Assignment4

Lorenzo Ricciardulli

2024-02-07

Loading data and summary of multiple regression model

```
data = read.csv("yield.txt", header = TRUE)
ols = lm(data$Yield_kg_per_hectare ~
data$Rainfall_mm +
data$Sunny_Days +
data$Irrigation_Schedule +
data$Seed_Variety +
data$Fertilizer_Amount_kg_per_hectare +
data$Soil_Quality, data = data)
summary(ols)
```

```
##
## Call:
## lm(formula = data$Yield_kg_per_hectare ~ data$Rainfall_mm + data$Sunny_Days +
##     data$Irrigation_Schedule + data$Seed_Variety + data$Fertilizer_Amount_kg_per_hectare +
##     data$Soil_Quality, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -222.129  -33.848    0.256   33.572  183.481
##
## Coefficients:
##                                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                          48.334835   5.110428    9.458   <2e-16
## data$Rainfall_mm                     -0.505536   0.003956 -127.796   <2e-16
## data$Sunny_Days                       1.992305   0.039658   50.237   <2e-16
## data$Irrigation_Schedule             49.986721   0.177531  281.567   <2e-16
## data$Seed_Variety                   300.463701   0.865392  347.200   <2e-16
## data$Fertilizer_Amount_kg_per_hectare   0.808804   0.005500  147.057   <2e-16
## data$Soil_Quality                     1.543654   0.027243   56.662   <2e-16
##
## (Intercept)                         ***
## data$Rainfall_mm                    ***
## data$Sunny_Days                     ***
## data$Irrigation_Schedule            ***
## data$Seed_Variety                   ***
## data$Fertilizer_Amount_kg_per_hectare ***
## data$Soil_Quality                   ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.06 on 15993 degrees of freedom
## Multiple R-squared:  0.9387, Adjusted R-squared:  0.9387
## F-statistic: 4.084e+04 on 6 and 15993 DF,  p-value: < 2.2e-16
```

1. Consider your multiple linear regression model fitted in the previous assignment and make sure to include at least three predictors. Formulate a joint null hypothesis to test whether the effect of two of such variables is the same and the remaining one has a negligible effect on the response. Formulate this hypothesis in the linear form A = b, clearly reporting the definition on A and b. Finally, compute the p-value of the test and comment the results.

```
mod1 = lm(data$Yield_kg_per_hectare ~
data$Rainfall_mm +
data$Sunny_Days +
data$Fertilizer_Amount_kg_per_hectare +
data$Soil_Quality, data = data)
A = matrix(0, nrow = 2, ncol = 5)
A[1,2]=A[1,4] = 1
A[2,3]= 1
A[2,5] = -1
b = c(0,0)
linearHypothesis(mod1,A,b)
```

```
## Linear hypothesis test
##
## Hypothesis:
## data$Rainfall_mm  + data$Fertilizer_Amount_kg_per_hectare = 0
## data$Sunny_Days - data$Soil_Quality = 0
##
## Model 1: restricted model
## Model 2: data$Yield_kg_per_hectare ~ data$Rainfall_mm + data$Sunny_Days +
##     data$Fertilizer_Amount_kg_per_hectare + data$Soil_Quality
##
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1  15997 549300589
## 2  15995 544464394  2   4836195 71.038 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
A
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    0    1    0    1    0
## [2,]    0    0    1    0   -1
```

The matrix A comprehends all the values of the coefficients of variables in the model. The vector b is the response vector of the linear system Ax = b. If the aim is to test the joint hypothesis: coeff.var1, coeff.var3 = 0 AND coeff.var2 = coeff.var4, the matrix A will comprehend 2 rows (as we want to test a joint hypothesis of two single hypothesis) and 5 columns (4 variables plus the intercept). The first row null hypothesis that states that the coefficients of var1 and var3 are equal to 0 is expressed putting 1 in both positions of variables in the model. A similar rationale is valid for the other hypothesis, the second row one. The response vector b should comprehend only 0 values and should be of dimension 2 rows and one column.

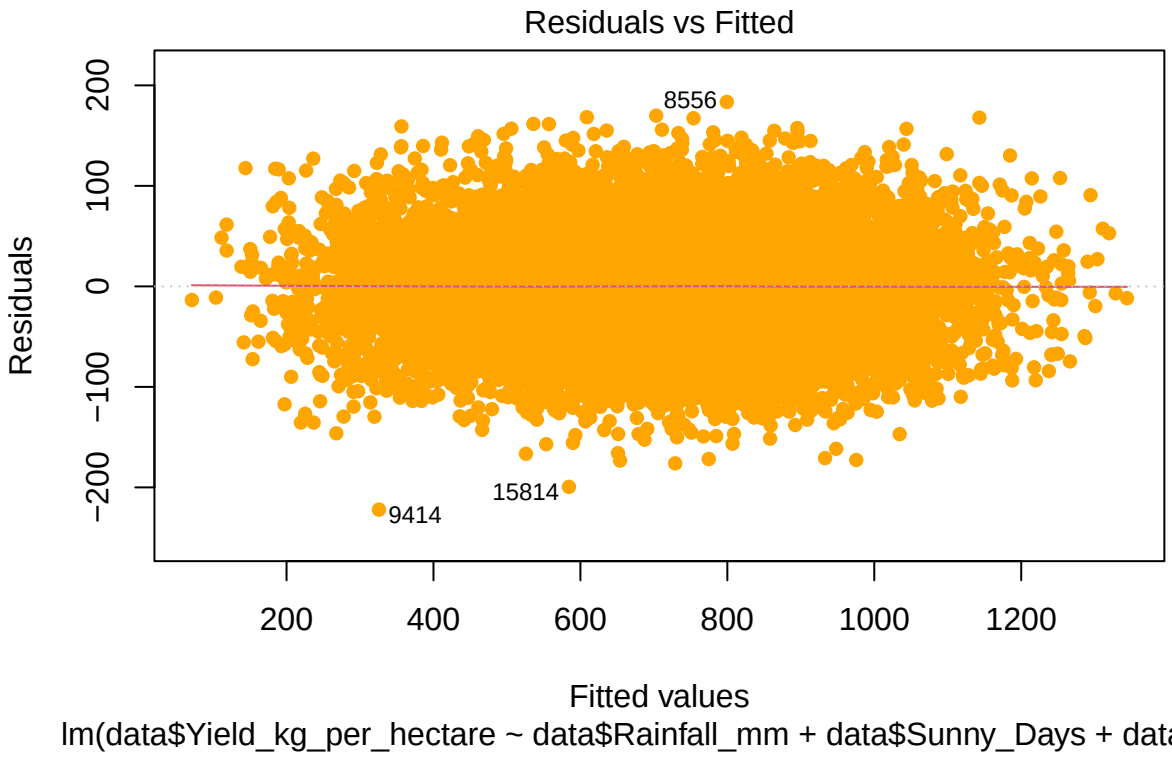
The F-statistic is high and the p-value is very low (close to zero), indicating there is a substantial difference in the explained variation between the two models and that this difference is statistically significant. Consequently, we can reject the null hypothesis that the restrictive model is sufficient and accept the alternative hypothesis that the full model is significantly better at explaining the variation in agricultural yield data. In summary, there is statistical evidence that the additional predictors in the full model (Model 2) have a significant impact on the variation in agricultural yield compared to the more restrictive model (with the joint hypothesis).

2. Given the multiple linear model fitted in the previous assignment, perform regression diagnostics to:

(a) Check the constant variance assumption for the errors.
(b) Check the structure of the relationship between the predictors and the response.
(c) Check the normality assumption.
(d) Check for large leverage points.
(e) Check for outliers.
(f) Check for inuential points.

2.a. Check the constant variance assumption for the errors.

```
plot(ols, which = 1, pch =16, col = "orange")
```



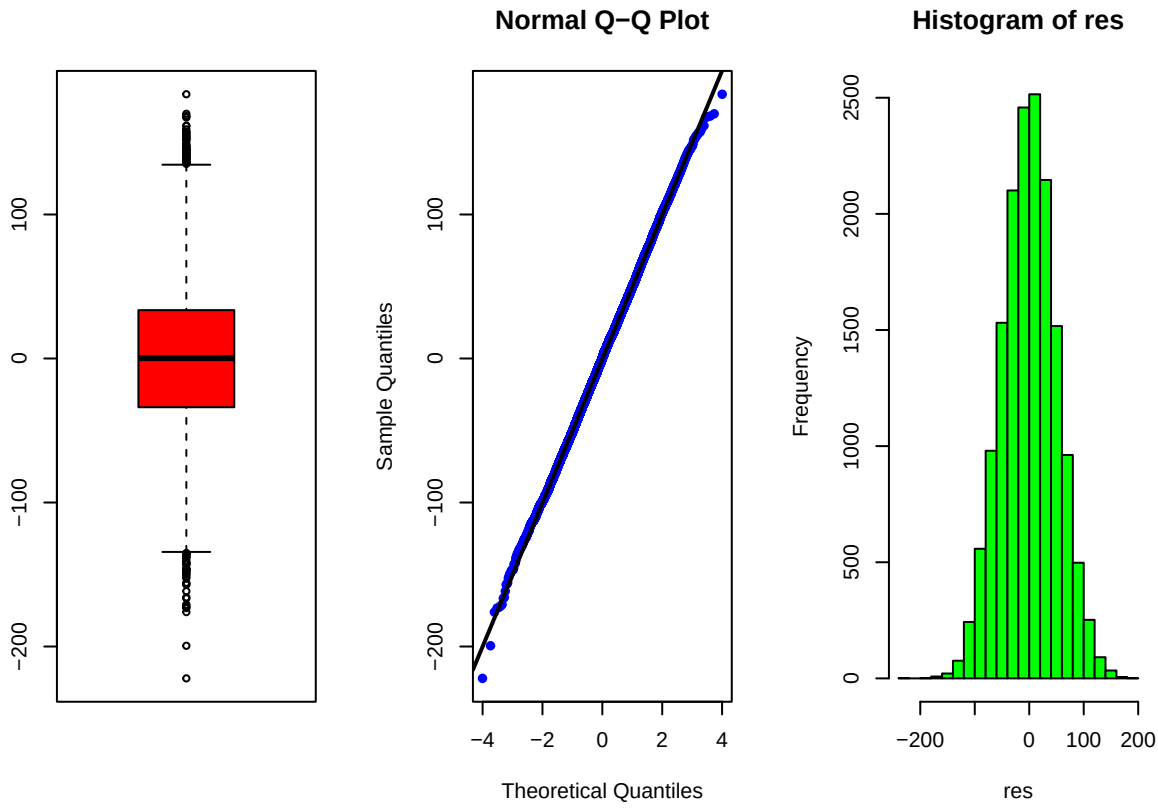
Residuals vs. Fitted Values Plot: This plot helps to assess whether the spread of residuals changes as the predicted values (fitted values) change. There is clearly a random scatter of points with no discernible pattern. If it can noticed any specific pattern (e.g., funnel shape, cone shape), it suggests heteroscedasticity, which violates the constant variance assumption, but it is not the case.

2.b. Check the structure of the relationship between the predictors and the response.

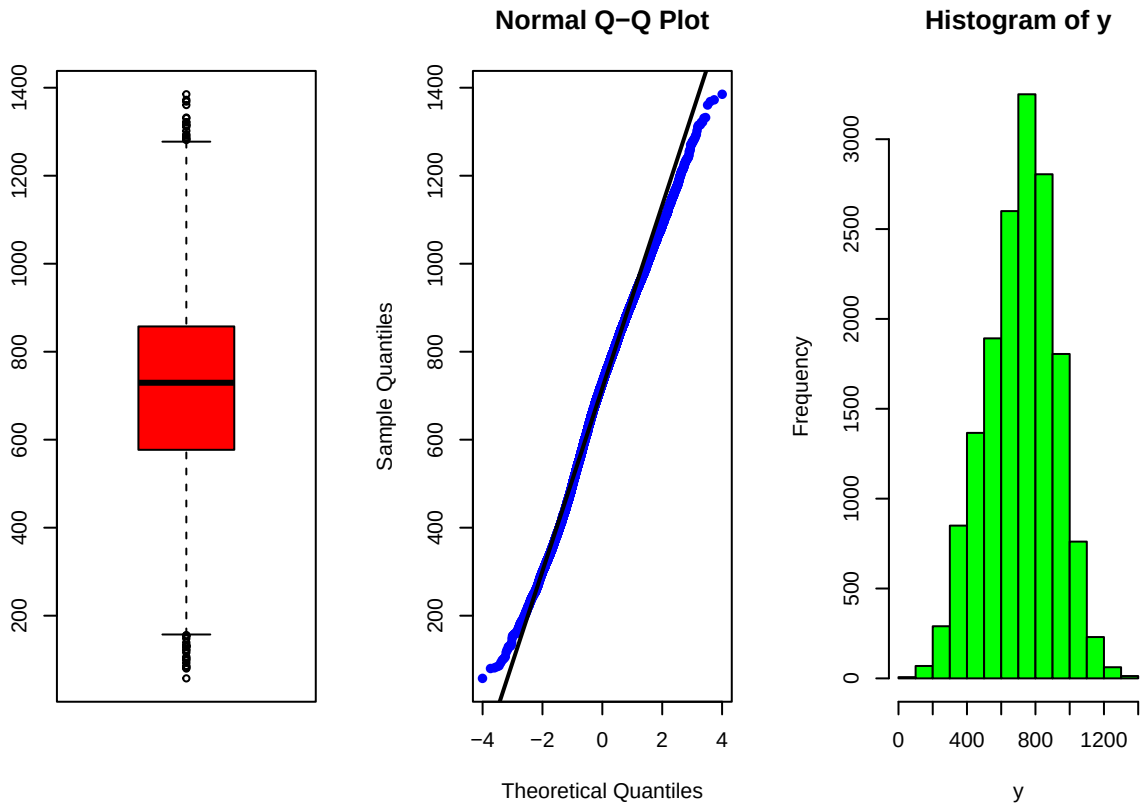
As it can be seen from the plot above (of the residuals vs fitted values), the linearity assumption seems to be respected: the values seems to be dispersed randomly and no relevant pattern can be stressed.

2.c. Check the normality assumption

```
res = residuals(ols)
par(mfrow = c(1,3))
boxplot(res,col = "red")
qqnorm(res, pch = 16, col = "blue")
qqline(res, lwd = 2, col = "black")
hist(res,col = "green")
```

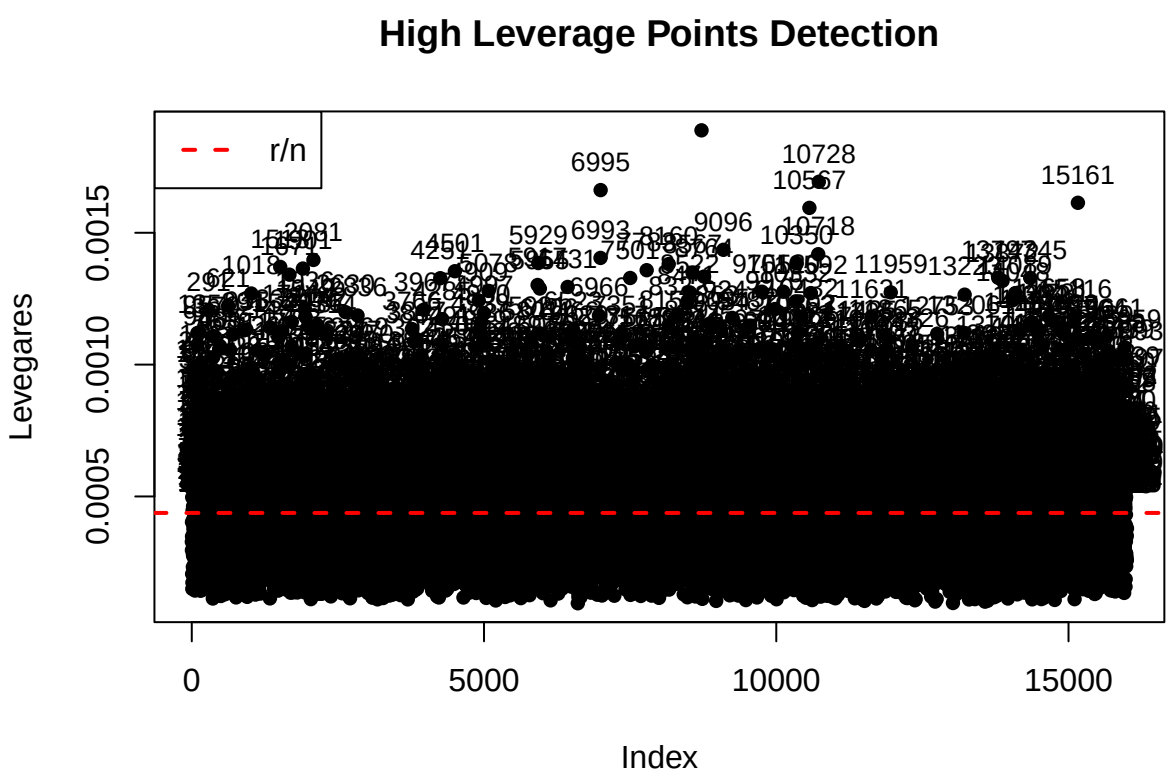



```
y = data$Yield_kg_per_hectare
par(mfrow = c(1,3))
boxplot(y,col = "red")
qqnorm(y, pch = 16, col = "blue")
qqline(y, lwd = 2, col = "black")
hist(y,col = "green")
```

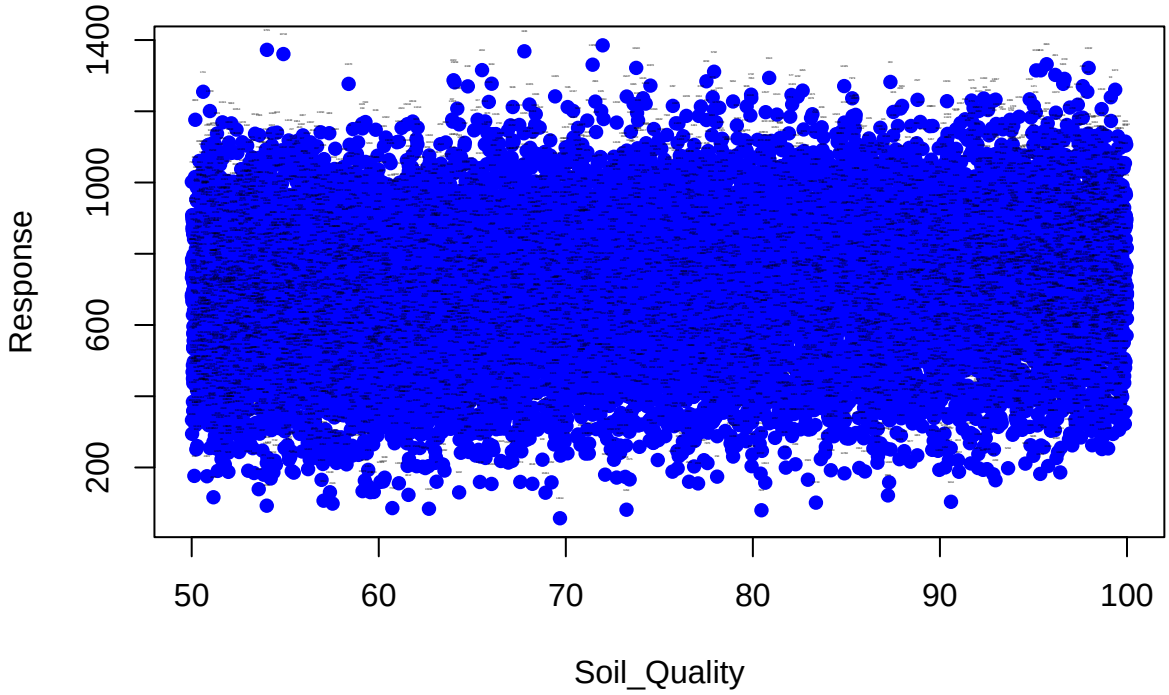


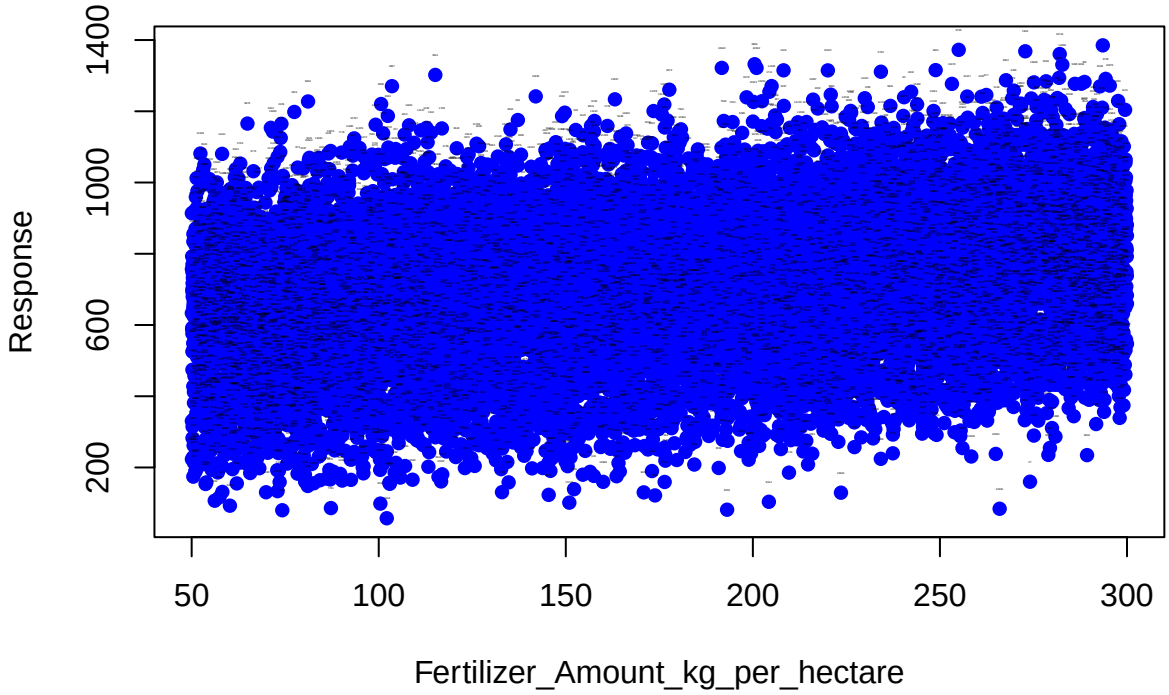
2.d. Check for large leverage points.

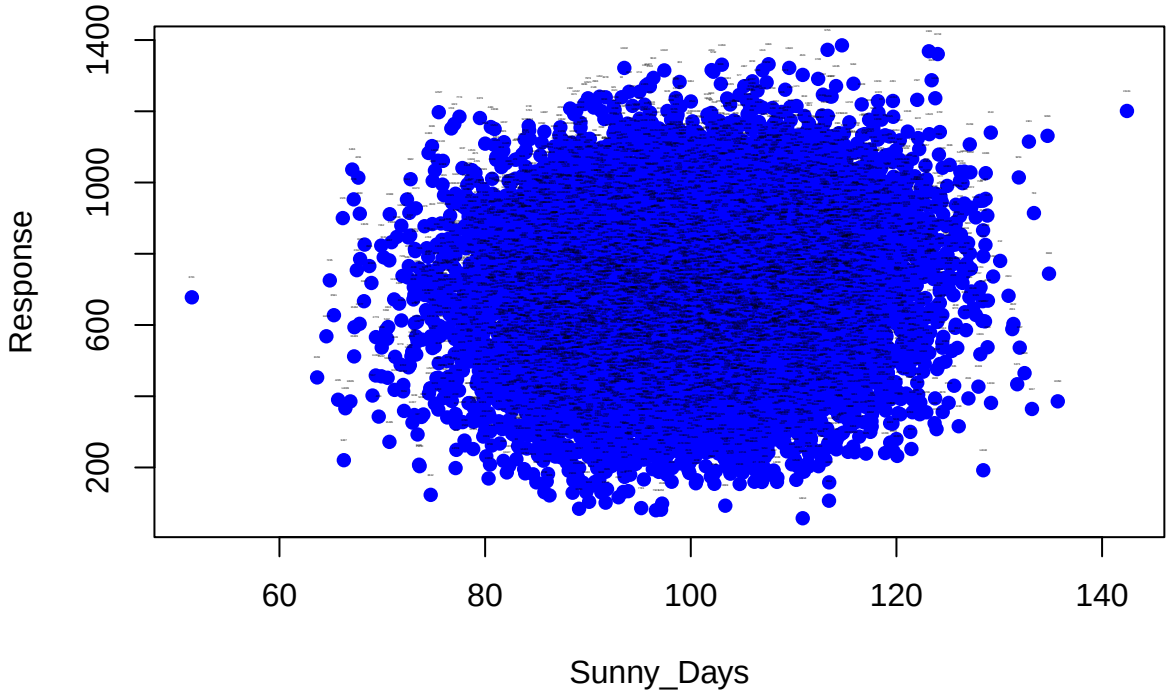
```
n = 16000
 r = length(coef(ols))
 h_values = hatvalues(ols)
 high_leverages_indexes = which( h_values > (r/n) )
 plot( h_values, pch = 16, ylab = "Levegares",
 main = "High Leverage Points Detection")
 abline(h = r/n, col = 'red', lty = 2, lwd = 2)
 legend(x = "topleft", legend = "r/n", lty = 2, lwd = 2, col = 'red' )
 text(high_leverages_indexes, h_values[high_leverages_indexes],
 labels = high_leverages_indexes, cex = 0.8, pos = 3)
```

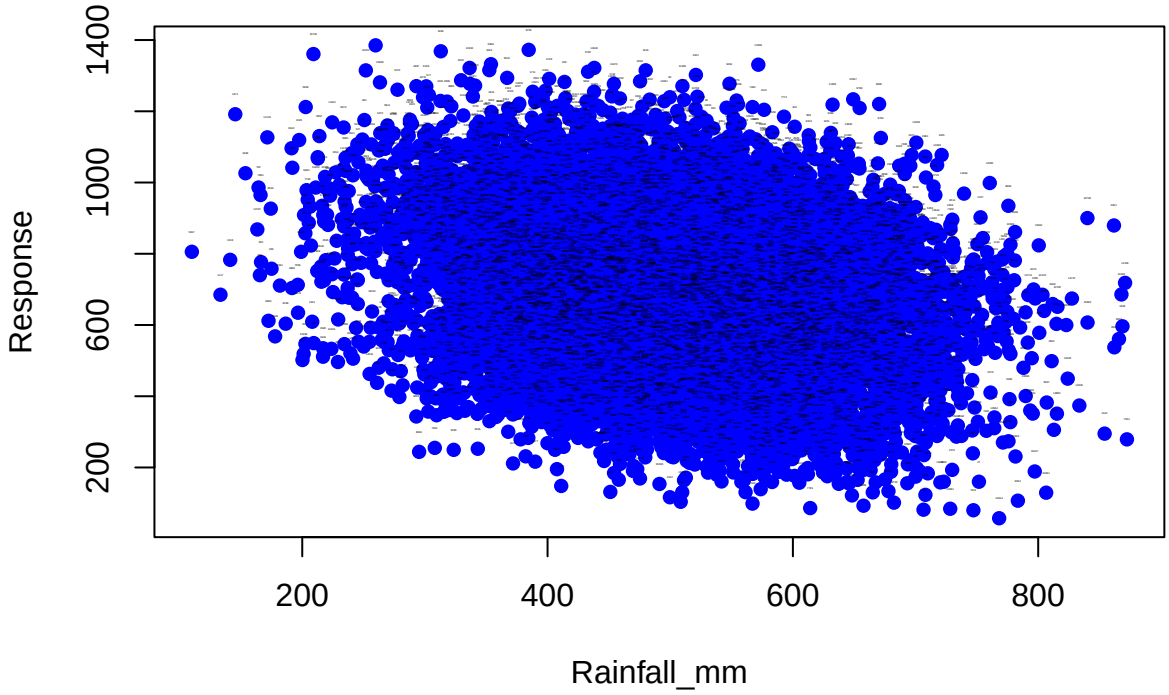


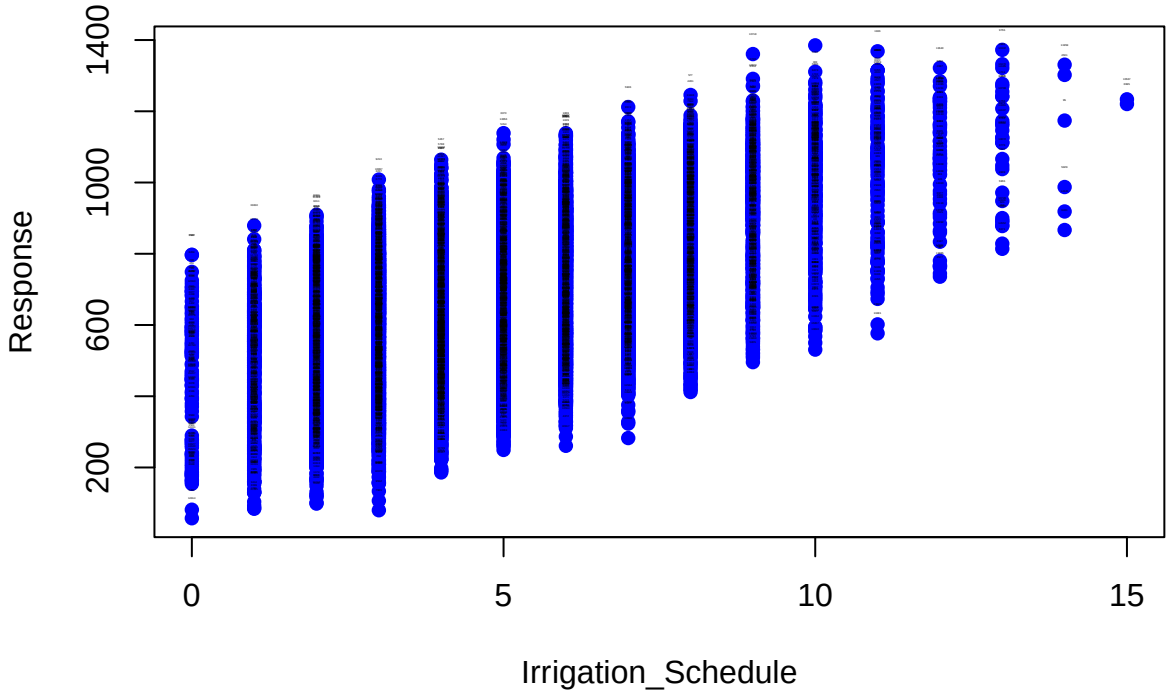

```
for(i in c(1,3,4,5,6)){
 plot(data[,i],data[,7],pch=16,xlab=colnames(data)[i],ylab="Response", col ="blue")
 text(data[high_leverages_indexes,i],data[high_leverages_indexes,7],
 labels=high_leverages_indexes,cex=0.1,pos=3)
 }
```


Response
200
600
1000
1400
50
60
70
80
90
100
Soil_Quality


Response
200
600
1000
1400
50
100
150
200
250
300
Fertilizer_Amount_kg_per_hectare


Response
200
600
1000
1400
60
80
100
120
140
Sunny_Days


Response
1400
1000
600
200
200
400
600
800
Rainfall_mm



2.e. Check for outliers.

```
i=1
studentized_res= rstandard(ols)
res_stud2=res/sqrt( sum(res^2)/(ols$df.residual) )
alpha=0.05
threshold=qt(1-alpha/2,df=ols$df.residual)
abs(studentized_res[i])
```

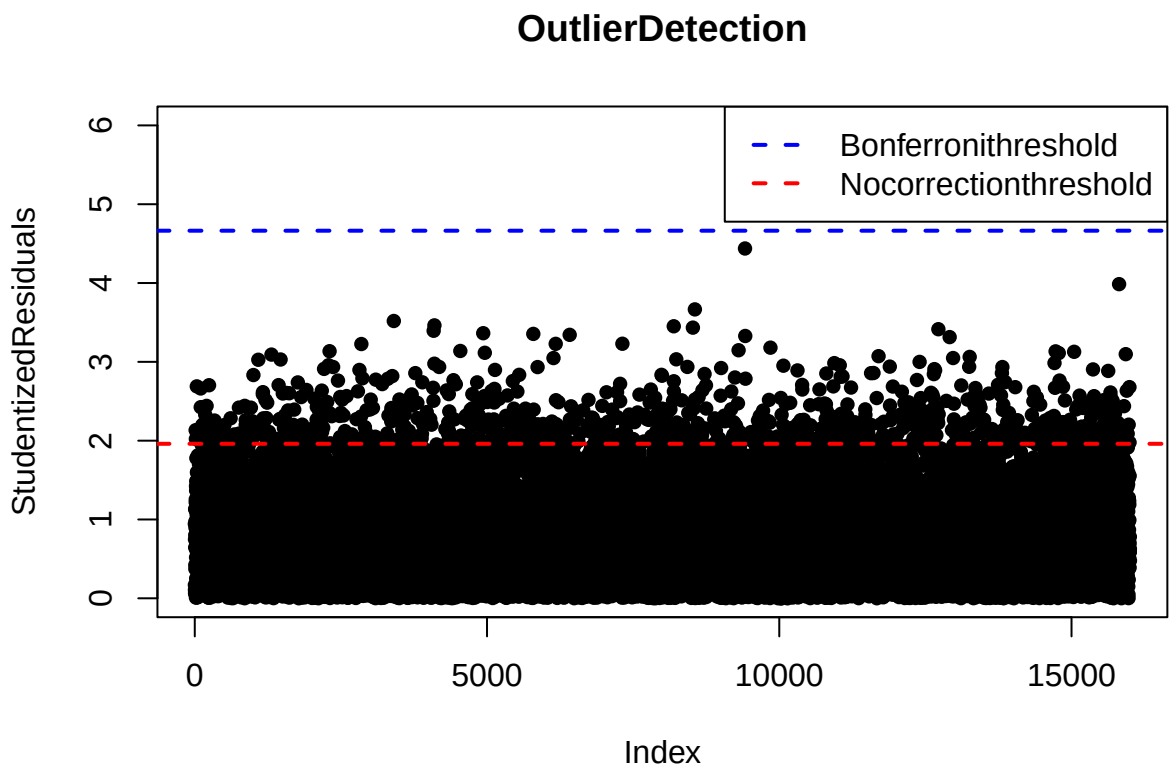
```
##         1
## 0.9446379
```

```
abs(studentized_res[i]) < threshold
```

```
##    1
## TRUE
```

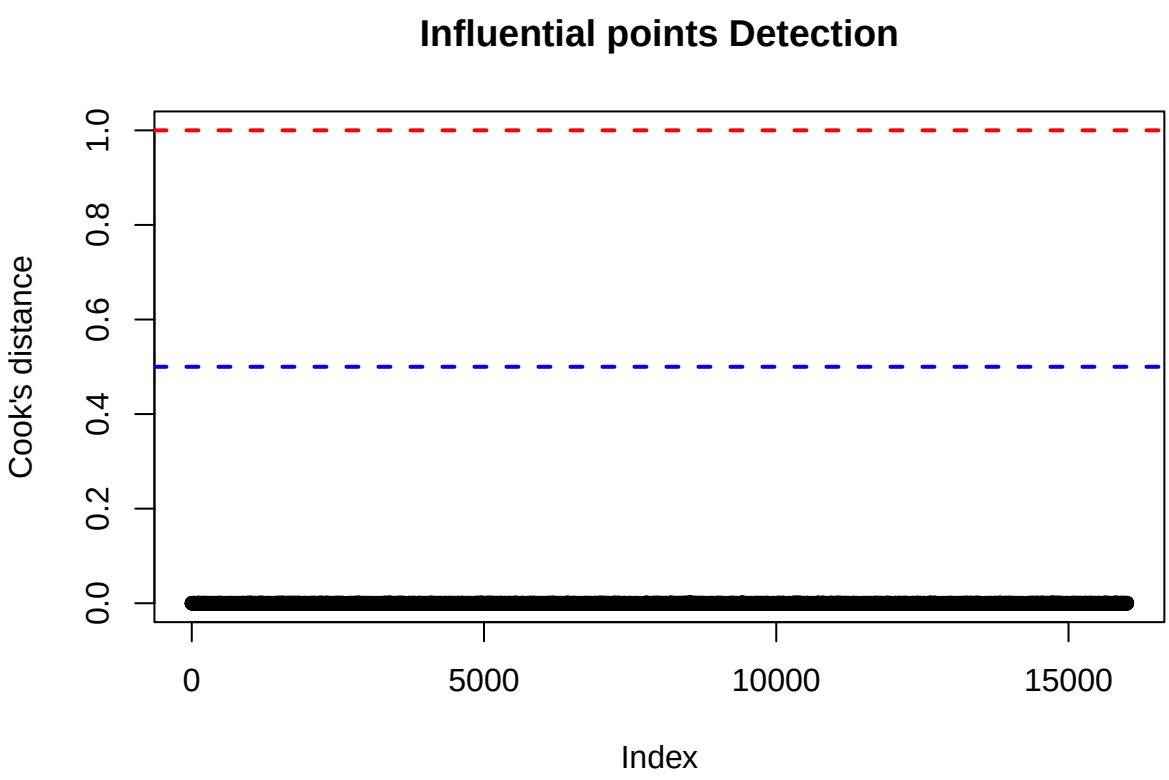
```
alpha=0.05
bonferroni_quantile = qt(1-alpha/(2*n),df=ols$df.residual)
plot( abs(studentized_res),pch=16, ylim= c(0,6),
main="OutlierDetection",ylab="StudentizedResiduals")
abline(h=threshold,
```

```
col='red',lty=2,lwd=2)
abline(h=bonferroni_quantile,
col='blue',lty=2,lwd=2)
legend("topright",
legend=c("Bonferronithreshold",
 "Nocorrectionthreshold"),
lty=c(2,2),lwd= c(2,2),col= c('blue','red'))
```



2.f. Check for influential points.

```
cook = cooks.distance(ols)
plot( cook, pch = 16, ylim = c(0,1), ylab = "Cook's distance",
main = "Influential points Detection")
abline(h = 1, col = 'red', lty = 2, lwd = 2)
abline(h = 0.5, col = 'blue', lty = 2, lwd = 2)
```



3. Using the findings of your diagnostics analysis, improve your model.

The results of the diagnostics analysis are the following:

-From the points 2.a and 2.b, it can be concluded that the plot of residuals vs fitted values does not show any discernible pattern. Therefore, the constant variance assumption and the linearity assumption seems to be respected.

-In the point 2.c, the normality assumption was checked. Since the boxplots are centered, the histograms resemble a uniform distribution and the values of QQ plots are very close to the abline represented in the plots, it could be said that the normality assumption seems to be valid in this context. It should be said that form the plots of the response variable, in particular in the QQ plot, a slight short tailed distribution could be noticed, but this fact does not invalidate the normality assumption.

-In the point 2.d an analysis to find high leverage points was performed. As it can be seen in the plot, a great number of high leverage points was found. The solution could imply the record of more observations close to the leverage points so that they are no longer far in the regressor space. It is useful to identify them because this may have been wrongly reported.

-In the point 2.e the objective was to identify outliers in the dataset. No outliers were found in the analysis.

-In the final point 2.f, the aim was to find influential points (points whose removal causes major changes in the analysis). No such kind of points was found in the analysis.

Eventually, it can be concluded that the diagnostics analysis performed in the previous point showed that no changes in the dataset are needed.