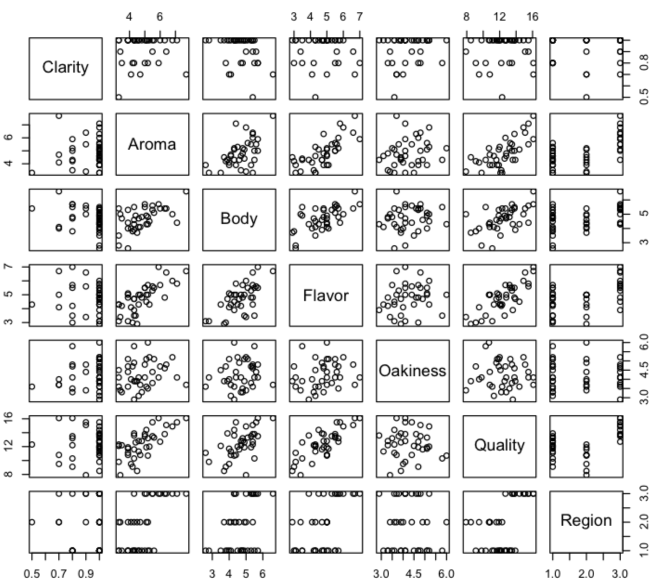
Linear Statistical Analysis : Homework 6

Problem 9.14

Table B.11 presents data on the quality of Pinot Noir wine.

a. Build an appropriate regression model for quality y using the all-possible-regressions approach. Use C_p as the model selection criterion, and incorporate the region information by using indicator variables.

```
Pinot=read.csv("data-table-B11.csv")

plot(Pinot)
```



```
library(leaps)
sets=regsubsets(Quality~Clarity+Aroma+Body+Flavor+Oakiness+factor(Region),nbest=2,
data=Pinot,method="exhaustive")
p1=summary(sets)
```

```
> p1$outmat
         Clarity Aroma Body Flavor Oakiness factor(Region)2
1  ( 1 ) " "     " "   " "  "*"    " "      " "
1  ( 2 ) " "     " "   " "  " "    " "      " "
2  ( 1 ) " "     " "   " "  "*"    " "      "*"
2  ( 2 ) " "     " "   " "  "*"    " "      " "
```

```
3  ( 1 ) " "      " "   " "  "*"    " "       "*"
3  ( 2 ) " "      " "   " "  "*"    "*"       "*"
4  ( 1 ) " "      " "   " "  "*"    "*"       "*"
4  ( 2 ) "*"      " "   " "  "*"    " "       "*"
5  ( 1 ) " "      "*"   " "  "*"    "*"       "*"
5  ( 2 ) " "      " "   "*"  "*"    "*"       "*"
6  ( 1 ) " "      "*"   "*"  "*"    "*"       "*"
6  ( 2 ) "*"      "*"   " "  "*"    "*"       "*"
7  ( 1 ) "*"      "*"   "*"  "*"    "*"       "*"
         factor(Region)3
1  ( 1 ) " "
1  ( 2 ) "*"
2  ( 1 ) " "
2  ( 2 ) "*"
3  ( 1 ) "*"
3  ( 2 ) " "
4  ( 1 ) "*"
4  ( 2 ) "*"
5  ( 1 ) "*"
5  ( 2 ) "*"
6  ( 1 ) "*"
6  ( 2 ) "*"
7  ( 1 ) "*"
```

```
> p1$adjr2
 [1] 0.6137349 0.5087726 0.7630989 0.7196368 0.8086792 0.7816549
 [7] 0.8164362 0.8044750 0.8115597 0.8114228 0.8061475 0.8055761
[13] 0.7996867
```

```
> cbind(p1$outmat, p1$cp, p1$bic, p1$adjr2)
         Clarity Aroma Body Flavor Oakiness factor(Region)2 factor(Region)3
1  ( 1 ) " "      " "   " "  "*"    " "       " "             " "
"35.4189781709754" "-29.9127780555374" "0.613734877270133"
1  ( 2 ) " "      " "   " "  " "    " "       " "             "*"
"54.2826430935391" "-20.7782153145939" "0.508772573899674"
2  ( 1 ) " "      " "   " "  "*"    " "       "*"             " "
"9.39285850527169" "-45.9231650354481" "0.763098857850499"
2  ( 2 ) " "      " "   " "  "*"    " "       " "             "*"
"16.9868283042521" "-39.5223277758607" "0.719636768403399"
3  ( 1 ) " "      " "   " "  "*"    " "       "*"             "*"
"2.47367160324065" "-51.5073698394105" "0.808679169502184"
3  ( 2 ) " "      " "   " "  "*"    "*"       "*"             " "
"7.06060656940056" "-46.4866224446383" "0.781654932209669"
4  ( 1 ) " "      " "   " "  "*"    "*"       "*"             "*"
"2.24065871817499" "-50.5769919318537" "0.816436179914452"
4  ( 2 ) "*"      " "   " "  "*"    " "       "*"             "*"
"4.21116242404491" "-48.1782146658171" "0.804475025525814"
5  ( 1 ) " "      "*"   " "  "*"    "*"       "*"             "*"
"4.10329364568146" "-47.1124100771656" "0.811559687118535"
5  ( 2 ) " "      " "   "*"  "*"    "*"       "*"             "*"
"4.125157851137"   "-47.0848204638463" "0.811422821771991"
6  ( 1 ) " "      "*"   "*"  "*"    "*"       "*"             "*"
"6.00013708327379" "-43.6052641659579" "0.80614753609534"
6  ( 2 ) "*"      "*"   " "  "*"    "*"       "*"             "*"
"6.08856424859786" "-43.4934216853078" "0.805576144577141"
7  ( 1 ) "*"      "*"   "*"  "*"    "*"       "*"             "*"                "8"
"-39.9678516446483" "0.799686702618604"
```

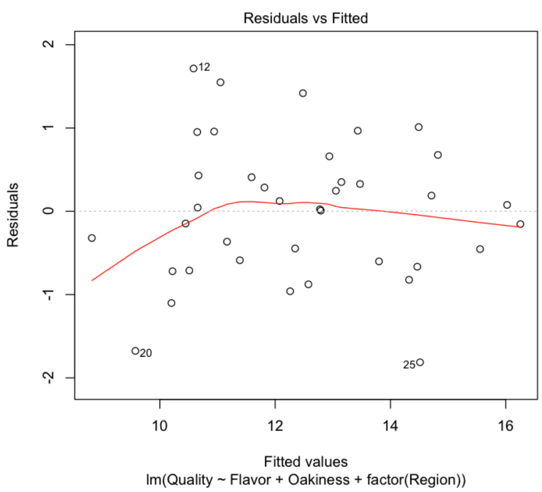
```
> p1$cp
 [1] 35.418978 54.282643  9.392859 16.986828  2.473672  7.060607
 [7]  2.240659  4.211162  4.103294  4.125158  6.000137  6.088564
[13]  8.000000
```

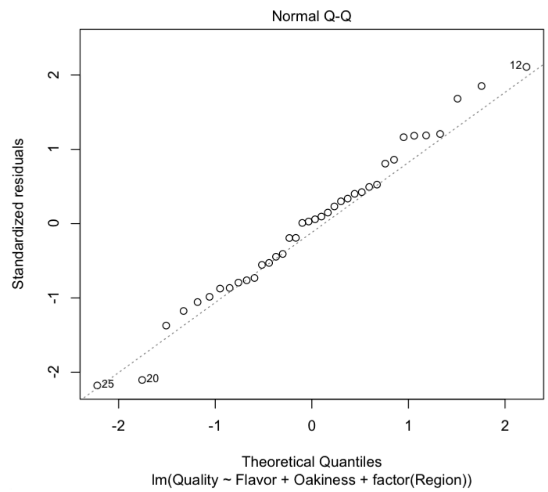
We find that the best model in terms of the Cp value is the model with x4, x5, r1, and r2.

b. For the best two models in terms of C_p, investigate model adequacy by using residual plots. Is there any practical basis for selecting between these models?

The best model includes Oakiness, Flavor and the regions.

```
Pinot.fit.1=lm(Quality~ Flavor+Oakiness +factor(Region),data=Pinot)

> plot(Pinot.fit.1)
```

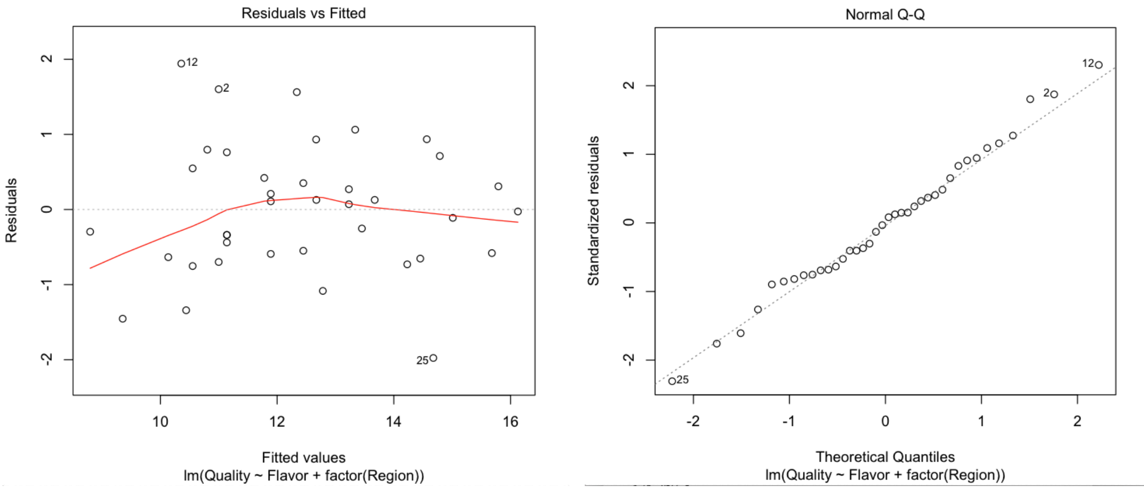




The residual and normality plots look good. There is a slight drift in the residuals on the left of the graph, but it is very slight.

The second best model includes Flavor and the regions.

```
Pinot.fit.2=lm(Quality~ Flavor+factor(Region),data=Pinot)
```



In the second best model the residuals have a slightly higher drift. But, residuals and normality look alright in both models.

c. Is there any difference between the two models in part b in terms of the PRESS statistic?

```
library(MPV)

> PRESS(Pinot.fit.1)
[1] 33.08173

> PRESS(Pinot.fit.2)
[1] 33.99346
```

Looks like the first model is better, judging by the smaller PRESS value.

Problem 9.15

Use the wine quality data in Table B.11 to construct a regression model for quality using the stepwise regression approach. Compare this model to the one you found in Problem 9.14, part a.

```
Pinot.Full=lm(Quality~Clarity+Aroma+Body+Flavor+Oakiness+factor(Region),data=Pinot)


> step(Pinot.Full,data=Pinot,direction="backward")
Start:  AIC=0.3
Quality ~ Clarity + Aroma + Body + Flavor + Oakiness + factor(Region)

                 Df Sum of Sq    RSS     AIC
- Clarity         1    0.0001 25.140 -1.6984
- Body            1    0.0742 25.214 -1.5866
- Aroma           1    0.1041 25.244 -1.5415
<none>                        25.140  0.3014
```

```
- Oakiness         1    1.8525 26.993  1.0031
- factor(Region)   2   18.1079 43.248 16.9159
- Flavor           1   18.1210 43.261 18.9274

Step:  AIC=-1.7
Quality ~ Aroma + Body + Flavor + Oakiness + factor(Region)

                 Df Sum of Sq    RSS     AIC
- Body            1    0.0864 25.227 -3.5680
- Aroma           1    0.1048 25.245 -3.5404
<none>                        25.140 -1.6984
- Oakiness        1    2.0316 27.172 -0.7454
- Flavor          1   18.1527 43.293 16.9554
- factor(Region)  2   20.5655 45.706 17.0162

Step:  AIC=-3.57
Quality ~ Aroma + Flavor + Oakiness + factor(Region)

                 Df Sum of Sq    RSS     AIC
- Aroma           1    0.1151 25.342 -5.3949
<none>                        25.227 -3.5680
- Oakiness        1    1.9841 27.211 -2.6909
- factor(Region)  2   20.6267 45.853 15.1388
- Flavor          1   23.2503 48.477 19.2531

Step:  AIC=-5.39
Quality ~ Flavor + Oakiness + factor(Region)

                 Df Sum of Sq    RSS     AIC
<none>                        25.342 -5.3949
- Oakiness        1     1.871 27.213 -4.6877
- factor(Region)  2    27.114 52.456 18.2508
- Flavor          1    34.753 60.095 25.4169

Call:
lm(formula = Quality ~ Flavor + Oakiness + factor(Region), data = Pinot)

Coefficients:
    (Intercept)           Flavor         Oakiness  factor(Region)2  factor(Region)3
         8.1208           1.1920          -0.3183          -1.5155           1.0935
```

The best model by the stepwise regression approach is the same model that was chosen by the Cp criterion and includes Flavor, Oakiness, and regions.

Problem 9.16

Rework Problem 9.14, part a, but exclude the region information.

a. **Comment on the difference in the models you have found. Is there indication that the region information substantially improves the model?**

```
library(leaps)
sets_no_region=regsubsets(Quality~Clarity+Aroma+Body+Flavor+Oakiness,nbest=2,
data=Pinot,method="exhaustive")


p2=summary(sets_no_region)


> cbind(p2$outmat, p2$cp, p2$bic, p2$adjr2)
         Clarity Aroma Body Flavor Oakiness
1  ( 1 ) " "     " "   " "  "*"    " "      "9.04360465152383" "-29.9127780555374"
"0.613734877270133"
1  ( 2 ) " "     "*"   " "  " "    " "      "23.2301678158677" "-19.0878128587384"
"0.486427357229608"
2  ( 1 ) " "     " "   " "  "*"    "*"      "6.81316027068609" "-30.2064549990156"
"0.641746597249027"
2  ( 2 ) " "     "*"   " "  "*"    " "      "7.10637349837282" "-29.9204636767974"
"0.639040179224373"
3  ( 1 ) " "     "*"   " "  "*"    "*"      "3.92778999698004" "-31.6808161362702"
"0.677628962375389"
3  ( 2 ) "*"     " "   " "  "*"    "*"      "6.63660071389462" "-28.7619100621486"
"0.651890707050826"
4  ( 1 ) "*"     "*"   " "  "*"    "*"      "4.67467766317753" "-29.4733237990903"
"0.68012762500036"
4  ( 2 ) " "     "*"   "*"  "*"    "*"      "5.81851311681163" "-28.1658204311365"
"0.668929921171898"
5  ( 1 ) "*"     "*"   "*"  "*"    "*"      "6"                "-26.6285883249032"
"0.676942845961602"



> p2$cp
[1]  9.043605 23.230168  6.813160  7.106373  3.927790  6.636601  4.674678  5.818513
6.000000



> PRESS(Pinot.fit.1)
[1] 33.08173


Pinot.fit.3=lm(Quality~ Aroma+Flavor+Oakiness,data=Pinot)
> PRESS(Pinot.fit.3)
[1] 56.05239
```

When we do not include region the best model by the Cp selection criterion includes Aroma, Flavor, and Oakiness. Still, this lowest Cp value is 3.927790, while the lowest Cp value that includes region is 2.473672. Also, the PRESS value for this fit is higher than the fit that includes region.

b. Calculate confidence intervals as mean quality for all points in the data set using the models from part a of this problem and Problem 9.14, part a. Based on this analysis, which model would you prefer?

```
> confint(Pinot.fit.1)
                       2.5 %      97.5 %
(Intercept)       6.0530084 10.18862503
Flavor            0.8315302  1.55254852
Oakiness         -0.7331876  0.09655457
factor(Region)2 -2.2508187 -0.78014931
factor(Region)3  0.2780048  1.90909084
```

```
> confint(Pinot.fit.3)
                  2.5 %     97.5 %
(Intercept)  3.75864235  9.1757473
Aroma        0.04729651  1.1129440
Flavor       0.64106744  1.7583182
Oakiness    -1.13965261 -0.0649967
```

```
> Conf1=confint(Pinot.fit.1)
> Conf3=confint(Pinot.fit.3)
```

```
conf_interval_1=Conf1[,2]-Conf1[,1]
> conf_interval_1
    (Intercept)          Flavor        Oakiness factor(Region)2 factor(Region)3
      4.1356166       0.7210184       0.8297422       1.4706694       1.6310860
```

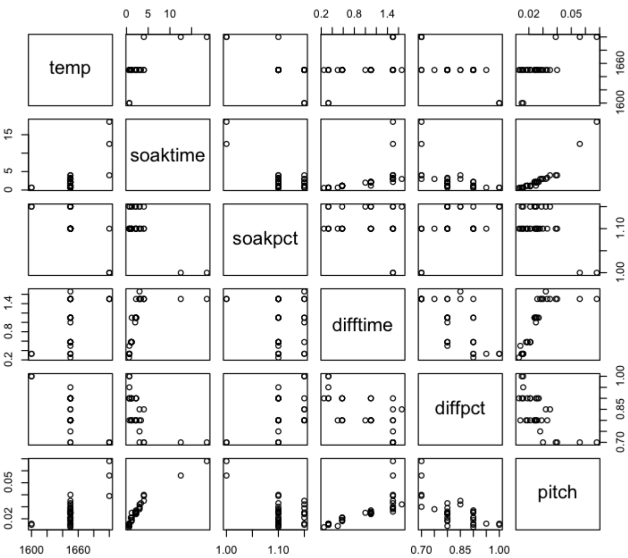
```
conf_interval_3=Conf3[,2]-Conf3[,1]
> conf_interval_3
(Intercept)       Aroma      Flavor    Oakiness
   5.417105    1.065647    1.117251    1.074656
```

The intervals on the regressors for the model that includes region are smaller than the intervals on the regressors for the model without region. This means that the model that includes region has a more precise estimate and is the better model.

Problem 9.17

Table B.12 presents data on a heat treating process used to carburize gears. The thickness of the carburized layer is a critical factor in overall reliability of this component. The response variable y=PITCH is the result of a carbon analysis of the gear pitch for a cross-sectioned part. Use all possible regressions and the C_p criterion to find an appropriate regression model for these data. Investigate model adequacy using residual plots.

```
Pitch=read.csv("data-table-B12.csv")

plot(Pitch)
```



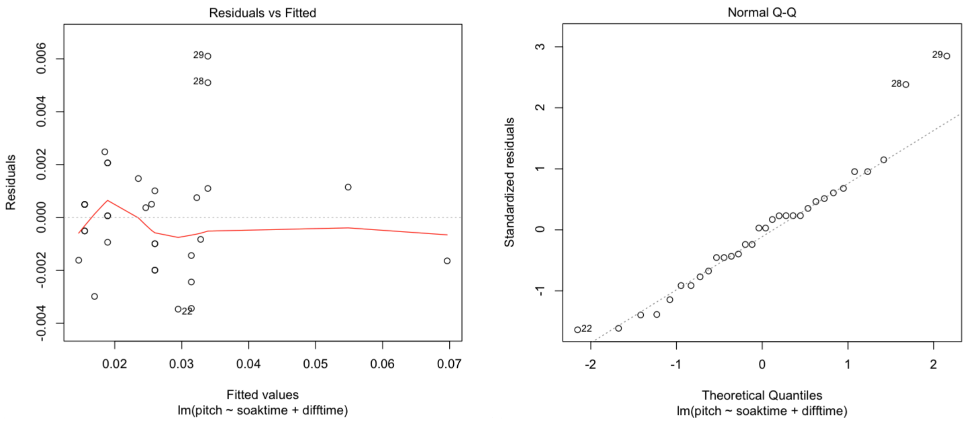

```
sets=regsubsets(pitch~.,nbest=2, data=Pitch,method="exhaustive")
p3=summary(sets)
```

```
> cbind(p3$outmat,p3$cp,p3$bic,p3$adjr2)
         temp soaktime soakpct difftime diffpct
1  ( 1 ) " "  "*"      " "     " "      " "     "77.7627821532872" "-60.4371834517814"
"0.874126853624969"
1  ( 2 ) " "  " "      " "     "*"      " "     "350.164024664589" "-19.665044343417"
"0.549929619273945"
2  ( 1 ) " "  "*"      " "     "*"      " "     "2.75075868204608" "-98.6525523877656"
"0.964602483918959"
2  ( 2 ) " "  "*"      " "     " "      "*"     "49.9845236513393" "-67.5528555062561"
"0.90644895922214"
3  ( 1 ) "*"  "*"      " "     "*"      " "     "3.03085383239283" "-97.1607431320902"
"0.965531434541384"
3  ( 2 ) " "  "*"      " "     "*"      "*"     "3.70088997009585" "-96.3772041417471"
"0.964677033692071"
4  ( 1 ) "*"  "*"      "*"     "*"      " "     "4.27929968143168" "-94.5973241316677"
"0.965248664478841"
4  ( 2 ) "*"  "*"      " "     "*"      "*"     "4.86523168261949" "-93.8916792296786"
"0.964473837150384"
5  ( 1 ) "*"  "*"      "*"     "*"      "*"     "6"                "-91.4735084494281"
"0.964295621632023"
```

```
> p3$cp
[1]  77.762782 350.164025   2.750759  49.984524   3.030854   3.700890   4.279300
4.865232   6.000000
```

The best model by the lowest Cp value includes soaktime and difftime.

```
Pitch.fit.1=lm(pitch~soaktime+difftime ,data=Pitch)


> summary(Pitch.fit.1)

Call:
lm(formula = pitch ~ soaktime + difftime, data = Pitch)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0034655 -0.0014819  0.0000639  0.0010311  0.0060980 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.0110147  0.0009165  12.018 8.74e-13 ***
soaktime    0.0024647  0.0001313  18.773  < 2e-16 ***
difftime    0.0086856  0.0009855   8.814 1.07e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.002223 on 29 degrees of freedom
Multiple R-squared:  0.9669,  Adjusted R-squared:  0.9646 
F-statistic: 423.4 on 2 and 29 DF,  p-value: < 2.2e-16




> anova(Pitch.fit.1)
Analysis of Variance Table

Response: pitch
          Df    Sum Sq   Mean Sq F value    Pr(>F)    
soaktime   1 0.0037994 0.0037994  769.09 < 2.2e-16 ***
difftime   1 0.0003838 0.0003838   77.68 1.068e-09 ***
Residuals 29 0.0001433 0.0000049                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 




r=resid(Pitch.fit.1)
pr=r/(1-lm.influence(Pitch.fit.1)$hat)
> sum(r^2)
[1] 0.0001432658
> sum(pr^2)
[1] 0.0001982679
```



All of the residuals seem to be to the left of the residual plot. But, they are scattered above and below the 0 line pretty evenly. The normality plot looks good. We would maybe consider getting rid of outliers 22, 28, and 29.

Problem 9.19

Repeat Problem 9.17 using the two cross-product variables defined in Problem 9.18 as additional candidate regressors. Comment on the model that you find.

```
Pitch.fit.2=lm(pitch~difftime+temp+soaktime+soakpct+difftime+diffpct+soaktime:soakpct+
difftime:diffpct,data=Pitch)


> summary(Pitch.fit.2)

Call:
lm(formula = pitch ~ difftime + temp + soaktime + soakpct + difftime + 
    diffpct + soaktime:soakpct + difftime:diffpct, data = Pitch)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0035276 -0.0010199 -0.0000239  0.0011133  0.0045464 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)  
(Intercept)      -8.809e-02  8.047e-02  -1.095   0.2845  
difftime          2.537e-02  1.046e-02   2.426   0.0232 *
temp              5.258e-05  3.566e-05   1.475   0.1533  
soaktime         -6.258e-03  6.870e-03  -0.911   0.3713  
soakpct          -3.570e-03  2.526e-02  -0.141   0.8888  
diffpct           1.980e-02  1.398e-02   1.417   0.1695  
soaktime:soakpct  8.599e-03  6.930e-03   1.241   0.2266  
difftime:diffpct -2.287e-02  1.154e-02  -1.982   0.0591 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002065 on 24 degrees of freedom
```

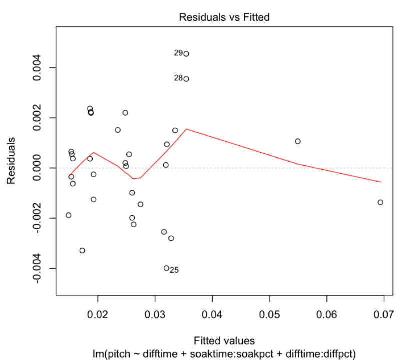
```
Multiple R-squared:  0.9763,  Adjusted R-squared:  0.9694
F-statistic: 141.5 on 7 and 24 DF,  p-value: < 2.2e-16


sets=regsubsets(pitch ~ difftime + temp + soaktime + soakpct + difftime + diffpct +
soaktime:soakpct + difftime:diffpct,nbest=2, data=Pitch,method="exhaustive")


p4=summary(sets)


> cbind(p4$outmat,p4$cp,p4$bic,p4$adjr2)
         difftime temp soaktime soakpct diffpct soaktime:soakpct difftime:diffpct
1  ( 1 ) " "      " "  " "      " "     " "     "*"              " "
"75.3209945931876" "-66.1693253817025" "0.894770341055221"
1  ( 2 ) " "      " "  "*"      " "     " "     " "              " "
"95.5900487225458" "-60.4371834517813" "0.874126853624969"
2  ( 1 ) "*"      " "  " "      " "     " "     "*"              " "
"6.2514889613979"  "-99.9604337348099" "0.966020059732141"
2  ( 2 ) "*"      " "  "*"      " "     " "     " "              " "
"7.59695720914603" "-98.6525523877656" "0.964602483918959"
3  ( 1 ) "*"      " "  " "      " "     " "     "*"              "*"
"4.66950679466354" "-100.262062139574" "0.968715225434827"
3  ( 2 ) "*"      " "  "*"      " "     " "     " "              "*"
"6.03473158187606" "-98.7734107953372" "0.96722546315867"
4  ( 1 ) "*"      " "  "*"      " "     " "     "*"              "*"
"5.34091955379785" "-98.3147138262597" "0.969060008340341"
4  ( 2 ) "*"      "*"  " "      " "     " "     "*"              "*"
"5.77784447067293" "-97.8073769709484" "0.968565568010439"
5  ( 1 ) "*"      "*"  " "      " "     "*"     "*"              "*"
"5.31780090508549" "-97.3090330698475" "0.970247499459586"
5  ( 2 ) "*"      "*"  "*"      " "     " "     "*"              "*"
"6.03644618603724" "-96.4133661443461" "0.96940297531667"
6  ( 1 ) "*"      "*"  "*"      " "     "*"     "*"              "*"
"6.01997750062293" "-95.527196733651"  "0.970643557388847"
6  ( 2 ) "*"      "*"  " "      "*"     "*"     "*"              "*"
"6.82997326811478" "-94.4658961587614" "0.96965360665875"
7  ( 1 ) "*"      "*"  "*"      "*"     "*"     "*"              "*"              "8"
"-92.0880864183708" "0.969445805465064"


p4$cp
 [1] 75.320995 95.590049  6.251489  7.596957  4.669507  6.034732  5.340920  5.777844
5.317801  6.036446  6.019978  6.829973  8.000000
```

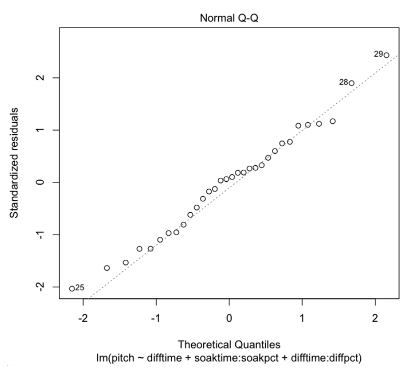
Now, we find that the model that includes difftime, soaktime:soakpct, and difftime:diffpct is the best model because it has the lowest Cp value of 4.669507.

```
Pitch.fit.3=lm(pitch~difftime+soaktime:soakpct+difftime:diffpct,data=Pitch)


> summary(Pitch.fit.3)

Call:
lm(formula = pitch ~ difftime + soaktime:soakpct + difftime:diffpct,
    data = Pitch)
```

```
Residuals:
       Min         1Q     Median         3Q        Max 
-0.0039923 -0.0013881  0.0001607  0.0011741  0.0045521 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       0.0116913  0.0009224  12.674 4.06e-13 ***
difftime          0.0163346  0.0046106   3.543  0.00141 ** 
soaktime:soakpct  0.0024057  0.0001435  16.761 3.95e-16 ***
difftime:diffpct -0.0107909  0.0057694  -1.870  0.07192 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00209 on 28 degrees of freedom
Multiple R-squared:  0.9717,	Adjusted R-squared:  0.9687 
F-statistic:   321 on 3 and 28 DF,  p-value: < 2.2e-16
```

The F-statistic is very high and the p-value: < 2.2e-16 is very low for this model of pitch.





The residuals are again to the left of the graph, but are spread about the zero line well. The normality plot looks good and is even better than before with the previous model.

```
> PRESS(Pitch.fit.3)
[1] 0.0001879593

> PRESS(Pitch.fit.1)
[1] 0.0001982679
```

The PRESS statistic is smaller for the new model with the cross product variables.