Assignment 7

Biplav Timalsina

April 15, 2018

STAT 757 Applied Regression Analysis

Exercise 7.5.3 [60 points]

This is a continuation of Exercise 5 in Chapter 6. The golf fan was so impressed with your answers to part 1 that your advice has been sought re the next stage in the data analysis, namely using model selection to remove the redundancy in full the model developed in part 1.

$log(Y) = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \beta_4 x4 + \beta_5 x5 + \beta_6 x6 + \beta_7 x7 + \epsilon$............(7.10)

where

Y = PrizeMoney ; x1 = Driving Accuracy ; x2 = GIR ; x3 = PuttingAverage ; x4 = BirdieConversion ; x5 = SandSaves ; x6 = Scrambling; and x7 = PuttsPerRound.

Interest centers on using variable selection to choose a subset of the predictors to model the transformed version of Y . Throughout this question we shall assume that model (7.10) is a valid model for the data.

Let us first make a model fitting all the covariates.

We can observe that only two covariate, GIR and BirdieConversion are the statistically significant, and Residual standard erros is 0.6639 which is fair. Now we move to variable selection.

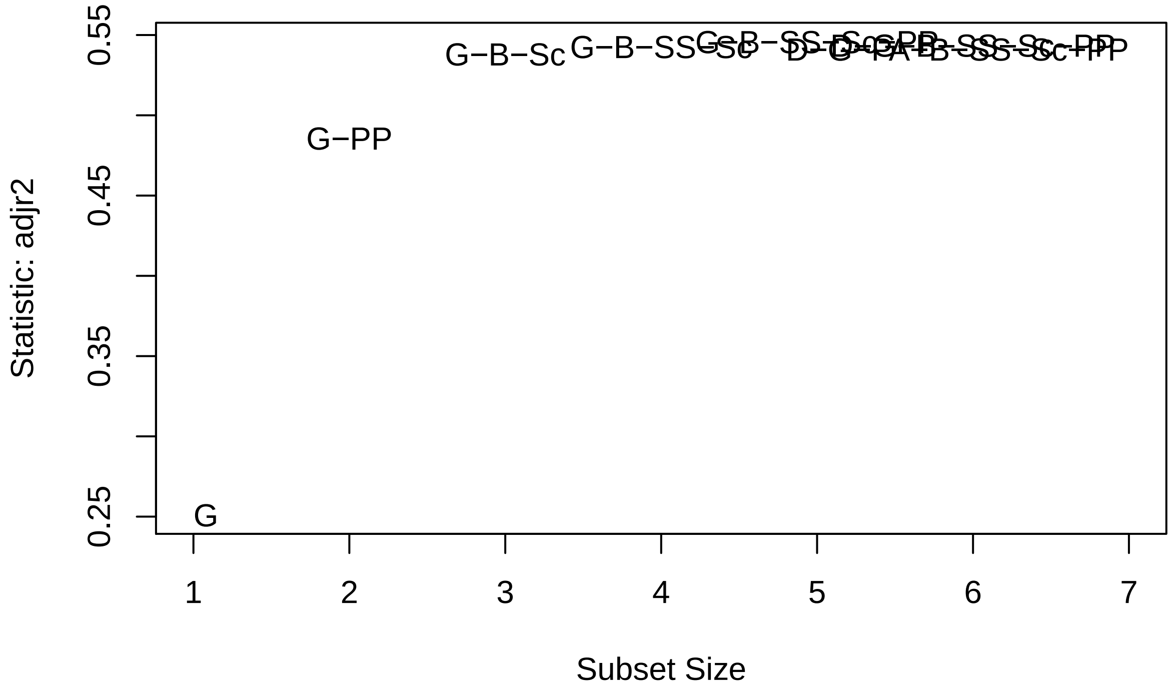
```r
#Creating all possible subsets
X <- cbind(DrivingAccuracy,GIR,PuttingAverage,BirdieConversion,SandSaves,Scrambling,PuttsPerRound)
library(leaps)
b <- regsubsets(as.matrix(X),log(PrizeMoney))
rs <- summary(b)
rs
```

```
## Subset selection object
## 7 Variables  (and intercept)
##                  Forced in Forced out
## DrivingAccuracy      FALSE      FALSE
## GIR                  FALSE      FALSE
## PuttingAverage       FALSE      FALSE
## BirdieConversion     FALSE      FALSE
## SandSaves            FALSE      FALSE
## Scrambling           FALSE      FALSE
## PuttsPerRound        FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          DrivingAccuracy GIR PuttingAverage BirdieConversion SandSaves
## 1  ( 1 ) " "             "*" " "            " "              " "      
## 2  ( 1 ) " "             "*" " "            " "              " "      
## 3  ( 1 ) " "             "*" " "            "*"              " "      
## 4  ( 1 ) " "             "*" " "            "*"              "*"      
## 5  ( 1 ) " "             "*" " "            "*"              "*"      
## 6  ( 1 ) "*"             "*" " "            "*"              "*"      
## 7  ( 1 ) "*"             "*" "*"            "*"              "*"      
##          Scrambling PuttsPerRound
## 1  ( 1 ) " "        " "          
## 2  ( 1 ) " "        "*"          
## 3  ( 1 ) "*"        " "          
## 4  ( 1 ) "*"        " "          
## 5  ( 1 ) "*"        "*"          
## 6  ( 1 ) "*"        "*"          
## 7  ( 1 ) "*"        "*"
```

```
par(mfrow=c(1,1))
library(car)
```

```
## Loading required package: carData
```

```
subsets(b,statistic=c("adjr2"),legend=FALSE)
```



```
##                    Abbreviation
## DrivingAccuracy               D
## GIR                           G
## PuttingAverage               PA
## BirdieConversion              B
## SandSaves                    SS
## Scrambling                   Sc
## PuttsPerRound                PP
```

We can observe above the best subsets of predictors for different values of number of covariates. Accordingly, we can develop models selecting corresponding covariates, and observing the values of $R^2_A dj$, AIC,AICc and BIC. Below we calculate all these values.

```
#Table values of ${R^2}adj$ , AIC, $AIC_C$, BIC
#Calculate adjusted R-squared
rs$adjr2
```

```
## [1] 0.2510765 0.4857746 0.5381917 0.5427792 0.5458520 0.5436566 0.5412404
```

```
om1 <- lm(log(PrizeMoney)~GIR)
#summary(om1)
```

```
om2 <- lm(log(PrizeMoney)~GIR+PuttsPerRound)
#summary(om2)
om3 <-lm(log(PrizeMoney)~GIR+BirdieConversion+Scrambling)
#summary(om3)
om4 <- lm(log(PrizeMoney)~GIR+BirdieConversion+Scrambling+SandSaves)
#summary(om4)
om5 <- lm(log(PrizeMoney)~GIR+BirdieConversion+Scrambling+SandSaves+PuttsPerRound)
#summary(om5)
om6 <- lm(log(PrizeMoney)~GIR+BirdieConversion+Scrambling+SandSaves+PuttsPerRound+DrivingAccuracy)
#summary(om6)
om7 <- lm(log(PrizeMoney)~DrivingAccuracy+GIR+PuttingAverage+BirdieConversion+SandSaves+Scrambling+Putts
#summary(om7)
```

These are all the models developed. Now we extract the values of AIC, BIC and AICc for various models. GIR GIR+PPR GIR+BC+SC GIR+BC+SC+SS GIR+BC+SC+SS+PPR GIR+BC+SC+SS+PPR+DA GIR+BC+SC+SS+PPR+DA+PA

```
#Subset size=1
n <- length(om1$residuals)
npar <- length(om1$coefficients) +1
#Calculate AIC
extractAIC(om1,k=2)
```

```
## [1]   2.00000 -62.51622
```

```
#Calculate AICc
extractAIC(om1,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]   2.12500 -62.39122
```

```
#Calculate BIC
extractAIC(om1,k=log(n))
```

```
## [1]   2.00000 -55.95999
```

```
#Subset size=2
npar <- length(om2$coefficients) +1
#Calculate AIC
extractAIC(om2,k=2)
```

```
## [1]    3.0000 -135.2203
```

```
#Calculate AICc
extractAIC(om2,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    3.209424 -135.010849
```

```
#Calculate BIC
extractAIC(om2,k=log(n))
```

```
## [1]    3.0000 -125.3859
```

```
#Subset size=3
npar <- length(om3$coefficients) +1
#Calculate AIC
extractAIC(om3,k=2)
```

```
## [1]    4.0000 -155.3108
```

```r
#Calculate AICc
extractAIC(om3,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    4.315789 -154.994976
```

```r
#Calculate BIC
extractAIC(om3,k=log(n))
```

```
## [1]    4.0000 -142.1983
```

```r
#Subset size=4
npar <- length(om4$coefficients) +1
#Calculate AIC
extractAIC(om4,k=2)
```

```
## [1]    5.000 -156.291
```

```r
#Calculate AICc
extractAIC(om4,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    5.444444 -155.846577
```

```r
#Calculate BIC
extractAIC(om4,k=log(n))
```

```
## [1]    5.0000 -139.9004
```

```r
#Subset size=5
npar <- length(om5$coefficients) +1
#Calculate AIC
extractAIC(om5,k=2)
```

```
## [1]    6.0000 -156.6416
```

```r
#Calculate AICc
extractAIC(om5,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    6.595745 -156.045867
```

```r
#Calculate BIC
extractAIC(om5,k=log(n))
```

```
## [1]    6.0000 -136.9729
```

```r
#Subset size=6
npar <- length(om6$coefficients) +1
#Calculate AIC
extractAIC(om6,k=2)
```

```
## [1]    7.0000 -154.7307
```

```r
#Calculate AICc
extractAIC(om6,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    7.770053 -153.960655
```

```r
#Calculate BIC
extractAIC(om6,k=log(n))
```

```
## [1]    7.0000 -131.7839
```

```r
#Subset size=7
npar <- length(om7$coefficients) +1
#Calculate AIC
extractAIC(om7,k=2)
```

```
## [1]    8.0000 -152.7355
```

```r
#Calculate AICc
extractAIC(om7,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    8.967742 -151.767720
```

```r
#Calculate BIC
extractAIC(om7,k=log(n))
```

```
## [1]    8.0000 -126.5105
```

(a) Identify the optimal model or models based on R^2Adj , AIC, AICC, BIC from the approach based on all possible subsets.

Interpretation: From the calculation above, we can populate the following table.

Subset size	Predictors	$R^2{}_{Adj}$	AIC	AIC_C	BIC
1	GIR	0.251	-62.516	-61.54	-55.95
2	GIR+PPR	0.486	-135.220	-135.011	-125.386
3	GIR+BC+SC	0.538	-155.311	-154.99	**-142.198**
4	GIR+BC+SC+SS	0.543	-156.291	-155.846	-139.901
5	GIR+BC+SC+SS+PPR	**0.546**	**-156.642**	**-156.0432**	-136.973
6	GIR+BC+SC+SS+PPR+DA	0.544	-154.731	-153.960	-131.784
7	GIR+BC+SC+SS+PPR+DA+PA	0.541	-152.736	-151.76	-126.511

The table above gives the values of $R^2{}_{Adj}$, AIC, AIC_C and BIC for the best subset of each size. From the table, we notice that $R^2 - Adj$, AIC & AIC_C judge the predictor subset of size 5 to be "best" while BIC judges the subset of size 3 to be the "best."

```r
om3 <-lm(log(PrizeMoney)~GIR+BirdieConversion+Scrambling)
summary(om3)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ GIR + BirdieConversion + Scrambling)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71081 -0.50717 -0.06683  0.41975  2.04147
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -11.08314    1.45712  -7.606 1.23e-12 ***
## GIR                0.15658    0.01787   8.761 1.01e-15 ***
## BirdieConversion   0.20625    0.02164   9.531  < 2e-16 ***
## Scrambling         0.09178    0.01539   5.965 1.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6661 on 192 degrees of freedom
## Multiple R-squared:  0.5453, Adjusted R-squared:  0.5382
## F-statistic: 76.75 on 3 and 192 DF,  p-value: < 2.2e-16
```

```
om5 <- lm(log(PrizeMoney)~GIR+BirdieConversion+Scrambling+SandSaves+PuttsPerRound)
summary(om5)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ GIR + BirdieConversion + Scrambling +
##     SandSaves + PuttsPerRound)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71291 -0.48168 -0.09097  0.44843  2.15763
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.583181   7.158721  -0.081   0.9352
## GIR               0.197022   0.028711   6.862 9.31e-11 ***
## BirdieConversion  0.162752   0.032672   4.981 1.41e-06 ***
## Scrambling        0.049635   0.024738   2.006   0.0462 *
## SandSaves         0.015524   0.009743   1.593   0.1127
## PuttsPerRound    -0.349738   0.230995  -1.514   0.1317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6606 on 190 degrees of freedom
## Multiple R-squared:  0.5575, Adjusted R-squared:  0.5459
## F-statistic: 47.88 on 5 and 190 DF,  p-value: < 2.2e-16
```

Notice that three predictor variables are judged to be statistically significant for both models (Scrambling is only slightly significant in 5 variable model).

A popular data anlaysis strategy which I have adopted according to book is to calculate R^2_{Adj}, AIC, AIC_C and BIC and then seelct the model which minimizes AIC, AIC_C and BIC. Accordingly, since Model5 has the maximum R^2_{Adj} with 0.546 and minimum AIC_C and AIC, but Model3 only has minimized BIC with -142.198, I think BIC has penalized model 5 because it has larger number of covariate in it. So, I think model5 is better in this case. But for model simplicity,we could choose model3.

(b) Identify the optimal model or models based on AIC and BIC from the approach based on backward selection.

```
#Stepwise Subsets
#Output from R: Backward Elimination based on AIC and BIC
backAIC <- step(om7,direction="backward", data=PGA)
```

```
## Start:  AIC=-152.74
## log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                    Df Sum of Sq    RSS     AIC
## - PuttingAverage    1    0.0020 82.868 -154.73
## - DrivingAccuracy   1    0.0396 82.905 -154.64
## - PuttsPerRound     1    0.2314 83.097 -154.19
## <none>                          82.866 -152.74
## - SandSaves         1    1.0436 83.909 -152.28
## - Scrambling        1    1.1576 84.023 -152.02
## - BirdieConversion  1    6.6928 89.558 -139.51
## - GIR               1    9.1200 91.986 -134.27
##
## Step:  AIC=-154.73
## log(PrizeMoney) ~ DrivingAccuracy + GIR + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                    Df Sum of Sq    RSS     AIC
## - DrivingAccuracy   1    0.0377 82.905 -156.64
## <none>                          82.868 -154.73
## - PuttsPerRound     1    1.0263 83.894 -154.32
## - SandSaves         1    1.0461 83.914 -154.27
## - Scrambling        1    1.7855 84.653 -152.55
## - BirdieConversion  1    8.6663 91.534 -137.24
## - GIR               1   17.0549 99.922 -120.05
##
## Step:  AIC=-156.64
## log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves + Scrambling +
##     PuttsPerRound
##
##                    Df Sum of Sq     RSS     AIC
## <none>                           82.905 -156.64
## - PuttsPerRound     1    1.0003  83.905 -156.29
## - SandSaves         1    1.1078  84.013 -156.04
## - Scrambling        1    1.7566  84.662 -154.53
## - BirdieConversion  1   10.8275  93.733 -134.58
## - GIR               1   20.5479 103.453 -115.24
```

```
backBIC <- step(om7,direction="backward", data=PGA, k=log(n))
```

```
## Start:  AIC=-126.51
## log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                    Df Sum of Sq    RSS     AIC
## - PuttingAverage    1    0.0020 82.868 -131.78
## - DrivingAccuracy   1    0.0396 82.905 -131.69
## - PuttsPerRound     1    0.2314 83.097 -131.24
## - SandSaves         1    1.0436 83.909 -129.34
## - Scrambling        1    1.1576 84.023 -129.07
## <none>                          82.866 -126.51
## - BirdieConversion  1    6.6928 89.558 -116.56
## - GIR               1    9.1200 91.986 -111.32
##
## Step:  AIC=-131.78
## log(PrizeMoney) ~ DrivingAccuracy + GIR + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                    Df Sum of Sq    RSS     AIC
## - DrivingAccuracy   1    0.0377 82.905 -136.97
## - PuttsPerRound     1    1.0263 83.894 -134.65
## - SandSaves         1    1.0461 83.914 -134.60
## - Scrambling        1    1.7855 84.653 -132.88
## <none>                          82.868 -131.78
## - BirdieConversion  1    8.6663 91.534 -117.57
## - GIR               1   17.0549 99.922 -100.38
```

```
##
## Step:  AIC=-136.97
## log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves + Scrambling +
##     PuttsPerRound
##
##                    Df Sum of Sq     RSS      AIC
## - PuttsPerRound     1    1.0003  83.905 -139.900
## - SandSaves         1    1.1078  84.013 -139.649
## - Scrambling        1    1.7566  84.662 -138.141
## <none>                           82.905 -136.973
## - BirdieConversion  1   10.8275  93.733 -118.192
## - GIR               1   20.5479 103.453  -98.853
##
## Step:  AIC=-139.9
## log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves + Scrambling
##
##                    Df Sum of Sq     RSS      AIC
## - SandSaves         1     1.286  85.191 -142.198
## <none>                           83.905 -139.900
## - Scrambling        1     7.595  91.501 -128.194
## - GIR               1    35.317 119.222  -76.324
## - BirdieConversion  1    36.555 120.460  -74.299
##
## Step:  AIC=-142.2
## log(PrizeMoney) ~ GIR + BirdieConversion + Scrambling
##
##                    Df Sum of Sq     RSS      AIC
## <none>                           85.191 -142.198
## - Scrambling        1    15.786 100.977 -114.157
## - GIR               1    34.057 119.248  -81.560
## - BirdieConversion  1    40.308 125.499  -71.545
```

Interpretation:

From the backward elimination based on AIC, we choose the model with the five predictors GIR, BirdieConversiton, SandSaves, Scrambling and PuttsPerRound. It can be seen that backward elimination based on BIC chooses the model with the three predictors GIR, BirdieConversion and Scrambling.

(c) Identify the optimal model or models based on AIC and BIC from the approach based on forward selection.

```r
#Output from R: Forward Selection based on AIC and BIC
mint <- lm(log(PrizeMoney)~1,data=PGA)
forwardAIC <- step(mint,scope=list(lower=~1,
upper=~DrivingAccuracy+GIR+PuttingAverage+BirdieConversion+SandSaves+Scrambling+PuttsPerRound),
direction="forward", data=PGA)
```

```
## Start:  AIC=-6.84
## log(PrizeMoney) ~ 1
##
##                    Df Sum of Sq    RSS     AIC
## + GIR               1    47.760 139.59 -62.516
## + BirdieConversion  1    40.930 146.43 -53.154
## + PuttingAverage    1    34.660 152.69 -44.936
## + Scrambling        1    25.260 162.09 -33.227
## + SandSaves         1    10.926 176.43 -16.618
## + PuttsPerRound     1     6.295 181.06 -11.540
## + DrivingAccuracy   1     6.184 181.17 -11.419
## <none>                          187.35  -6.841
##
## Step:  AIC=-62.52
## log(PrizeMoney) ~ GIR
##
##                     Df Sum of Sq     RSS      AIC
## + PuttsPerRound     1    44.240  95.355 -135.220
## + PuttingAverage    1    39.748  99.847 -126.197
## + BirdieConversion  1    38.618 100.977 -123.991
## + SandSaves         1    15.043 124.552  -82.864
## + Scrambling        1    14.096 125.499  -81.380
## <none>                          139.595  -62.516
## + DrivingAccuracy   1     0.185 139.410  -60.776
##
## Step:  AIC=-135.22
## log(PrizeMoney) ~ GIR + PuttsPerRound
##
##                     Df Sum of Sq    RSS     AIC
## + BirdieConversion  1    8.1732 87.181 -150.78
## + DrivingAccuracy   1    2.6309 92.724 -138.70
## + SandSaves         1    1.1746 94.180 -135.65
## + PuttingAverage    1    1.0592 94.295 -135.41
## <none>                          95.355 -135.22
## + Scrambling        1    0.0510 95.304 -133.32
##
## Step:  AIC=-150.78
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion
##
##                    Df Sum of Sq    RSS     AIC
## + Scrambling       1    3.1684 84.013 -156.04
## + SandSaves        1    2.5196 84.662 -154.53
## + PuttingAverage   1    1.2574 85.924 -151.63
## <none>                         87.181 -150.78
## + DrivingAccuracy  1    0.0611 87.120 -148.92
##
## Step:  AIC=-156.04
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling
##
##                    Df Sum of Sq    RSS     AIC
## + SandSaves        1   1.10778 82.905 -156.64
## <none>                         84.013 -156.04
## + DrivingAccuracy  1   0.09937 83.914 -154.27
## + PuttingAverage   1   0.00033 84.013 -154.04
##
## Step:  AIC=-156.64
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling +
##     SandSaves
##
##                    Df Sum of Sq    RSS     AIC
## <none>                         82.905 -156.64
## + DrivingAccuracy  1  0.037678 82.868 -154.73
## + PuttingAverage   1  0.000062 82.905 -154.64
```

```
forwardBIC <- step(mint,scope=list(lower=~1,
upper=~DrivingAccuracy+GIR+PuttingAverage+BirdieConversion+SandSaves+Scrambling+PuttsPerRound),
direction="forward", data=PGA,k=log(n))
```

```
## Start:  AIC=-3.56
## log(PrizeMoney) ~ 1
##
##                     Df Sum of Sq    RSS     AIC
## + GIR                1    47.760 139.59 -55.960
## + BirdieConversion   1    40.930 146.43 -46.597
## + PuttingAverage     1    34.660 152.69 -38.379
## + Scrambling         1    25.260 162.09 -26.671
## + SandSaves          1    10.926 176.43 -10.062
## + PuttsPerRound      1     6.295 181.06  -4.983
## + DrivingAccuracy    1     6.184 181.17  -4.863
## <none>                          187.35  -3.563
##
## Step:  AIC=-55.96
## log(PrizeMoney) ~ GIR
##
##                     Df Sum of Sq     RSS      AIC
## + PuttsPerRound      1    44.240  95.355 -125.386
## + PuttingAverage     1    39.748  99.847 -116.362
## + BirdieConversion   1    38.618 100.977 -114.157
## + SandSaves          1    15.043 124.552  -73.030
## + Scrambling         1    14.096 125.499  -71.545
## <none>                          139.595  -55.960
## + DrivingAccuracy    1     0.185 139.410  -50.941
##
## Step:  AIC=-125.39
## log(PrizeMoney) ~ GIR + PuttsPerRound
##
##                     Df Sum of Sq    RSS     AIC
## + BirdieConversion   1    8.1732 87.181 -137.67
## + DrivingAccuracy    1    2.6309 92.724 -125.59
## <none>                          95.355 -125.39
## + SandSaves          1    1.1746 94.180 -122.54
## + PuttingAverage     1    1.0592 94.295 -122.30
## + Scrambling         1    0.0510 95.304 -120.21
##
## Step:  AIC=-137.67
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion
##
##                   Df Sum of Sq    RSS     AIC
## + Scrambling       1    3.1684 84.013 -139.65
## + SandSaves        1    2.5196 84.662 -138.14
## <none>                        87.181 -137.67
## + PuttingAverage   1    1.2574 85.924 -135.24
## + DrivingAccuracy  1    0.0611 87.120 -132.53
##
## Step:  AIC=-139.65
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling
##
##                   Df Sum of Sq    RSS     AIC
## <none>                           84.013 -139.65
## + SandSaves         1   1.10778 82.905 -136.97
## + DrivingAccuracy   1   0.09937 83.914 -134.60
## + PuttingAverage    1   0.00033 84.013 -134.37
```

```
#detach(PGA)
```

Interpretation: We notice that both for ForwardAIC, we get model with 5 covariates as follows. $log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling + SandSaves $

We notice that both for ForwardBIC, we get model with 4 covariates as follows. $log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling $

(d) Carefully explain why the models chosen in (a) & (c) are not the same while those in (a) and (b) are the same. For a, we go over all possible subsets of the models and hence it is exhaustive and has considered everything. For b, backward selection first starts with all the covariates, and removes those variates which are not significant one by one. Adding a variable now might confound the contribution of future covariates (the interplay between variable might be missed. Thus, only those variates are removed who do not contribute much to the model. For c, forward selection first starts with one variable that contributes least p-value. While doing so once a variable is added, it might negate the presence of other variates coming in.

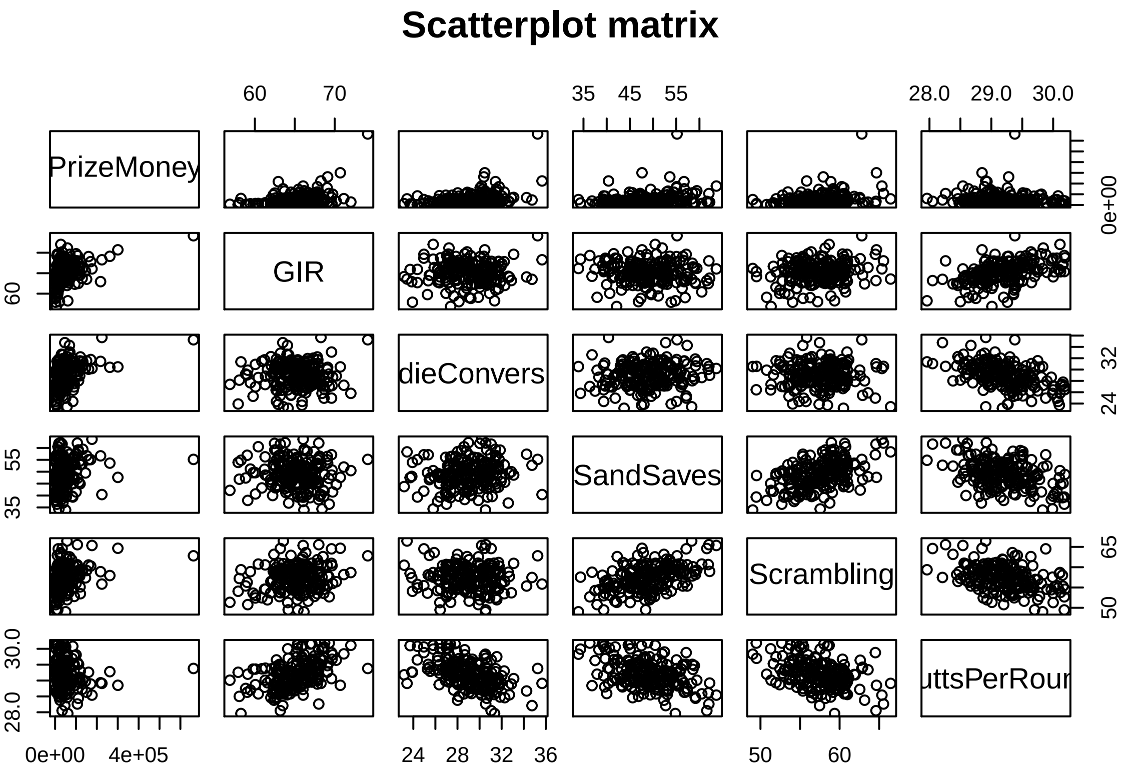
Since a and c does its work in a different style, (a considers possible subset of all predictors, but c doesn't consider all possible covariate,), hence they are different.

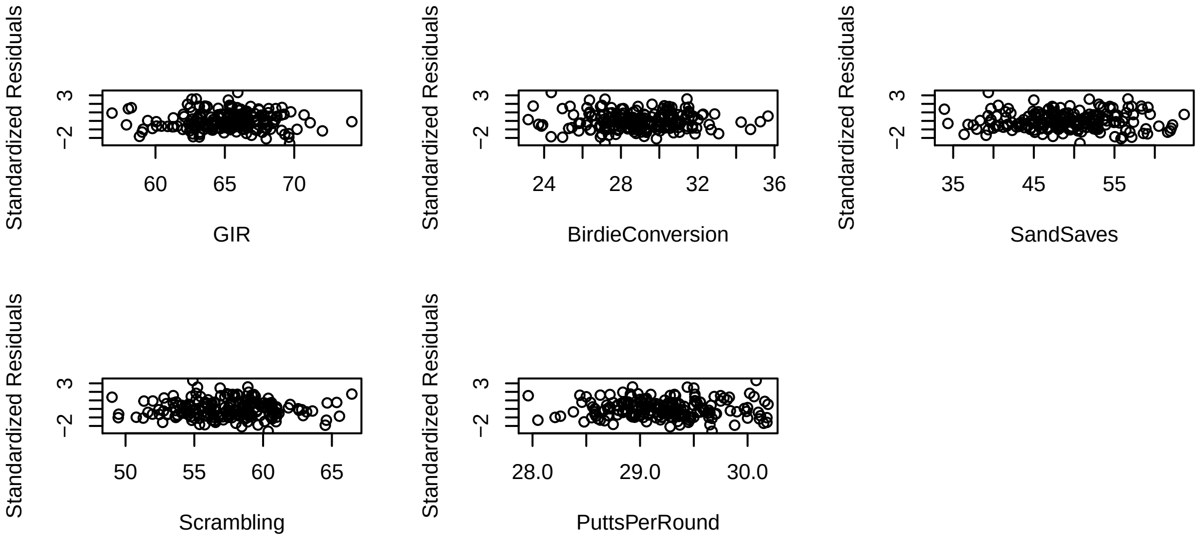
But b starts with all the covariates and removes the variable with largest p value while there is interplay within variables.

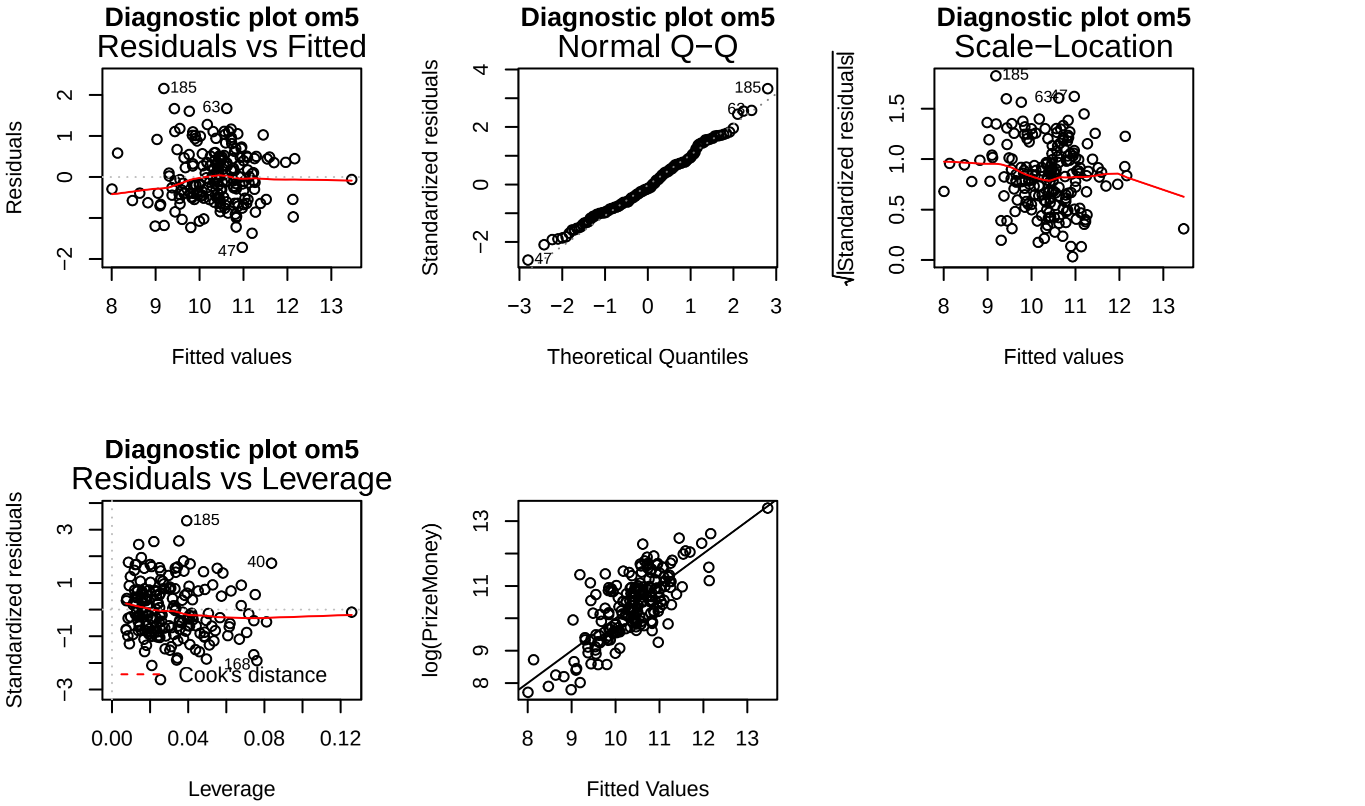
(e) Recommend a final model. Give detailed reasons to support your choice.
Given all the finding and data earlier, I would recommend model with 5 variables. The reasons are:

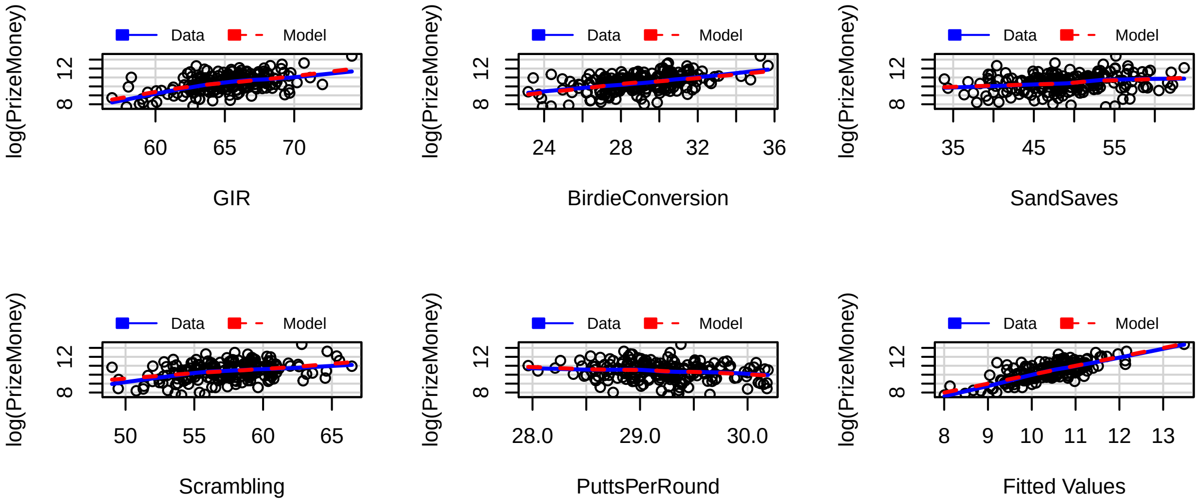
1) R_adj is maximum for this model.
2) AIC_c and AIC is lowest for this model.
3) Although from BIC, this model is not the most suitable, this might also be because of it penalizing more for 5 variables compared to 3 variables.

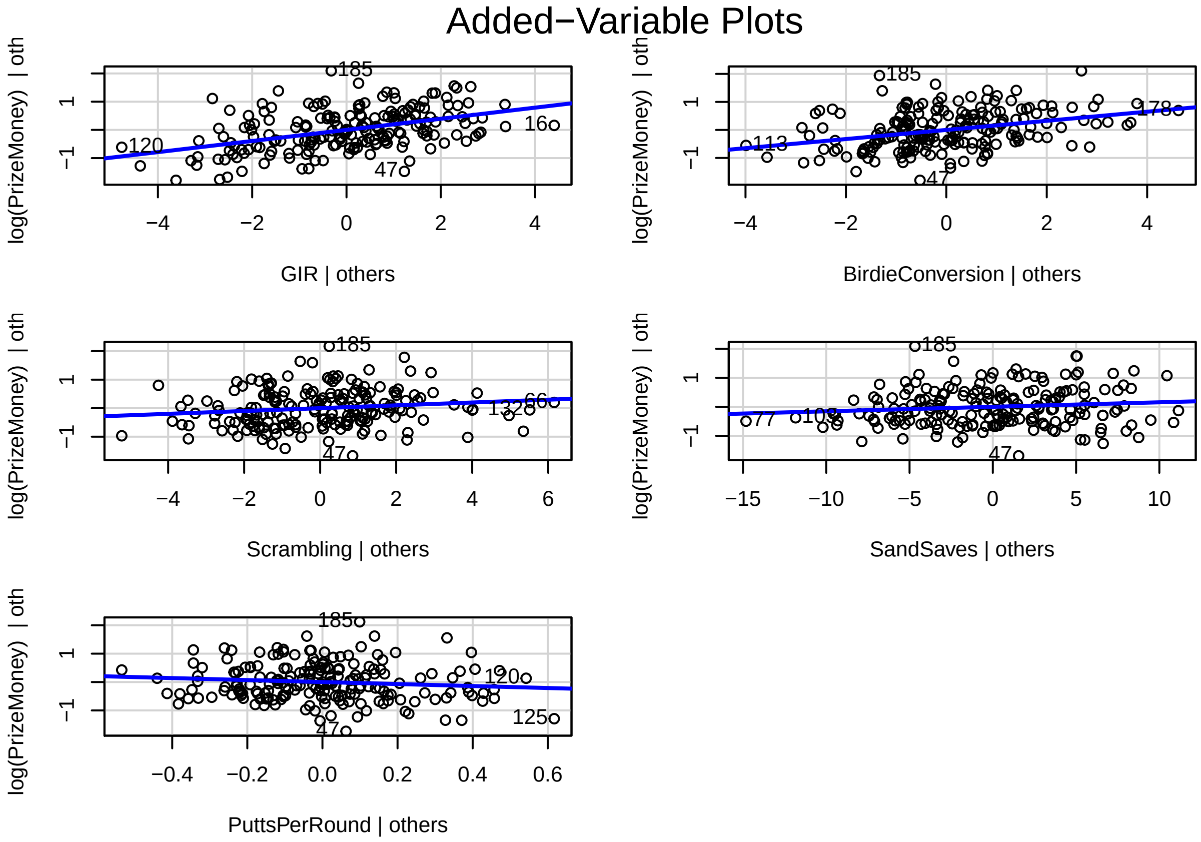
Scatter plot matrix






Diagnostic plot om5
Residuals vs Fitted
Residuals
Fitted values
185
63
47
Diagnostic plot om5
Normal Q−Q
Standardized residuals
Theoretical Quantiles
185
63
47
Diagnostic plot om5
Scale−Location
√|Standardized residuals|
Fitted values
185
63
47
Diagnostic plot om5
Residuals vs Leverage
Standardized residuals
Leverage
185
40
168
Cook's distance
log(PrizeMoney)
Fitted Values


Data
Model
log(PrizeMoney)
8
12
60
65
70
GIR
24
28
32
36
BirdieConversion
35
45
55
SandSaves
50
55
60
65
Scrambling
28.0
29.0
30.0
PuttsPerRound
8
9
10
11
12
13
Fitted Values



We can see from above added variable plots that GIR and BirdieConversion are changing and all others are almost constant. 4) From marginal model plots, the data and model lines are almost same,which indicates good fit. 5)

```
##              GIR BirdieConversion       Scrambling        SandSaves
##         2.730165         2.322693         2.734765         1.441054
##    PuttsPerRound
##         4.652336
```

We can see that our variables are not inflated, but later in our model we see - sign for PuttsPerSecond, which implies sign error.

(f) Interpret the regression coefficients in the final model. Is it necessary to be cautious about taking these results to literally?

The final model is: Log(Prizemoney)= -0.58+0.197(GIR)+ 0.162(BirdieConversion)+ 0.049*(Scrambling)+ 0.015(SandSaves)-0.349(PuttsPerRound)

For every unit change in GIR, PrizeMoney roughly changes with e^(0.197). Similarly, for 0.162, 0.049, 0.015 and -0.349 increase in BirdieCOnversion, Scrambling, SandSaves and PuttsPerSecond, there is roughly e^(0.162), e^(0.049), e^(0.015) and e^(-0.349) increase in Prize Money.

Our model that we have selected was not the best according to BIC, hence might be too complex to interpret and might have overfit the data as well. ALso the selection process changes the properties of the estimators as well as the standard inferential procedures such as test and confidence intervals. The regression coefficients obtained after variable selection are biased. Hence, yes it is definitely necessary to be cautious about taking these results too literally.

Project Milestones

1. Conduct your data analysis plan. . Apply your model to fake data and ensure a proper fit. . Apply your model to real data. . Decide whether model is valid for the real data.
2. Refine your model as needed until you are satisfied with the fit. . Don't make decisions based on p-values or other inferential devices! . Only consider the fit and whether your model addresses your research hypothesis.

Ans:

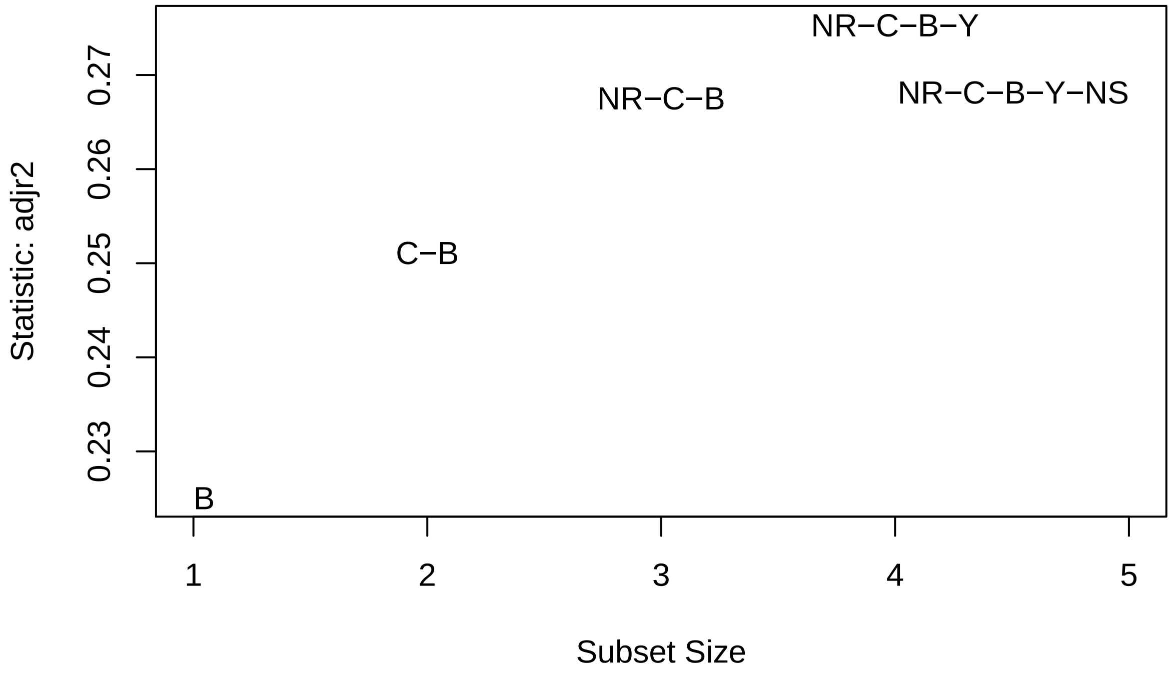
This project work is done along with Dhurba Neupane.

We start by fitting the data to generate a model. TO do so, lets use the method we used in this assignment. But before doing so, we divide the data into test and train data. We have a total of 128 data, out of which 100 is taken as train and 28 are used as training data.

```
NS<-read.csv("F:/unr/4th sem/applied regression analysis/Assignments/HW7/NStudy.csv",header = TRUE)
newFullData<-NS[,c(3,4,5,6,7,8,9,10)]
newData<-NS[1:100,c(3,4,5,6,7,8,9,10)]
attach(newData)
#Creating all possible subsets
X <- cbind(NRate,Cultivar,Block,Year,NSource)
library(leaps)
b <- regsubsets(as.matrix(X),Seedyield)
rs <- summary(b)
rs
```

```
## Subset selection object
## 5 Variables  (and intercept)
##          Forced in Forced out
## NRate        FALSE      FALSE
## Cultivar     FALSE      FALSE
## Block        FALSE      FALSE
## Year         FALSE      FALSE
## NSource      FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          NRate Cultivar Block Year NSource
## 1  ( 1 ) " "   " "      "*"   " "  " "
## 2  ( 1 ) " "   "*"      "*"   " "  " "
## 3  ( 1 ) "*"   "*"      "*"   " "  " "
## 4  ( 1 ) "*"   "*"      "*"   "*"  " "
## 5  ( 1 ) "*"   "*"      "*"   "*"  "*"
```

```
par(mfrow=c(1,1))
library(car)
subsets(b,statistic=c("adjr2"),legend=FALSE)
```



```
##          Abbreviation
## NRate              NR
## Cultivar            C
## Block               B
## Year                Y
## NSource            NS
```

We observed the step in which the model was generated in the previous step. Now we study the five models.

```
#Table values of ${R^2}adj$ , AIC, $AIC_C$, BIC
#Calculate adjusted R-squared
rs$adjr2
```

```
## [1] 0.2250729 0.2511338 0.2675698 0.2753340 0.2681911
```

```
om1 <- lm(Seedyield~Block)
#summary(om1)
om2 <- lm(Seedyield~Block+ Cultivar)
#summary(om2)
om3 <-lm(Seedyield~Block + Cultivar + Year)
#summary(om3)
om4 <- lm(Seedyield~Block+ Cultivar + Year + NRate)
#summary(om4)
om5 <- lm(Seedyield~Block+ Cultivar + Year + NRate+NSource)
#summary(om5)

#Calculate AIC, BIC, AICc
```

```
#Subset size=1
n <- length(om1$residuals)
npar <- length(om1$coefficients) +1
#Calculate AIC
extractAIC(om1,k=2)
```

```
## [1]    2.000 1255.577
```

```
#Calculate AICc
extractAIC(om1,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    2.250 1255.827
```

```
#Calculate BIC
extractAIC(om1,k=log(n))
```

```
## [1]    2.000 1260.787
```

```
#Subset size=2
npar <- length(om2$coefficients) +1
#Calculate AIC
extractAIC(om2,k=2)
```

```
## [1]    3.00 1253.13
```

```
#Calculate AICc
extractAIC(om2,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    3.421053 1253.551182
```

```
#Calculate BIC
extractAIC(om2,k=log(n))
```

```
## [1]    3.000 1260.946
```

```
#Subset size=3
npar <- length(om3$coefficients) +1
#Calculate AIC
extractAIC(om3,k=2)
```

```
## [1]    4.000 1253.294
```

```
#Calculate AICc
extractAIC(om3,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    4.638298 1253.932279
```

```
#Calculate BIC
extractAIC(om3,k=log(n))
```

```
## [1]    4.000 1263.715
```

```
#Subset size=4
npar <- length(om4$coefficients) +1
#Calculate AIC
extractAIC(om4,k=2)
```

```
## [1]    7.000 1246.217
```

```
#Calculate AICc
extractAIC(om4,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]    8.582418 1247.799910
```

```
#Calculate BIC
extractAIC(om4,k=log(n))
```

```
## [1]    7.000 1264.454
```

```
#Subset size=5
npar <- length(om5$coefficients) +1
#Calculate AIC
extractAIC(om5,k=2)
```

```
## [1]    8.000 1248.146
```

```
#Calculate AICc
extractAIC(om5,k=2)+2*npar*(npar+1)/(n-npar-1)
```

```
## [1]   10.000 1250.146
```

```
#Calculate BIC
extractAIC(om5,k=log(n))
```

```
## [1]    8.000 1268.988
```

```
detach(newData)
```

0.2250729 0.2511338 0.2675698 0.2753340 0.2681911

Subset size	Predictors	R^2_{Adj}	AIC	AIC_C	BIC
1	Block	0.2250729	1255.57	1255.82	**1260.787**
2	Block+Cultivar	0.2511338	1253.13	1253.551182	1260.946
3	Block + Cultivar + Year	0.2675698	1253.294	1253.932279	1263.715
4	Block + Cultiar + Year + NRate	**0.2753340**	**1246.217**	**1247.799910**	1264.454
5	Block +CUltivar+Year+NRate+NSource	0.2681911	1248.14	1250.146	1268.98

It seems from the above table that the fourth model with all the variables is best according to {R^2_adj}, AIC and AIC_C, while first model is best accordinng to BIC. A popular data anlaysis strategy which I have adopted according to book is to calculate R^2_{Adj}, AIC, AIC_C and BIC and then seelct the model which minimizes AIC, AIC_C and BIC. Accordingly, since Model4 has the maximum R^2_{Adj} with 0.2625 and minimum AIC_C and AIC, but Model1 only has minimized BIC with 1253.13, I think BIC has penalized model 4 because it has larger number of covariate in it. So, I think model4 is better in this case. But for model simplicity,we could choose model1.

```
testData<-NS[101:128,c(3,4,5,6,7)]
attach(testData)

pred<-predict.lm(om4, testData)
actualY<-NS[101:128,8]
rmse<-sqrt(mean((pred-actualY)**2)/28)
rmse
```

```
## [1] 68.47441
```

```
detach(testData)
```

We observed the root mean square erros as 68.47 which is acceptable as the range of SeedYield is [31.06,2479.09]. But for 31.06 this doesn't make sense. However, we understand that this is an exception case.

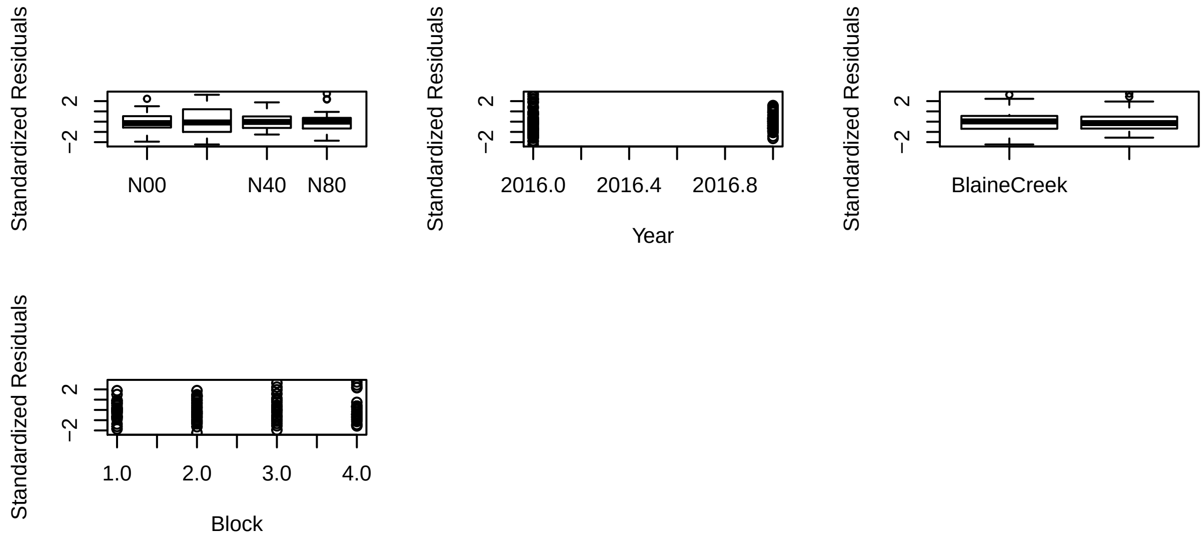
Scatter plot matrix

Hence, the model that we have chosen is:

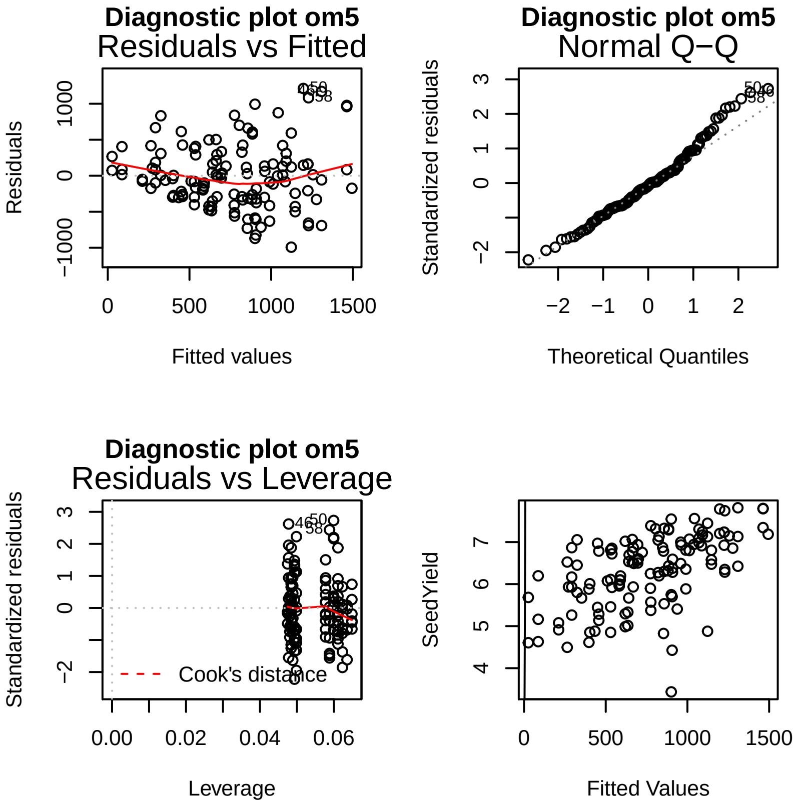
$SeedYield = 480353.88 + 185.47 Block - 265.64 Cultivar.Proghorn - 238.1 Year + 407.64 NRate120 + 60.48 NRate40 + 376.31 NRate80 $

```
# Residual plots against each predictor
newData<-NS[,c(3,4,5,6,7,8,9,10)]
attach(newData)
par(mfrow=c(3,3))
StanRes1 <- rstandard(om4)
plot(NRate,StanRes1, ylab="Standardized Residuals")
plot(Year,StanRes1, ylab="Standardized Residuals")
plot(Cultivar,StanRes1, ylab="Standardized Residuals")
plot(Block,StanRes1, ylab="Standardized Residuals")
#plot(NSource,StanRes1, ylab="Standardized Residuals")

# Diagnostic plots for the regression model
par(mfrow=c(2,3))
```

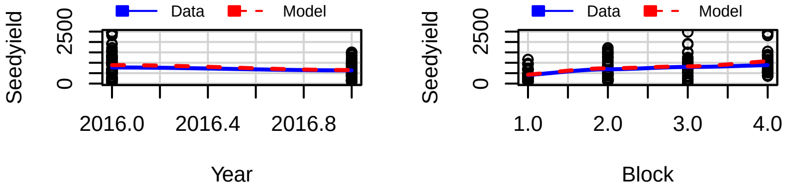


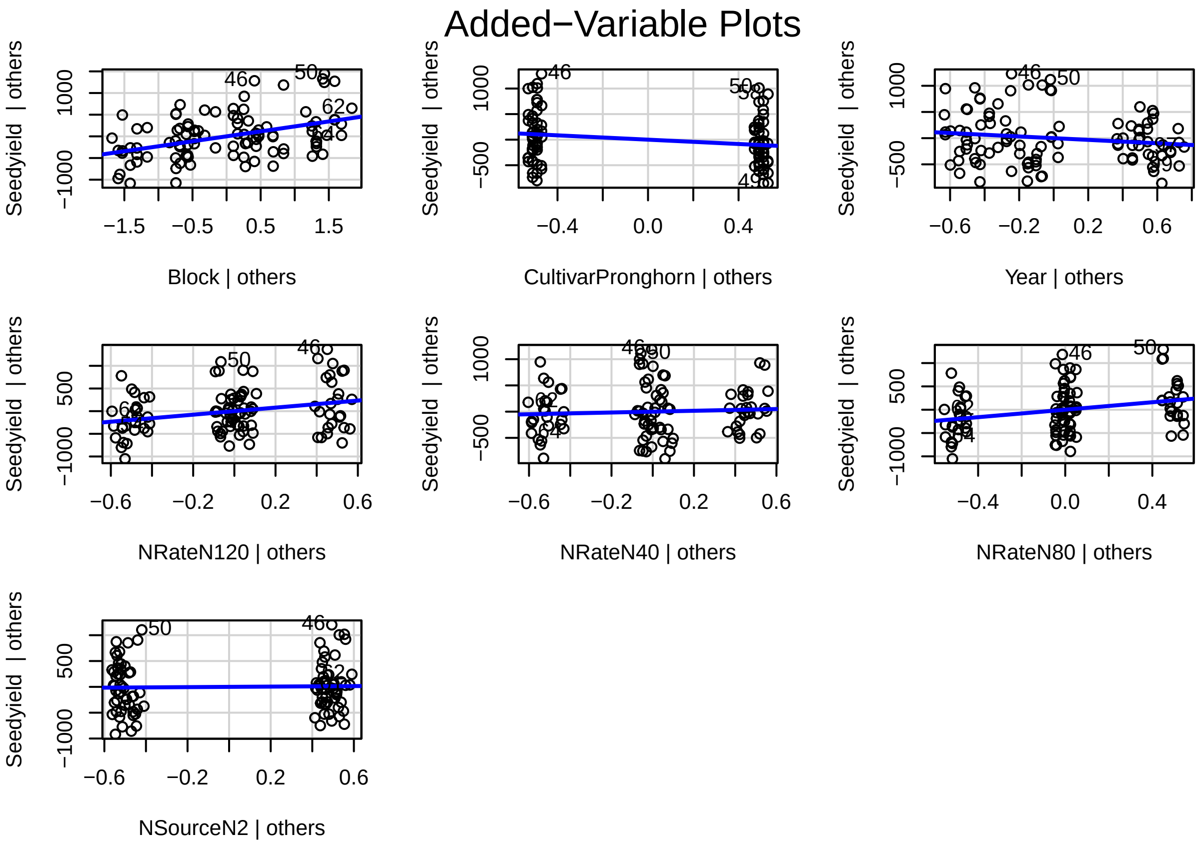
```
plot(om4,main="Diagnostic plot om5")
plot(om4$fitted.values,log(Seedyield),ylab ="SeedYield",xlab="Fitted Values")
abline(lsfit(om4$fitted.values,Seedyield))
par(mfrow=c(1,2))
```



```r
# Marginal model plots
par(mfrow=c(3,3))
mmp(om4,Year)
mmp(om4,Block)
#mmp(om4,Cultivar)
#mmp(om4,NRate)
#mmp(om4,NSource)

# Added-variable plots
avPlots(om5)
```


Data
Model
Seedyield
0
2500
2016.0
2016.4
2016.8
Year
Data
Model
Seedyield
0
2500
1.0
2.0
3.0
4.0
Block



From Added Variable plot, we an see that Block is the most signifant among all covariates. But, all other are also significant. From the residual plot, we can see that the variance is almost constant. None of the data are outliers.

From all the above observations, plots and root mean square error we can say that the model is a valid model.