Assignment 5

Biplav Timalsina

April 01, 2018

STAT 757 Applied Regression Analysis

3. Chateau Latour is widely acknowledged as one of the world's greatest wine estates with a rich history dating back to at least 1638. The Grand Vin de Chateau Latour is a wine of incredible power and longevity. At a tasting in New York in April 2000, the 1863 and 1899 vintages of Latour were rated alongside the 1945 and the 1961 vintages as the best in a line-up of 39 vintages ranging from 1863 to 1999 (*Wine Spectator*, August 31, 2000). Quality of a particular vintage of Chateau Latour has a huge impact on price. For example, in March 2007, the 1997 vintage of Chateau Latour could be purchased for as little as $159 per bottle while the 2000 vintage of Chateau Latour costs as least $700 per bottle (www.wine-searcher.com).

   While many studies have identified that the timing of the harvest of the grapes has an important effect on the quality of the vintage, with quality improving the earlier the harvest. A less explored issue of interest is the effect of unwanted rain at vintage time on the quality of icon wine like Chateau Latour. This question addresses this issue.

The Chateau Latour web site (www.chateau-latour.com) provides a rich source of data. In particular, data on the quality of each vintage, harvest dates and weather at harvest time were obtained from the site for the vintages from 1961 to 2004. An example of the information on weather at harvest time is given below for the 1994 vintage:

> Harvest began on the 13th September and lasted on the 29th, frequently interrupted by storm showers. But quite amazingly the dilution effect in the grapes was very limited …. (http://www.chateau-latour.com/commentaires/1994uk.html"; Accessed: March 16, 2007)

Each vintage was classified as having had "unwanted rain at harvest" (e.g., the 1994 vintage) or not (e.g., the 1996 vintage) on the basis of information like that reproduced above. Thus, the data consist of:

Vintage = year the grapes were harvested
Quality – on a scale from 1 (worst) to 5 (best) with some half points
End of harvest – measured as the number days since August 31
Rain – a dummy variable for unwanted rain at harvest = 1 if yes.
The data can be found on the book web site in the file latour.csv.
The first model considered was:

$$\begin{aligned}\text{Quality} = \beta_0 &+ \beta_1\text{End of Harvest} + \beta_2\text{Rain}\\ &+ \beta_3\text{End of Harvest} \times \text{Rain} + e \end{aligned} \tag{5.10}$$

A plot of the data and the two regression lines from model (5.10) can be found in Figure 5.8. In addition, numerical output appears below.

(a) Show that the coefficient of the interaction term in model (5.10) is statistically significant. In other words, show that the rate of change in quality rating depends on whether there has been any unwanted rain at vintage.

(b) Estimate the number of days of delay to the end of harvest it takes to decrease the quality rating by 1 point when there is:

(i) No unwanted rain at harvest

(ii) Some unwanted rain at harvest

Regression output from R

```
Call:
lm(formula = Quality ~ EndofHarvest + Rain +
Rain:EndofHarvest)
Coefficients:
                   Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         5.16122     0.68917    7.489  3.95e-09 ***
EndofHarvest       -0.03145     0.01760   -1.787    0.0816 .
Rain                1.78670     1.31740    1.356    0.1826
EndofHarvest:Rain  -0.08314     0.03160   -2.631    0.0120 *
---
Residual  standard  error: 0.7578  on  40 degrees  of freedom
Multiple  R-Squared:   0.6848,       Adjusted R-squared: 0.6612
F-statistic:   28.97 on 3 and   40 DF, p-value: 4.017e-10
```

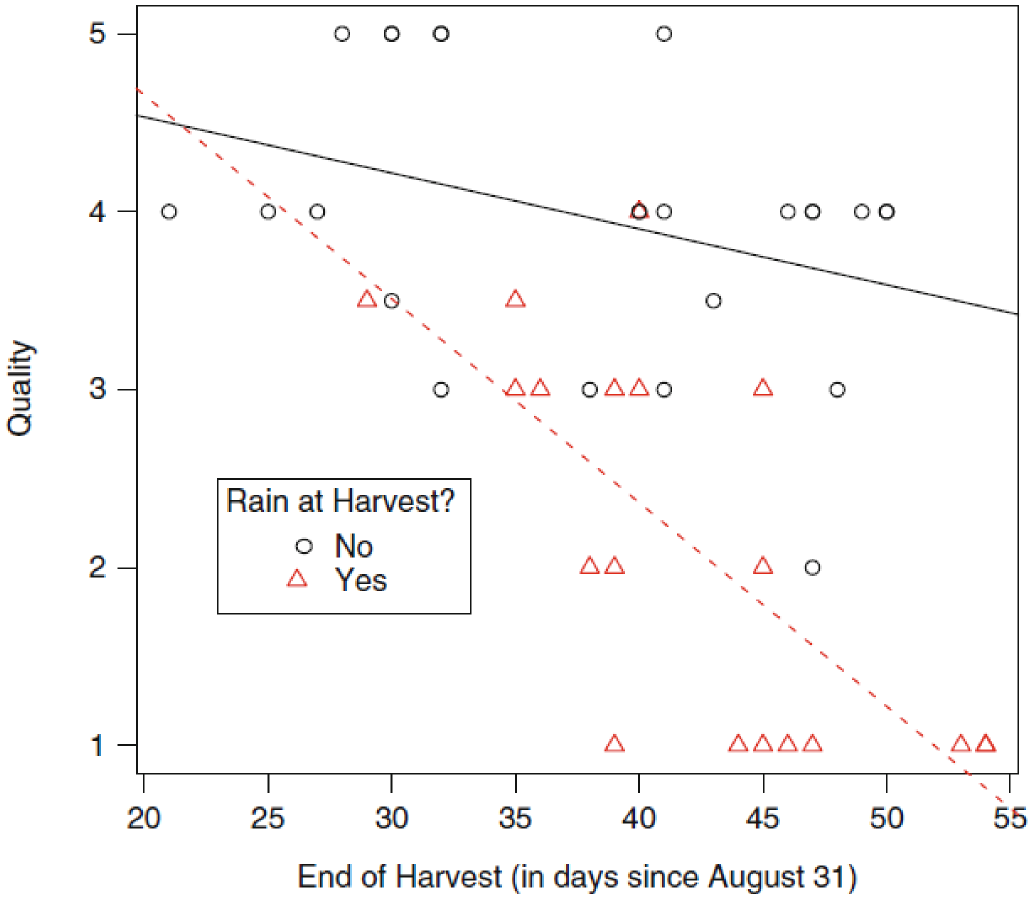
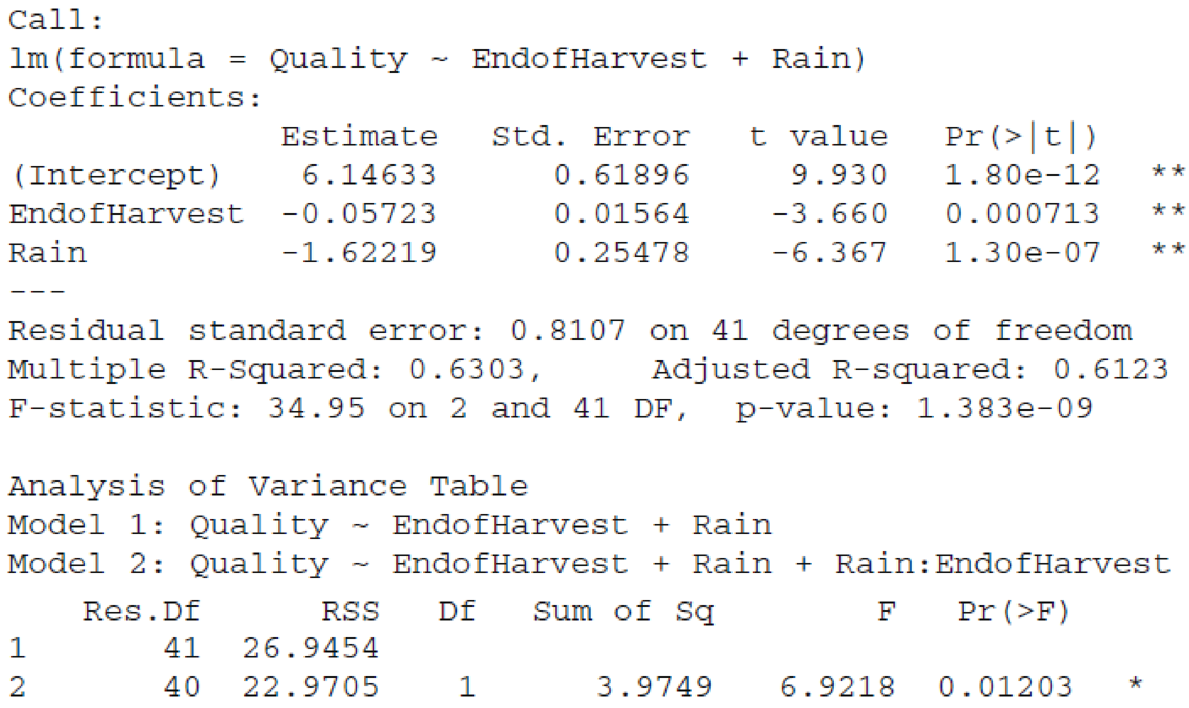


Figure 5.8 A scatter plot of Quality versus End of Harvest for Chateau Latour



```
Call:
lm(formula = Quality ~ EndofHarvest + Rain)
Coefficients:
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)      6.14633      0.61896     9.930   1.80e-12   ***
EndofHarvest    -0.05723      0.01564    -3.660   0.000713   ***
Rain            -1.62219      0.25478    -6.367   1.30e-07   ***
---
Residual standard error: 0.8107 on 41 degrees of freedom
Multiple R-Squared: 0.6303,     Adjusted R-squared: 0.6123
F-statistic: 34.95 on 2 and 41 DF,  p-value: 1.383e-09

Analysis of Variance Table
Model 1: Quality ~ EndofHarvest + Rain
Model 2: Quality ~ EndofHarvest + Rain + Rain:EndofHarvest
    Res.Df      RSS   Df   Sum of Sq        F    Pr(>F)
1       41  26.9454
2       40  22.9705    1       3.9749   6.9218  0.01203   *
```

#(a) Show that the coefficient of the interaction term in model (5.10) is statistically significant. In other words, show that the rate of change in quality rating depends on whether there has been any unwanted rain at vintage.

We can show that the coefficient of interaction term in model is statistically significant in following two ways:

1) Run full model with interaction term as in model (5.10) as $Quality = \beta_0 + \beta_1 EndofHarvest + \beta_2 Rain + \beta_3 EndofHarvest * Rain$

We can see that p value of interaction term is 0.0120 and is shown significant.

2) Fit a reduced model as $Quality = \beta_0 + \beta_1 EndofHarvest + \beta_2 Rain$

If we run the analysis of variance to see if two models are significantly different. From ANOVA table, we observe the p-value to be 0.0203. This implies that the two models are significantly different. Hence, we can conclude that the coefficient of the interaction term in the full model is statistically significant.

#(b) Estimate the number of days of delay to the end of harvest it takes to decrease the quality rating by 1 point when there is: We know, change in quality/change in end of harvest= slope.

Since quality is decreased by 1, therefore,

$Numberofdaysdelayed = -1|slope$ #(i)No unwanted rain at harvest If there is no unwanted rain at harvest, the value of Rain is 0. And, then using the definition of slope that we have defined, no of days delayed = -1/(-0.03145) =31.48

Therefore, it requires approximately 32 days to decrease the quality rating by 1 point when there is no unwanted rain at harvest.

#(ii)Some unwanted rain at harvest IF there is some unwanted rain at harvest, the value of rain is 1. The slope is additive here. From our definition earlier, number of days required= -1/(-0.03145-0.08314)=8.72.

Therefore, it requires approximately 8.7 days to decrease the quality rating by 1 point when there is some unwanted rain at harvest.

Project milestones [20 points]

1. Perform an exploratory data analysis

- Numerically summarize the variables.
- Make plots and explore relationships between variables.
- Identify any strange points or anything else that doesn't make sense.

2. Begin to think about how to model the relationships in your data.

Title: Nitrogen Rate, Nitrogen Source Effects on Camelina Genotypes in Semiarid Region of Northern Nevada Irrigated with Reclaimed Waste Water.

We begin by performing exploratory analysis on our data. This would be helpful to form some prelimiary notion of relationship between variables. That way we can have idea about which variable contributes more to the y-output, and which variables doens't contribute much. This is very essential to fine tuning the final regression relationship.

We first start with finding the structure of our data, minimum, maximum, mean value of different variables. This will help us to find if the scale of different variables are comparable. Then we perform correlation analysis and plot scatterplot. This will help us identify the relationships between different variables in pairs.

Then we plot boxplot for certain numeric variables to have and idea of how they look, what is the range and the likes, figuratively.

We then plot the yield with various independent variables to have an idea of different cases.

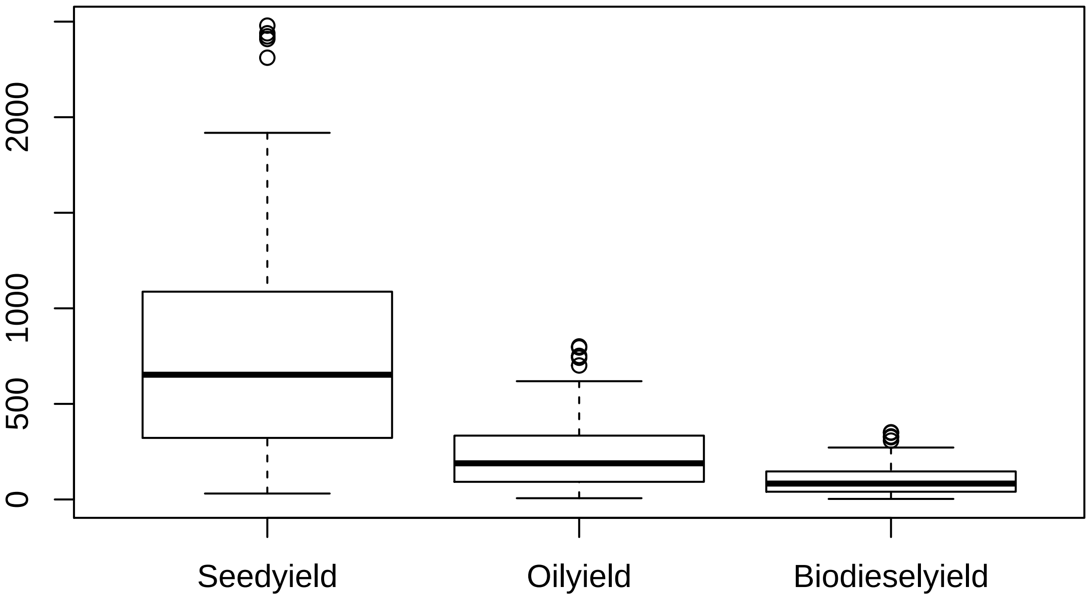
```
## Loading required package: ggplot2

## Loading required package: GGally

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'GGally'

## Loading required package: CCA

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'CCA'

##       Plot          Treatment         Block            Cultivar  NSource
##  Min.   :  1.00   Min.   : 1.00   Min.   :1.00   BlaineCreek:64   N1:64
##  1st Qu.: 32.75   1st Qu.: 4.75   1st Qu.:1.75   Pronghorn  :64   N2:64
##  Median : 64.50   Median : 8.50   Median :2.50
##  Mean   : 64.50   Mean   : 8.50   Mean   :2.50
##  3rd Qu.: 96.25   3rd Qu.:12.25   3rd Qu.:3.25
##  Max.   :128.00   Max.   :16.00   Max.   :4.00
##   NRate         Year        Seedyield          Oilyield
##  N00 :32   Min.   :2016   Min.   :  31.06   Min.   :  6.54
##  N120:32   1st Qu.:2016   1st Qu.: 326.17   1st Qu.: 92.62
##  N40 :32   Median :2016   Median : 652.72   Median :189.28
##  N80 :32   Mean   :2016   Mean   : 767.50   Mean   :235.10
##            3rd Qu.:2017   3rd Qu.:1083.99   3rd Qu.:333.39
##            Max.   :2017   Max.   :2479.09   Max.   :800.75
##  Biodieselyield
##  Min.   :  2.87
##  1st Qu.: 40.66
##  Median : 83.09
##  Mean   :103.21
##  3rd Qu.:146.36
##  Max.   :351.53
```

We now see the general structure of our data. For example: Seedyield has min value of 31.06, maximum of 2479.09; the variability is really high. The 1st quartile is 326.17, median is 652.72, mean is 757.5 and 3rd quartile is 1083.99.

Similarly we can observe such statistics for Oilyield and BiodieselYield.

Year, NSource, Block, Cultivar, Nrate, Treatment are all discrete variables with certain levels.

```
newData<-NS[,c(2,3,8,9,10)]
boxplot(newData[,3:5])
```

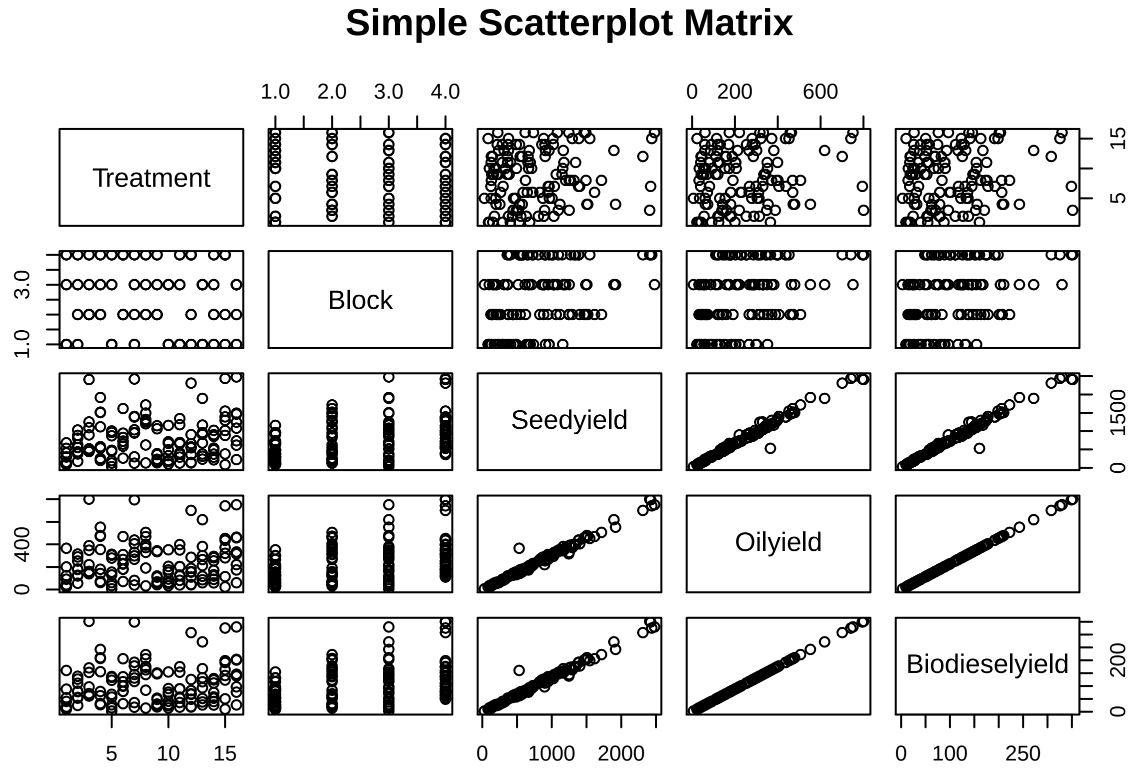


We can observe the box plot for the three different yield in the above figure. This will give us an idea of what the yield data looks like. In general, Seedyield is greater than Oilyeild is greater than BiodieselYield, which is correct to our notion.

```
newData<-NS[,c(2,3,8,9,10)]

cor(newData)
```

```
##                   Treatment       Block Seedyield  Oilyield Biodieselyield
## Treatment        1.00000000 -0.04850713 0.1541814 0.1251508      0.1251514
## Block           -0.04850713  1.00000000 0.4003525 0.4170700      0.4170753
## Seedyield        0.15418141  0.40035249 1.0000000 0.9903622      0.9903618
## Oilyield         0.12515080  0.41706997 0.9903622 1.0000000      1.0000000
## Biodieselyield   0.12515140  0.41707530 0.9903618 1.0000000      1.0000000
```

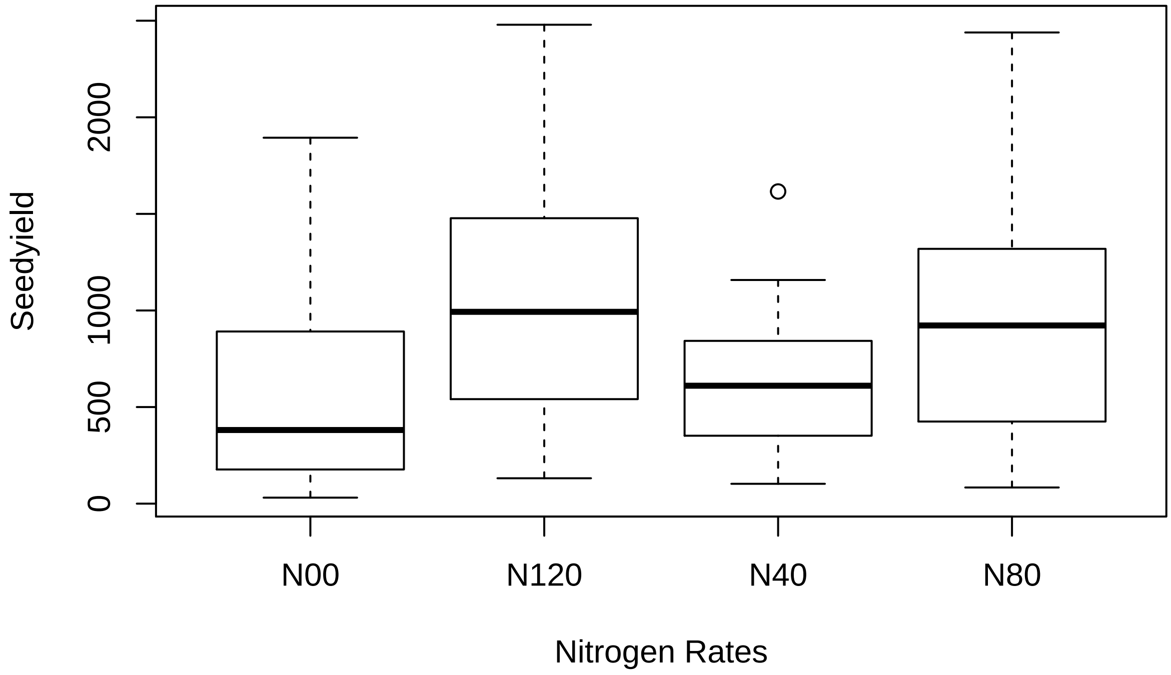
```
#pairs(newData[,1:5])
pairs(~Treatment+Block+Seedyield+Oilyield+Biodieselyield, data=newData,main="Simple Scatterplot Matrix"
```



From the corelation table, we see slightly negative corelation between treatment and block (but since this is in our control, we don't worry about it). Seedyield, Oilyield and Biodieselyield are highly correlated with each other. Treatment has positive corelation with all three kinds of yields. Block also has higher corelation with three different kinds of yields.

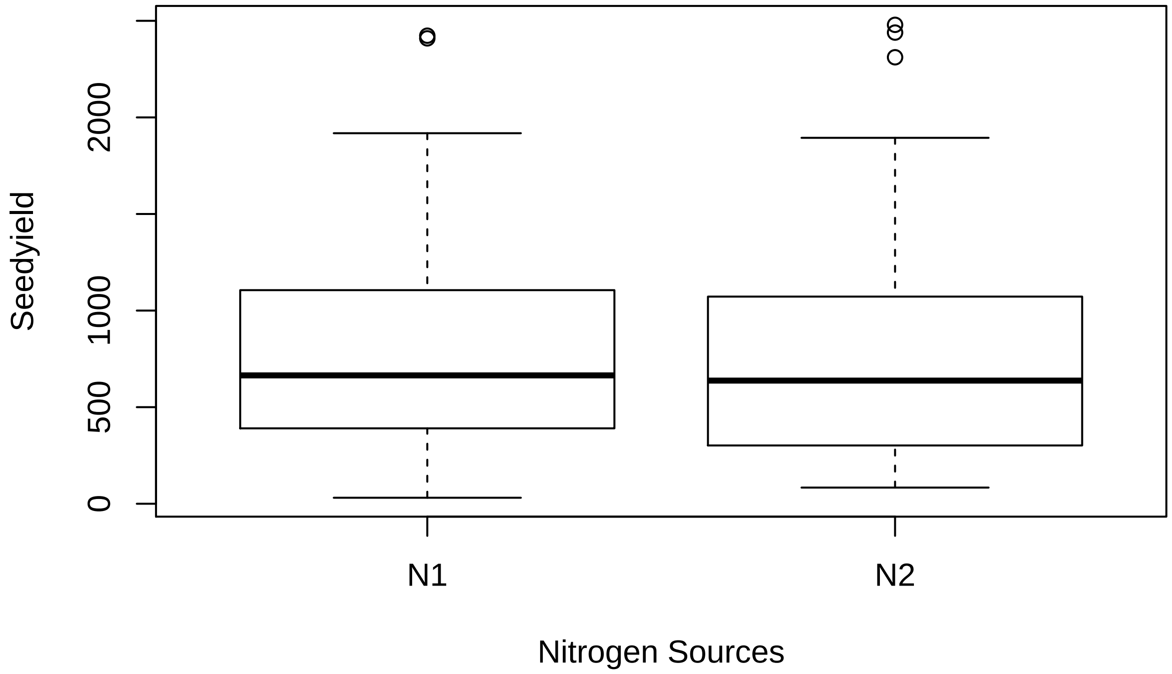
From the scatterplot above, we can see that there is strong relationship between the three yields, Seedyield, Oilyield and Biodieselyeild which re-inforces our notion that yeild are related to each other positively. If we compare treatment with the yield, we cannot see any definite pattern outright and that it may need to undergo some transformation. Comparing the yield for different blocks, we can see the average yeild for different blocks are roughly same. But, there is some variation in block 1 which has less yield most of the time, and block 4 has more higher yield compared to others. We might generate some findings that different blocks of treatment were able to generate different yield. Because other input variables were non-numeric, scatterplots couldn't be plotted. We should be able to convert them to some numeric units and then plot them to identify the relationships.

```
par(mfrow=c(1,1))
plot(NRate,Seedyield,ylab="Seedyield",xlab="Nitrogen Rates")
```



Comparing nitrogen rates with seedyield in the above BoxPlot, we can see that there is definitely some relationship between them. N120 and N80 cause more seedyeild compared to the counterparts N00 and N40.

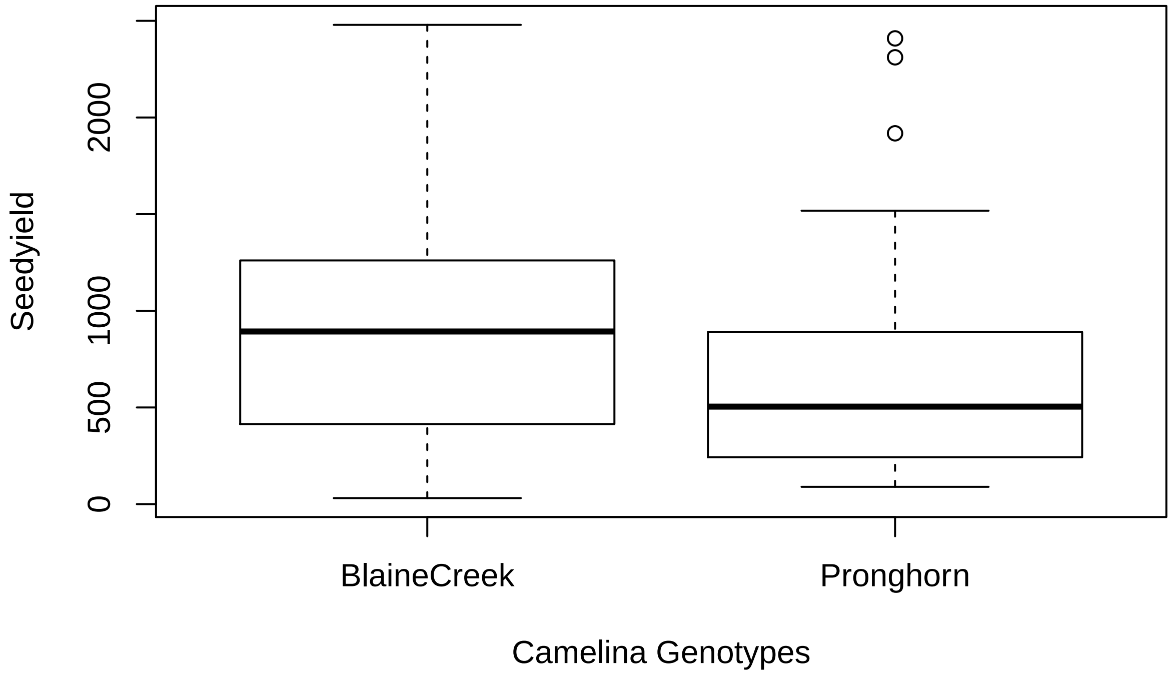
```
par(mfrow=c(1,1))
plot(NSource,Seedyield,ylab="Seedyield",xlab="Nitrogen Sources")
```



```
#plot(Cultivar,Seedyield,ylab="Seedyield",xlab="Camelina Genotypes")
```

It seems from above figure that the effect of Nitrogen Sources are almost similar on SeedYield. There are some outliers in both sources. So, we might not include nitrogen sources in the final relationship.

```
par(mfrow=c(1,1))
plot(Cultivar,Seedyield,ylab="Seedyield",xlab="Camelina Genotypes")
```



It seems that the genotypes also contribute some amount on the amount of seedyield. BlainCreek has comparatively more seedyield than Pronghorn. This should be used in the formula.

```
mfull<-lm(formula=Seedyield~NRate+NSource+Cultivar+NRate:NSource+NRate:Cultivar+NSource:Cultivar+NRate:
summary(mfull)
```

```
##
## Call:
## lm(formula = Seedyield ~ NRate + NSource + Cultivar + NRate:NSource +
##     NRate:Cultivar + NSource:Cultivar + NRate:NSource:Cultivar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1045.33  -322.90   -64.17   271.64  1607.45
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               515.06     186.23   2.766  0.00664
## NRateN120                                 610.23     263.37   2.317  0.02232
## NRateN40                                  276.70     263.37   1.051  0.29570
## NRateN80                                  557.57     263.37   2.117  0.03647
## NSourceN2                                 301.22     263.37   1.144  0.25519
## CultivarPronghorn                        -221.25     263.37  -0.840  0.40265
## NRateN120:NSourceN2                      -259.54     372.46  -0.697  0.48737
## NRateN40:NSourceN2                       -507.41     372.46  -1.362  0.17584
## NRateN80:NSourceN2                       -244.89     372.46  -0.657  0.51221
## NRateN120:CultivarPronghorn                48.78     372.46   0.131  0.89604
## NRateN40:CultivarPronghorn                 41.48     372.46   0.111  0.91153
## NRateN80:CultivarPronghorn                 92.21     372.46   0.248  0.80493
## NSourceN2:CultivarPronghorn              -133.70     372.46  -0.359  0.72029
## NRateN120:NSourceN2:CultivarPronghorn  -157.02     526.74  -0.298  0.76618
## NRateN40:NSourceN2:CultivarPronghorn    206.06     526.74   0.391  0.69640
## NRateN80:NSourceN2:CultivarPronghorn   -234.19     526.74  -0.445  0.65746
##
## (Intercept)                            **
## NRateN120                              *
## NRateN40
## NRateN80                               *
## NSourceN2
## CultivarPronghorn
## NRateN120:NSourceN2
## NRateN40:NSourceN2
## NRateN80:NSourceN2
## NRateN120:CultivarPronghorn
## NRateN40:CultivarPronghorn
## NRateN80:CultivarPronghorn
## NSourceN2:CultivarPronghorn
## NRateN120:NSourceN2:CultivarPronghorn
## NRateN40:NSourceN2:CultivarPronghorn
## NRateN80:NSourceN2:CultivarPronghorn
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 526.7 on 112 degrees of freedom
## Multiple R-squared:  0.2246, Adjusted R-squared:  0.1207
## F-statistic: 2.162 on 15 and 112 DF,  p-value: 0.01162
```

```r
mreduced<-lm(formula=Seedyield~NRate+NSource+Cultivar)
summary(mreduced)
```

```
##
## Call:
## lm(formula = Seedyield ~ NRate + NSource + Cultivar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1009.49  -322.64   -65.59   279.39  1576.85
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         675.31     111.84   6.038 1.73e-08 ***
## NRateN120           465.60     129.14   3.605 0.000453 ***
## NRateN40             95.25     129.14   0.738 0.462198
## NRateN80            422.68     129.14   3.273 0.001385 **
## NSourceN2           -41.74      91.32  -0.457 0.648439
## CultivarPronghorn  -265.64      91.32  -2.909 0.004311 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 516.6 on 122 degrees of freedom
## Multiple R-squared:  0.1876, Adjusted R-squared:  0.1543
## F-statistic: 5.636 on 5 and 122 DF,  p-value: 0.000104
```

```
anova(mreduced,mfull)
```

```
## Analysis of Variance Table
##
## Model 1: Seedyield ~ NRate + NSource + Cultivar
## Model 2: Seedyield ~ NRate + NSource + Cultivar + NRate:NSource + NRate:Cultivar +
##     NSource:Cultivar + NRate:NSource:Cultivar
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    122 32555364
## 2    112 31075271 10   1480094 0.5334 0.8633
```

```
detach(NS)
```

#(i)No unwanted rain at harvest