Assignment5

Lorenzo Ricciardulli

2024-02-15

Loading data and summary of ols 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)
```

(a) Consider all the available covariates to explain your response and assume a linear regression model. Perform a best subset selection to explore all the possible models and comment the results.

```
library(leaps)
```

```
## Warning: il pacchetto 'leaps' è stato creato con R versione 4.1.3
```

```
model1 = regsubsets (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(model1)
```

```
## Subset selection object
## Call: regsubsets.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)
## 6 Variables  (and intercept)
##                                       Forced in Forced out
```

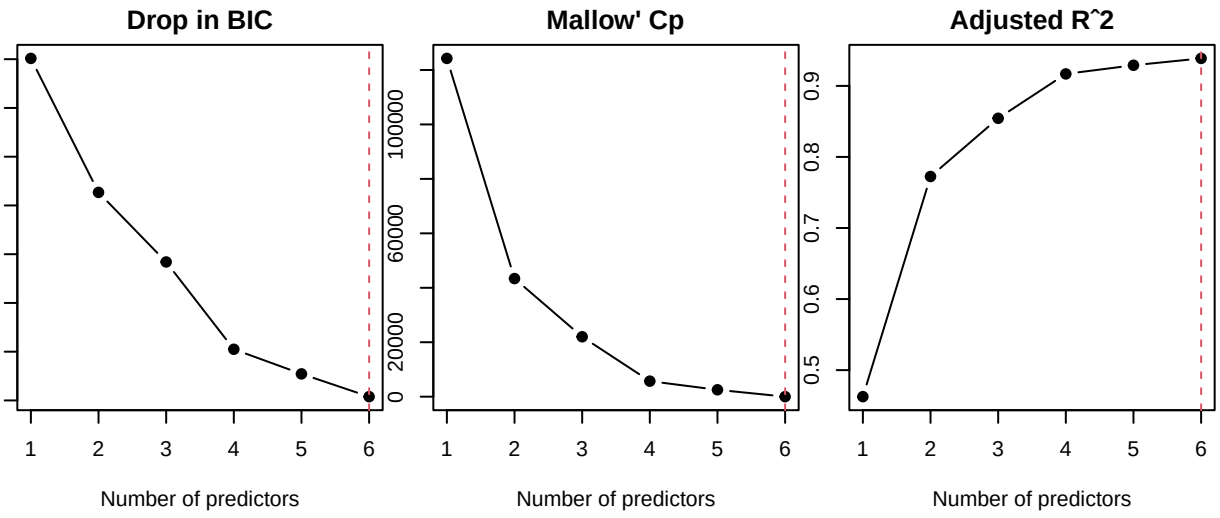
```
## data$Rainfall_mm                           FALSE      FALSE
## data$Sunny_Days                            FALSE      FALSE
## data$Irrigation_Schedule                   FALSE      FALSE
## data$Seed_Variety                          FALSE      FALSE
## data$Fertilizer_Amount_kg_per_hectare      FALSE      FALSE
## data$Soil_Quality                          FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          data$Rainfall_mm data$Sunny_Days data$Irrigation_Schedule
## 1  ( 1 ) " "              " "             " "
## 2  ( 1 ) " "              " "             "*"
## 3  ( 1 ) " "              " "             "*"
## 4  ( 1 ) "*"              " "             "*"
## 5  ( 1 ) "*"              " "             "*"
## 6  ( 1 ) "*"              "*"             "*"
##          data$Seed_Variety data$Fertilizer_Amount_kg_per_hectare
## 1  ( 1 ) "*"               " "
## 2  ( 1 ) "*"               " "
## 3  ( 1 ) "*"               "*"
## 4  ( 1 ) "*"               "*"
## 5  ( 1 ) "*"               "*"
## 6  ( 1 ) "*"               "*"
##          data$Soil_Quality
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
```

The asterisks indicate which variables build the best model, according to the number of variables chosen. For example the best 3 variable model is made of the three variables: "Seed_Variety", "Fertilizer_Amount_kg_per_hectare" and "Irrigation_schedule".

(b) What is the best overall model obtained according to AIC, BIC, adjusted R2 and Mallow's Cp? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained.

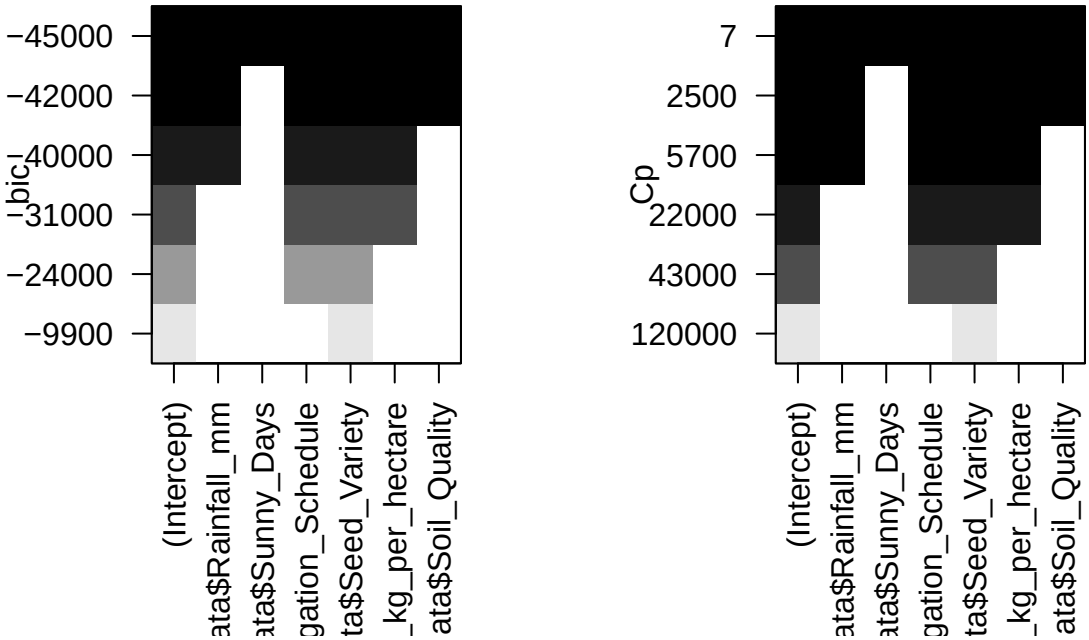
```r
summ = summary(model1)
par(pty="s",mfrow=c(1,3),mar=c(2,1,2,1))
# BIC
plot(summ$bic, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Drop in BIC")
abline (v=which.min(summ$bic),col = 2, lty=2)
# Cp
plot(summ$cp, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Mallow' Cp")
abline (v=which.min(summ$cp),col = 2, lty=2)
#R2
```

```
plot(summ$adjr2, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Adjusted R^2")
abline (v=which.max(summ$adjr2),col = 2, lty=2)
```



From the plots above, it can be stressed that the best model that minimizes BIC, Cp and at the same time guarantees the highest R squared is the one that comprehends all the six predictors considered in the dataset.

```
par(mfrow=c(1,2))
 plot(model1, scale = "bic")
 plot(model1, scale = "Cp")
```



Also the scale plot of BIC and Cp clearly shows that the model that minimizes the two statistics is the one the comprehends the six predictors (this fact may be seen in the model by the darkest squares at the top of the plot)

```
round(coef(model1,6),6)
```

```
##                          (Intercept)                        data$Rainfall_mm
##                            48.334835                              -0.505536
##                       data$Sunny_Days                 data$Irrigation_Schedule
##                             1.992305                              49.986721
##                    data$Seed_Variety data$Fertilizer_Amount_kg_per_hectare
##                           300.463701                              0.808804
##                    data$Soil_Quality
##                             1.543654
```

The coefficients of the best model according to the best selection procedure can be seen above. It may be also stressed that the coefficients coincide with the one used in the multiple regressions. Therefore, the choice of all predictors was right. Below, the main results of Cp and BIC procedures are shown.

BIC function computes the BIC criterion of a given model.

```
null = lm(data$Yield_kg_per_hectare ~ 1, data=data)
best = lm(data$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)
n = nrow(data)
p = 6
bic_best = n*log(deviance(best)/n) + (p+1+1)*log(n)
bic_null = n*log(deviance(null)/n) + 1*log(n)
```

Drop in BIC for best model (omitting the constant)

```
bic_best- bic_null
```

```
## [1] -44611.59
```

```
summ$bic[6]
```

```
## [1] -44611.59
```

BIC of model best (with constant)

```
bic_best
```

```
## [1] 125295
```

The Mallows' Cp statistic can be obtained as:

```
all = lm(data$Yield_kg_per_hectare ~ ., data=data)
deviance(best)/summary(all)$sigma^2+2*(p+1)-n
```

```
## [1] 7
```

```
summ$cp[6]
```

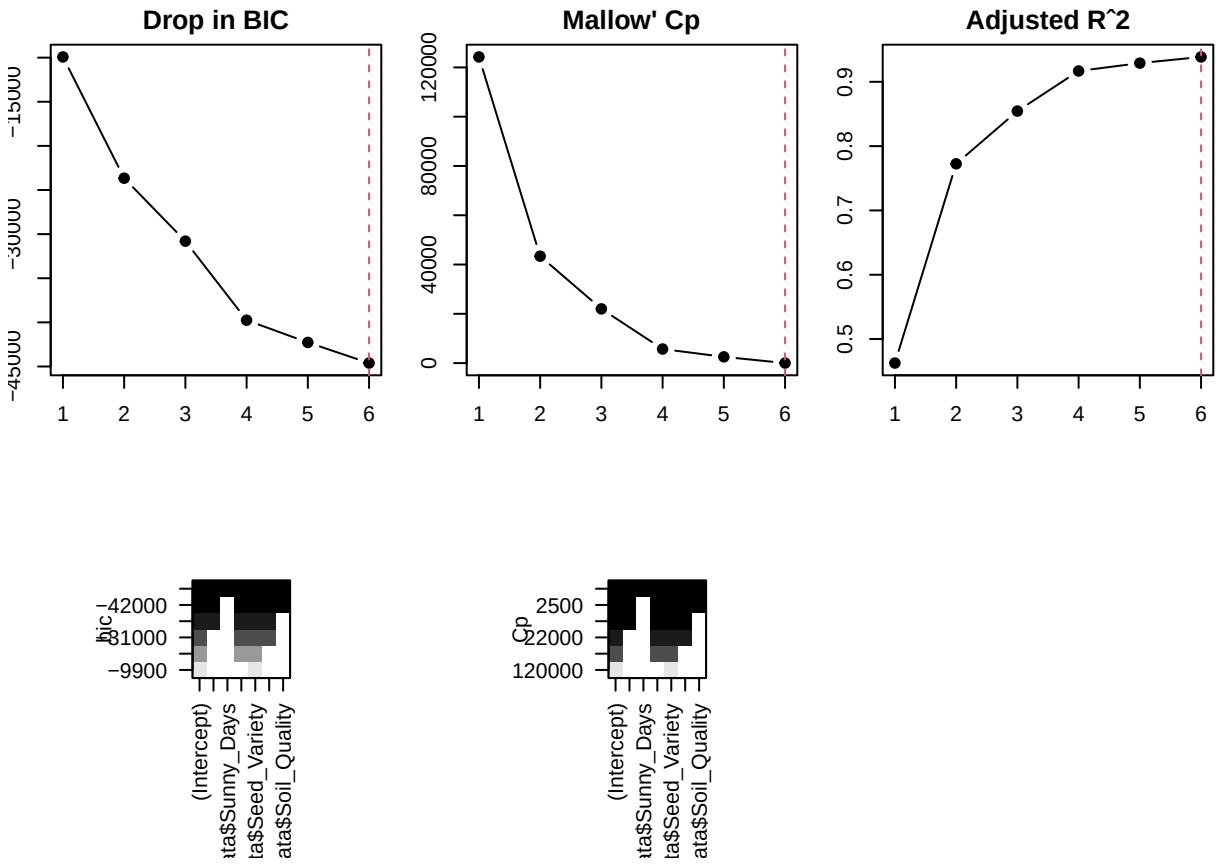
```
## [1] 7
```

(c) Perform a backward selection, comment the results and select the best overall model using BIC, adjusted R2 and Mallow's Cp. Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained.

```
model2 = regsubsets (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, method ="backward")
summary(model2)
```

```
## Subset selection object
## Call: regsubsets.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$Soil_Quality,
##     data = data, method = "backward")
## 6 Variables  (and intercept)
##                                       Forced in Forced out
## data$Rainfall_mm                          FALSE      FALSE
## data$Sunny_Days                           FALSE      FALSE
## data$Irrigation_Schedule                  FALSE      FALSE
## data$Seed_Variety                         FALSE      FALSE
## data$Fertilizer_Amount_kg_per_hectare     FALSE      FALSE
## data$Soil_Quality                         FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: backward
##          data$Rainfall_mm data$Sunny_Days data$Irrigation_Schedule
## 1  ( 1 ) " "              " "             " "
## 2  ( 1 ) " "              " "             "*"
## 3  ( 1 ) " "              " "             "*"
## 4  ( 1 ) "*"              " "             "*"
## 5  ( 1 ) "*"              " "             "*"
## 6  ( 1 ) "*"              "*"             "*"
##          data$Seed_Variety data$Fertilizer_Amount_kg_per_hectare
## 1  ( 1 ) "*"               " "
## 2  ( 1 ) "*"               " "
## 3  ( 1 ) "*"               "*"
## 4  ( 1 ) "*"               "*"
## 5  ( 1 ) "*"               "*"
## 6  ( 1 ) "*"               "*"
##          data$Soil_Quality
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
```

```r
summ2 = summary(model2)
par(pty="s",mfrow=c(2,3),mar=c(2,1,2,1))
# BIC
plot(summ2$bic, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Drop in BIC")
abline (v=which.min(summ2$bic),col = 2, lty=2)
# Cp
plot(summ2$cp, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Mallow' Cp")
abline (v=which.min(summ2$cp),col = 2, lty=2)
#R2
plot(summ2$adjr2, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Adjusted R^2")
abline (v=which.max(summ2$adjr2),col = 2, lty=2)
plot(model2, scale = "bic")
plot(model2, scale = "Cp")
```




```
round(coef(model2,6),6)
```

```
##                          (Intercept)                        data$Rainfall_mm
##                            48.334835                              -0.505536
##                       data$Sunny_Days                  data$Irrigation_Schedule
##                             1.992305                              49.986721
##                    data$Seed_Variety data$Fertilizer_Amount_kg_per_hectare
##                           300.463701                              0.808804
##                    data$Soil_Quality
##                             1.543654
```

According to the backward procedure, the best model is the six variables one (the same of the best overall model of the previous point). Therefore, since the selected variables are the same, the BIC, Cp and coefficients give the same results.

(d) Perform a forward selection, comment the results and select the best overall model using BIC, adjusted R2 and Mallow's Cp. Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained.

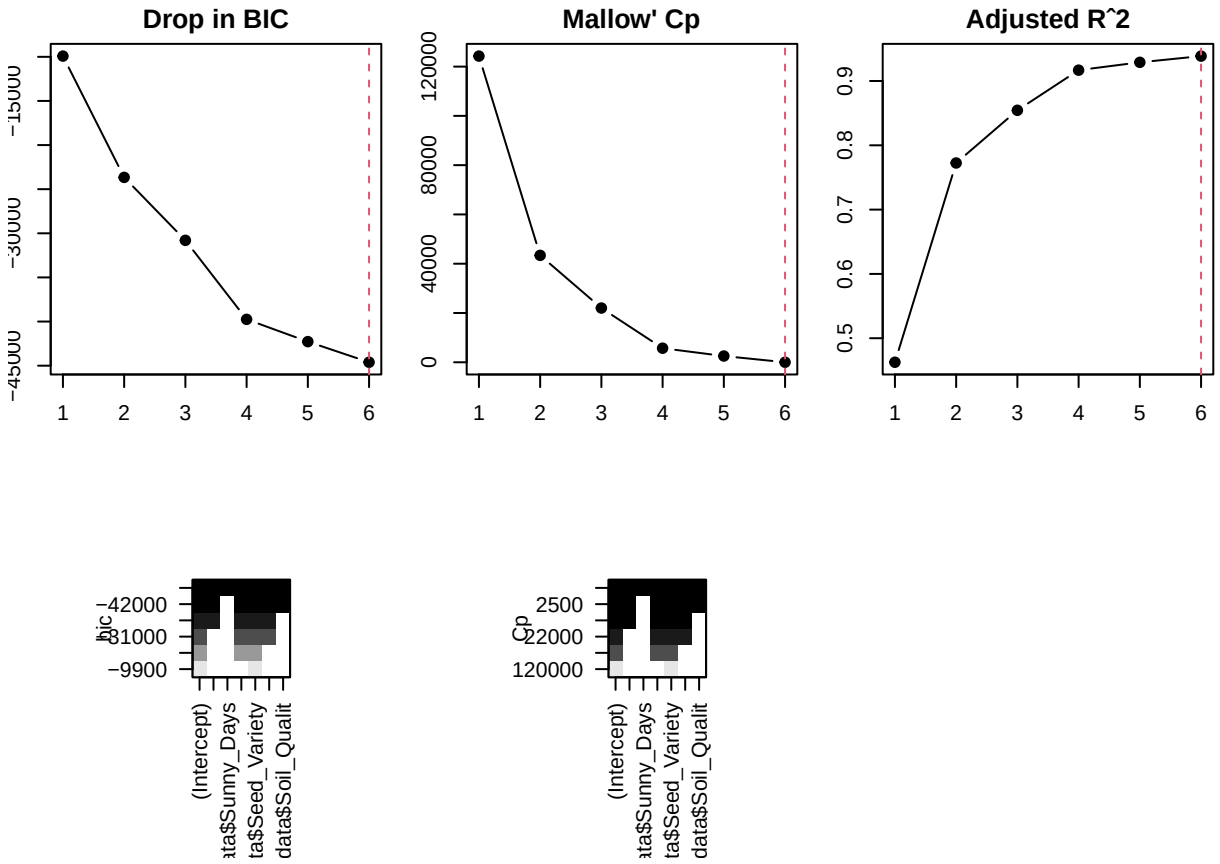
```r
model3 = regsubsets (data$Yield_kg_per_hectare ~
data$Rainfall_mm +
data$Sunny_Days +
data$Irrigation_Schedule +
data$Seed_Variety +
data$Fertilizer_Amount_kg_per_hectare +
data$Soil_Qualit, data=data, method ="forward")
summary(model3)
```

```
## Subset selection object
## Call: regsubsets.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$Soil_Qualit,
##     data = data, method = "forward")
## 6 Variables  (and intercept)
##                                      Forced in Forced out
## data$Rainfall_mm                         FALSE      FALSE
## data$Sunny_Days                          FALSE      FALSE
## data$Irrigation_Schedule                 FALSE      FALSE
## data$Seed_Variety                        FALSE      FALSE
## data$Fertilizer_Amount_kg_per_hectare    FALSE      FALSE
## data$Soil_Qualit                         FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: forward
##          data$Rainfall_mm data$Sunny_Days data$Irrigation_Schedule
## 1  ( 1 ) " "              " "             " "
## 2  ( 1 ) " "              " "             "*"
## 3  ( 1 ) " "              " "             "*"
## 4  ( 1 ) "*"              " "             "*"
## 5  ( 1 ) "*"              " "             "*"
## 6  ( 1 ) "*"              "*"             "*"
##          data$Seed_Variety data$Fertilizer_Amount_kg_per_hectare
## 1  ( 1 ) "*"               " "
## 2  ( 1 ) "*"               " "
## 3  ( 1 ) "*"               "*"
## 4  ( 1 ) "*"               "*"
## 5  ( 1 ) "*"               "*"
## 6  ( 1 ) "*"               "*"
##          data$Soil_Qualit
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
```

```r
summ3 = summary(model3)
par(pty="s",mfrow=c(2,3),mar=c(2,1,2,1))
# BIC
plot(summ3$bic, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Drop in BIC")
abline (v=which.min(summ3$bic),col = 2, lty=2)
```

```r
# Cp
plot(summ3$cp, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Mallow' Cp")
abline (v=which.min(summ3$cp),col = 2, lty=2)
#R2
plot(summ3$adjr2, type="b", pch=19,
xlab="Number of predictors", ylab="", main="Adjusted R^2")
abline (v=which.max(summ3$adjr2),col = 2, lty=2)

plot(model3, scale = "bic")
plot(model3, scale = "Cp")
```




```r
round(coef(model3,6),6)
```

```
##                          (Intercept)                        data$Rainfall_mm
##                            48.334835                             -0.505536
##                       data$Sunny_Days                  data$Irrigation_Schedule
##                             1.992305                             49.986721
##                    data$Seed_Variety data$Fertilizer_Amount_kg_per_hectare
##                           300.463701                              0.808804
##                     data$Soil_Qualit
##                             1.543654
```

According to the forward procedure, the best model is the six variables one (the same of the model obtained through the backwar procedure of the previous point). Therefore, since the selected variables are the same, the BIC, Cp and coefficients give the same results.

(e) Compare the results of points a), c) and d). Do the three procedure identify the best p models?

The results are the same. The procedures identify the right models because they include all the variables in order to obtain the lowest BIC and Cp values and the highest R squared. The model that includes all the variables is the best according to these algorithms.

(f) Using the validation set approach, compute the MSE of your model. Repeat the process three times, using three different splits of the observations into a training set and a validation set. Comment on your results.

```
library(rpart)
set.seed(1)
trainx = sample(nrow(data), nrow(data)/3, replace = FALSE )
testx = data[-trainx,]
olsx = lm(Yield_kg_per_hectare ~
Rainfall_mm +
Sunny_Days +
Irrigation_Schedule +
Seed_Variety +
Fertilizer_Amount_kg_per_hectare +
Soil_Quality,data=data,subset=trainx)
predx=predict(olsx,newdata=testx)
MSEx=mean((testx$Yield_kg_per_hectare-predx)^2)
MSEx
```

```
## [1] 2508.848
```

```
set.seed(2)
trainy = sample(nrow(data), nrow(data)*(1/3), replace = FALSE )
testy = data[-trainy,]
olsy = lm(Yield_kg_per_hectare ~
Rainfall_mm +
Sunny_Days +
Irrigation_Schedule +
Seed_Variety +
Fertilizer_Amount_kg_per_hectare +
Soil_Quality,data=data,subset=trainy)
predy=predict(olsy,newdata=testy)
MSEy=mean((testy$Yield_kg_per_hectare-predy)^2)
MSEy
```

```
## [1] 2526.593
```

```
set.seed(3)
trainz = sample(nrow(data), nrow(data)/3, replace = FALSE )
```

```r
testz = data[-trainz,]
olsz = lm(Yield_kg_per_hectare ~
Rainfall_mm +
Sunny_Days +
Irrigation_Schedule +
Seed_Variety +
Fertilizer_Amount_kg_per_hectare +
Soil_Quality,data=data,subset=trainz)
predz=predict(olsz,newdata=testz)
MSEz=mean((testz$Yield_kg_per_hectare-predz)^2)
MSEz
```

```
## [1] 2490.838
```