girth, Height, degree = 2)1.0
##                           3.27471581                          2.51881720
## polym(Girth, Height, degree = 2)2.0 polym(Girth, Height, degree = 2)0.1
##                          -0.24312237                          0.54248964
## polym(Girth, Height, degree = 2)1.1 polym(Girth, Height, degree = 2)0.2
##                          -0.11844598                         -0.05024754
```

```
##   (Intercept) polym(Girth, Height, degree = 2)1.0
## 1        TRUE                                TRUE
## 2        TRUE                                TRUE
## 3        TRUE                                TRUE
## 4        TRUE                                TRUE
## 5        TRUE                                TRUE
##   polym(Girth, Height, degree = 2)2.0 polym(Girth, Height, degree = 2)0.1
## 1                               FALSE                               FALSE
## 2                               FALSE                                TRUE
## 3                                TRUE                                TRUE
## 4                                TRUE                                TRUE
## 5                                TRUE                                TRUE
##   polym(Girth, Height, degree = 2)1.1 polym(Girth, Height, degree = 2)0.2
## 1                               FALSE                               FALSE
## 2                               FALSE                               FALSE
## 3                               FALSE                               FALSE
## 4                               FALSE                                TRUE
```

```
## 5                    TRUE                    TRUE
```


Mallow C_p
Cp
2.4
4
6
12
43
(Intercept)
t, degree = 2)1.0
t, degree = 2)2.0
t, degree = 2)0.1
t, degree = 2)1.1
t, degree = 2)0.2



The AIC criterion indicates the best model is the full model with all the polynomial terms, while the Mallow Cp indicates the best model is a reduced one with the first three terms of the polynomial expansion :

$log(Volume) \sim polym(Girth, Height, degree = 2)1.0 + polym(Girth, Height, degree = 2)2.0 + polym(Girth, Height, degree = 2)0.1$

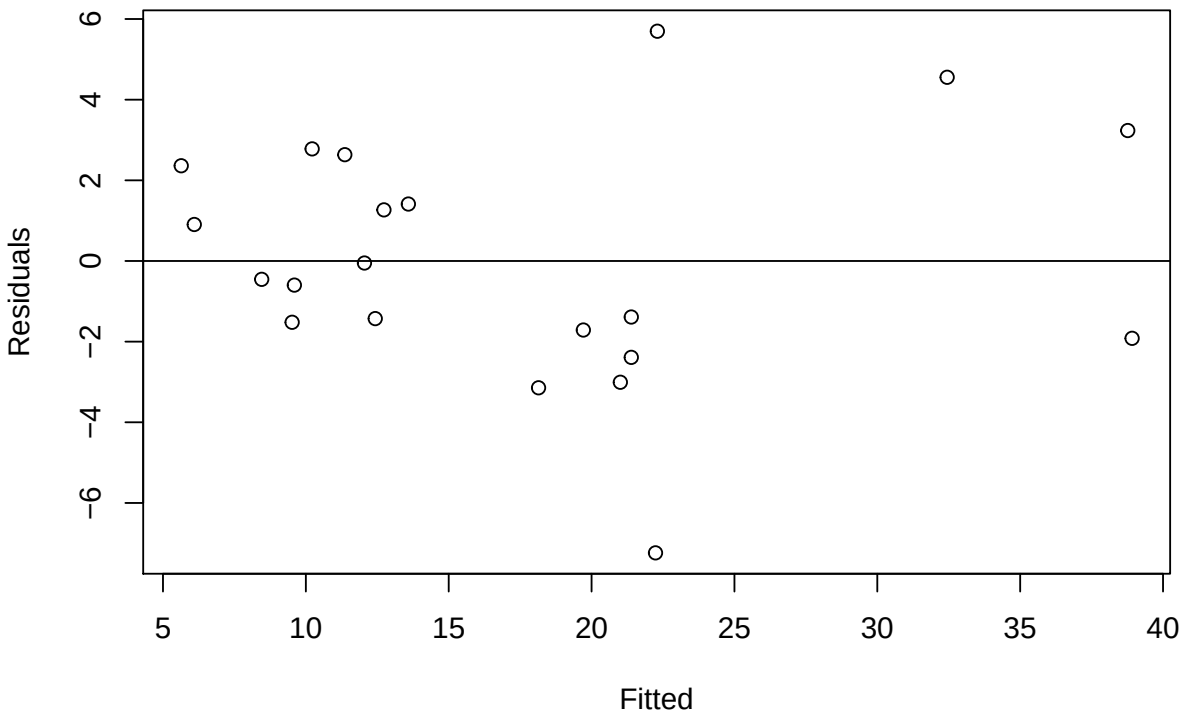
$log(Volume) \sim Girth + Girth^2, + Height$

10.5 Model reduction in stackloss data

Fit a linear model to the stackloss data with stack.loss as the predictor and the other variables as predictors.

```
##
## Call:
## lm(formula = stack.loss ~ ., data = stackloss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2377 -1.7117 -0.4551  2.3614  5.6978
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.9197    11.8960  -3.356  0.00375 **
## Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
## Water.Temp    1.2953     0.3680   3.520  0.00263 **
## Acid.Conc.   -0.1521     0.1563  -0.973  0.34405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.243 on 17 degrees of freedom
## Multiple R-squared:  0.9136, Adjusted R-squared:  0.8983
## F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

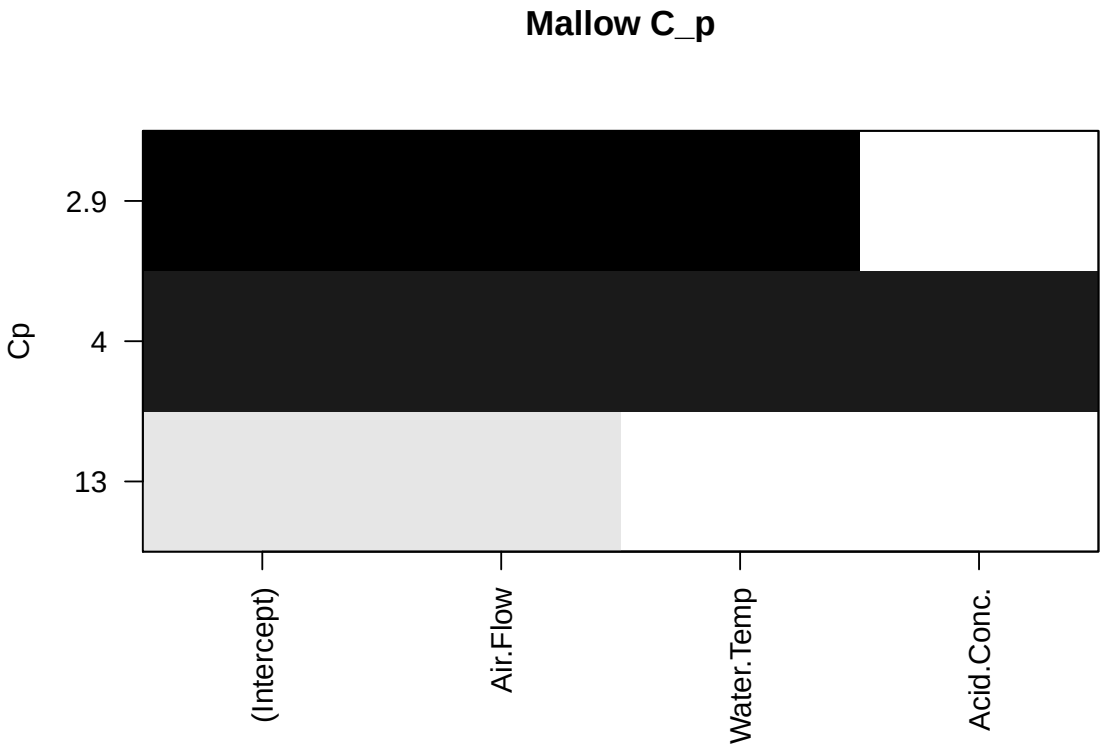


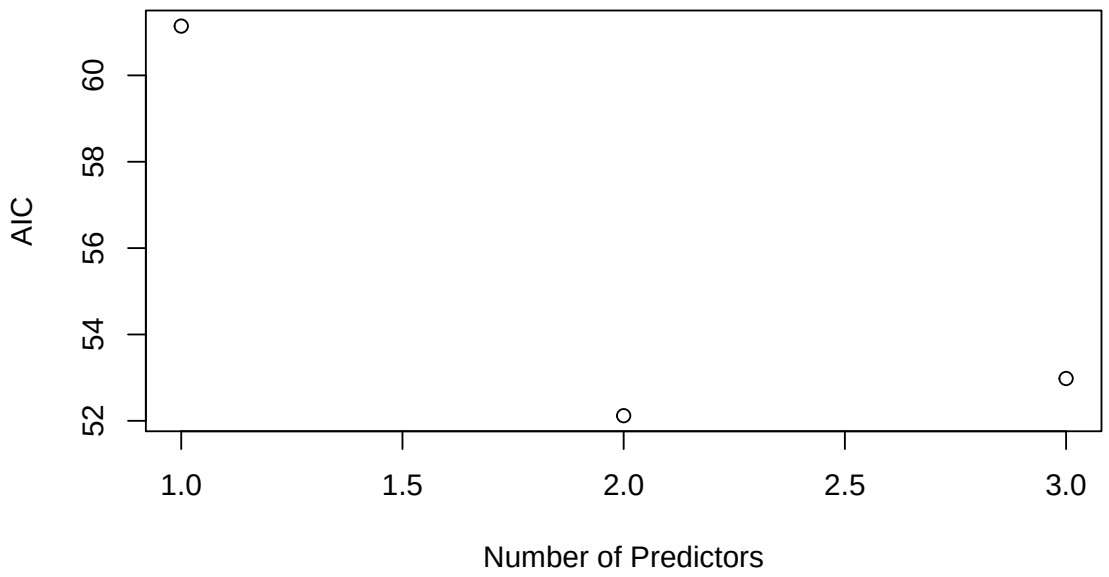
Simplify the model if possible.

Now we run subset selection. We'll use an exhaustive method since there are not too many predictors. And we'll use the Mallow C_p as our criteria.

```
##   (Intercept) Air.Flow Water.Temp Acid.Conc.
## 1        TRUE     TRUE      FALSE      FALSE
```

```
## 2        TRUE      TRUE        TRUE       FALSE
## 3        TRUE      TRUE        TRUE        TRUE
```





The reduced model with the lowest AIC has 2 variables and is $stack.loss \sim Air.Flow + Water.Temp$. We note this is the same model indicated by the Mallow Cp criterion.

Check the model for outliers and influential points.

Check for outliers.

Table 1: Range of Studentized residuals

range.residuals.left	range.residuals.right
-3.471	2.027

Table 2: Bonferroni corrected t-value

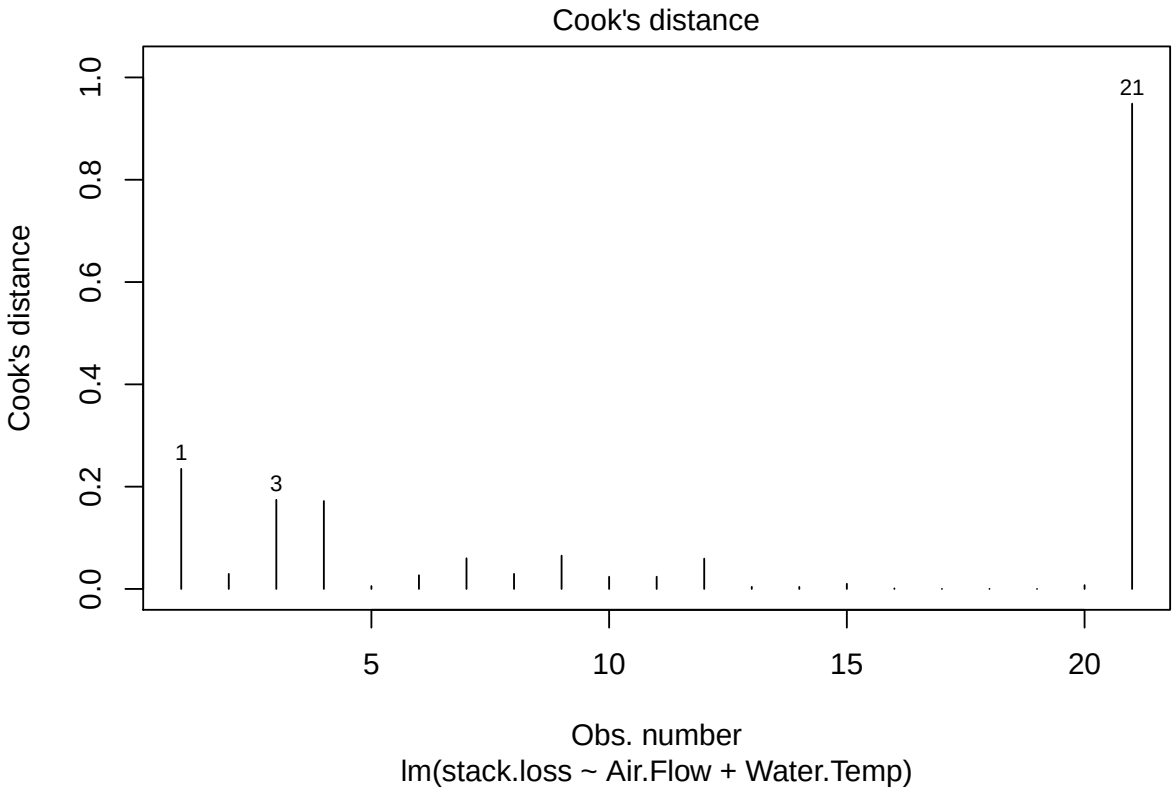
t.val.alpha
-3.565

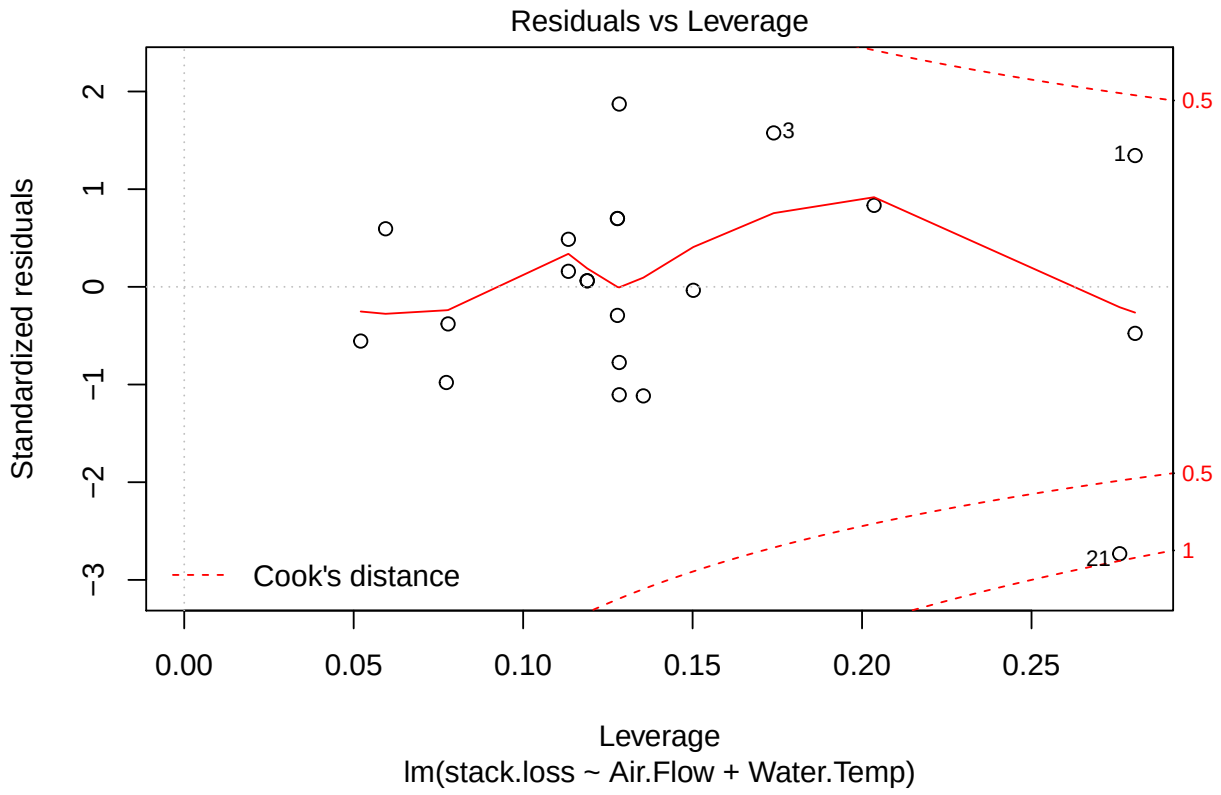
Here we look for studentized residuals that fall outside the interval given by the Bonferroni

corrected t-values. In the case of the reduced model we do not see any outliers.

Check for influential points.

We plot the Cook's distances and the residual-leverage plot with level set contours of the Cook distance.





We see that data element 21 is an influential point for the reduced model under the criteria $D_i > \frac{1}{2}$. Elements 1 and 3 are also influential under the criteria $D_i > \frac{4}{n}$

Now return to the full model, determine whether there are any outliers or influential points

Check for outliers.

Table 3: Range of Studentized residuals

range.residuals.left	range.residuals.right
-3.33	2.052

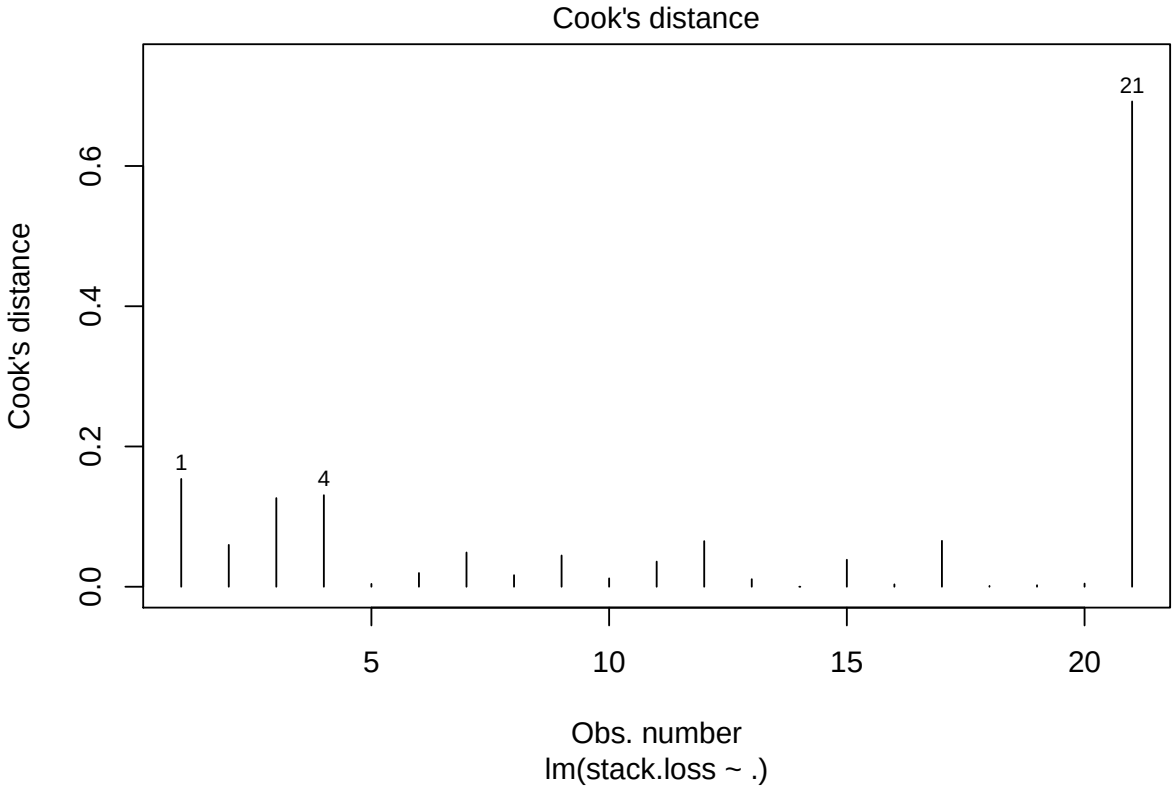
Table 4: Bonferroni corrected t-value

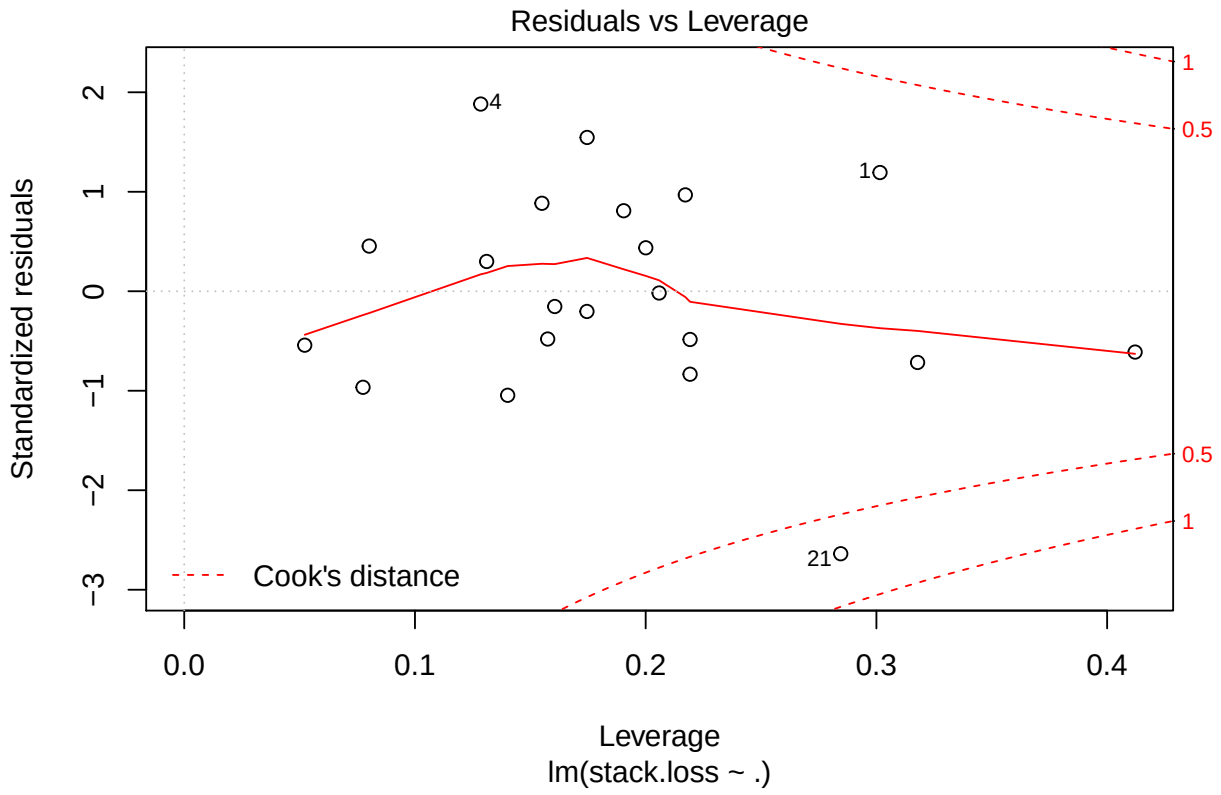
t.val.alpha
-3.604

Here we look for studentized residuals that fall outside the interval given by the Bonferroni corrected t-values. we see there are no outliers for the full model

Check for influential points.

We plot the Cook's distances and the residual-leverage plot with level set contours of the Cook distance.

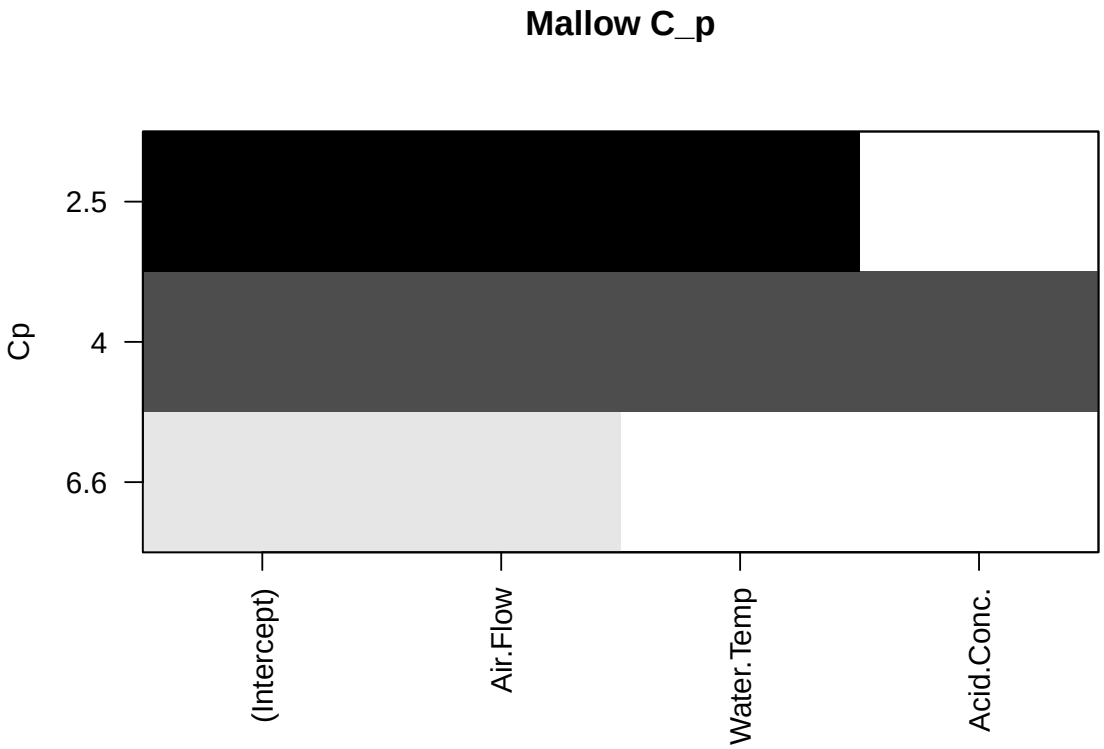


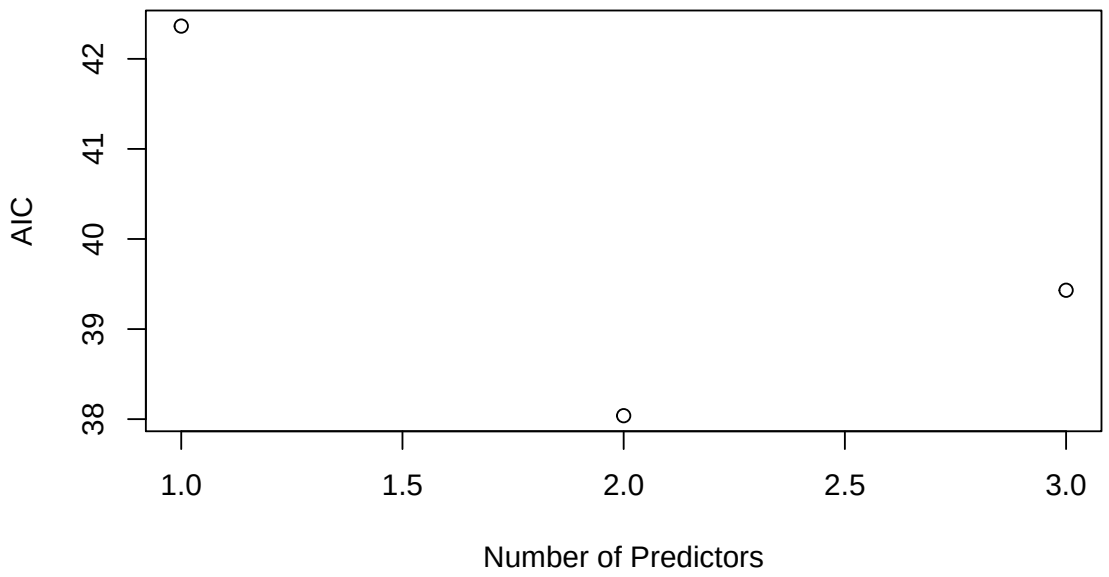


We see element 21 is an influential point, and that 1 and 4 are also influential under the criteria $D_i > \frac{4}{n}$. Since element 1 is an influential in both the full and reduced model we remove that along with element 21.

Eliminate the outliers and influential points for the full model and then repeat the variable selection procedures.

```
##   (Intercept) Air.Flow Water.Temp Acid.Conc.
## 1        TRUE     TRUE      FALSE      FALSE
## 2        TRUE     TRUE       TRUE      FALSE
## 3        TRUE     TRUE       TRUE       TRUE
```


Mallow C_p
Cp
2.5
4
6.6
(Intercept)
Air.Flow
Water.Temp
Acid.Conc.



We see that the subset selection routine has chosen the same model $stack.loss \sim Air.Flow + Water.Temp$.