Assignment6

Lorenzo Ricciardulli

2024-02-22

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

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

## Warning: il pacchetto 'alr4' è stato creato con R versione 4.1.3

## Caricamento del pacchetto richiesto: car

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

## Caricamento del pacchetto richiesto: carData

##
## Caricamento pacchetto: 'car'

## I seguenti oggetti sono mascherati da 'package:faraway':
##
##     logit, vif

## Caricamento del pacchetto richiesto: effects

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

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

##
## Caricamento pacchetto: 'alr4'

## I seguenti oggetti sono mascherati da 'package:faraway':
##
##     cathedral, pipeline, twins

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

## corrplot 0.92 loaded

##
## Caricamento pacchetto: 'ellipse'
```

```
## Il seguente oggetto è mascherato da 'package:car':
##
##     ellipse

## Il seguente oggetto è mascherato da 'package:graphics':
##
##     pairs

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

##
## Caricamento pacchetto: 'boot'

## Il seguente oggetto è mascherato da 'package:car':
##
##     logit

## I seguenti oggetti sono mascherati da 'package:faraway':
##
##     logit, melanoma

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

##
## Caricamento pacchetto: 'ROCit'

## Il seguente oggetto è mascherato da 'package:boot':
##
##     logit

## Il seguente oggetto è mascherato da 'package:car':
##
##     logit

## Il seguente oggetto è mascherato da 'package:faraway':
##
##     logit
```

Loading data

```
data = read.csv("yield.txt", header = TRUE)
```

(a) Given the dataset used in the previous assignments, consider now a binary response (e.g., you may split your continuous response variable into two categories and explicit the meaning of Y = 1 and Y =0). Provide a graphical representation of the data.

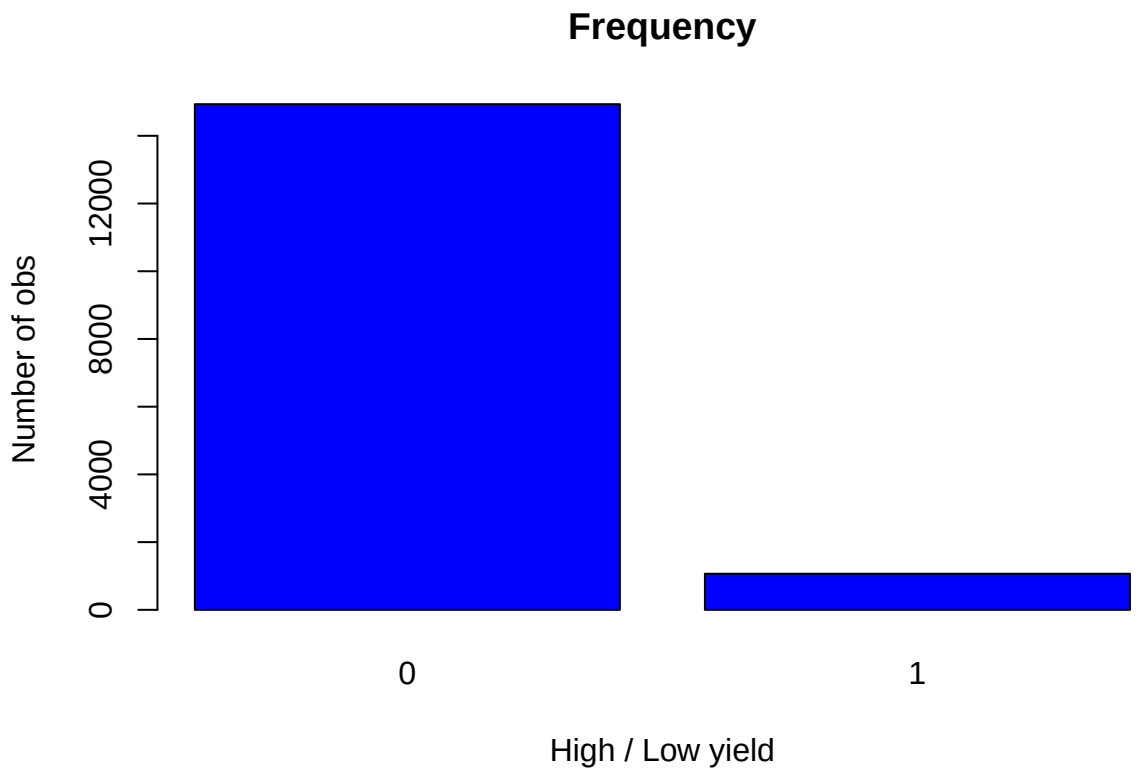
Splitting the response variable

The response variable could be split in 2 categories. For example, among the 16000 observations, the observations that have the value of yield_kg_per_hectare above (or equal) 1000 kg per hectare could be

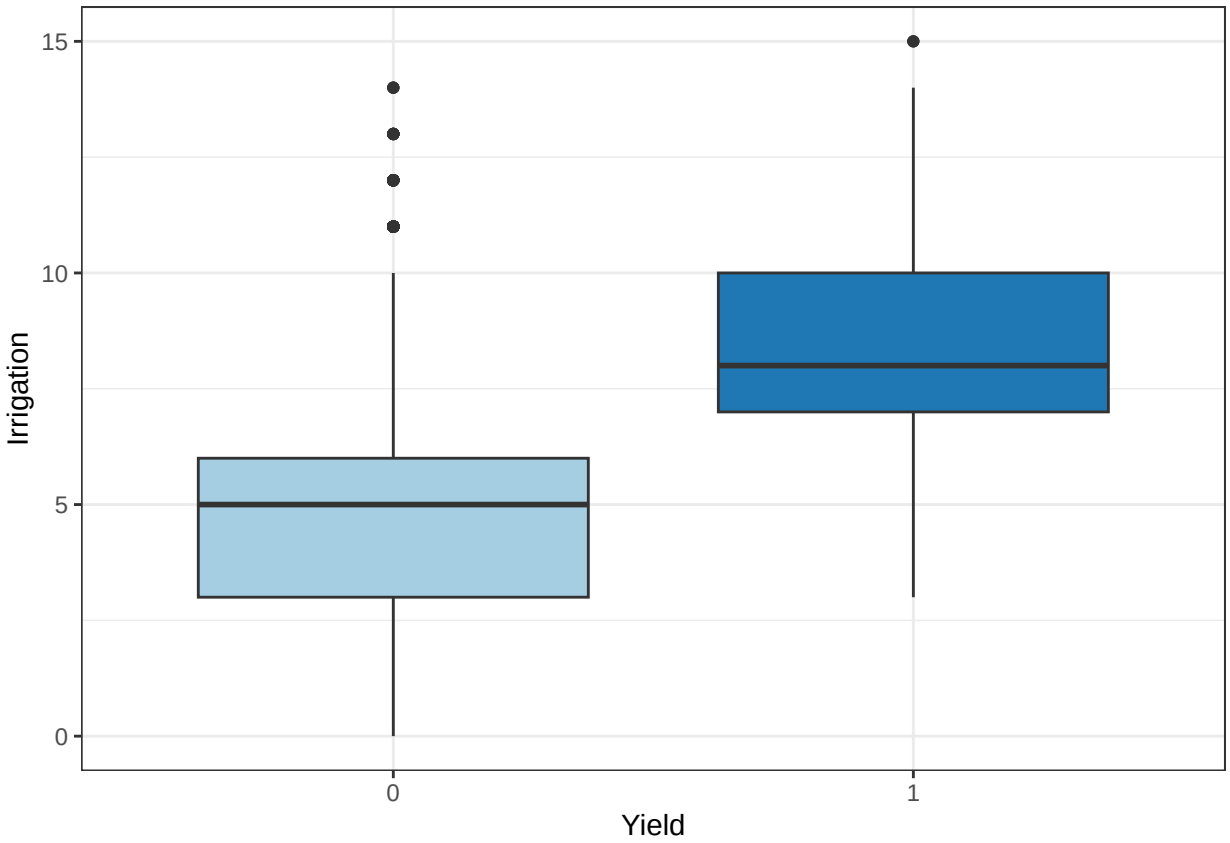
considered as “High Yield” and replaced with the value 1. The observations below 1000 kg per hectare could be considered as “Low Yield” and indicated by the value 0.

```
for (i in 1:nrow(data)){
  if (data$Yield_kg_per_hectare[i] < 1000) {
    data$Yield_kg_per_hectare[i] = 0
  } else {
    data$Yield_kg_per_hectare[i] = 1
  }
}
```

```
freq = table(data$Yield_kg_per_hectare)
barplot(freq, main = "Frequency", xlab = "High / Low yield", ylab = "Number of obs", col = "blue")
```




```
ggplot(data, aes(x = factor(Yield_kg_per_hectare), y = Irrigation_Schedule)) +
geom_boxplot( fill=c("#A6CEE3", "#1F78B4")) +
labs(x = "Yield", y = "Irrigation")+
theme_bw()
```



(b) Fit a logistic regression model using the best subset predictors selected in the previous assignment and interpret the coefficients and their uncertainties (hint: you may also provide confidence intervals for the coefficients). Which predictors are relevant?

According to BIC, Mallow's CP and maximization of R squared procedures, the best set of predictors was the complete one (with 6 predictors).

```
logitmod = glm(data$Yield_kg_per_hectare ~ data$Soil_Quality +
data$Seed_Variety +
data$Fertilizer_Amount_kg_per_hectare +
data$Sunny_Days +
data$Rainfall_mm +
data$Irrigation_Schedule,
family = binomial(link = logit), data = data)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logitmod)
```

```
##
## Call:
## glm(formula = data$Yield_kg_per_hectare ~ data$Soil_Quality +
```

```
##     data$Seed_Variety + data$Fertilizer_Amount_kg_per_hectare +
##     data$Sunny_Days + data$Rainfall_mm + data$Irrigation_Schedule,
##     family = binomial(link = logit), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5593  -0.0478  -0.0051  -0.0002   3.2489
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -3.631e+01  1.447e+00  -25.10   <2e-16
## data$Soil_Quality                    5.845e-02  4.220e-03   13.85   <2e-16
## data$Seed_Variety                    1.188e+01  7.470e-01   15.90   <2e-16
## data$Fertilizer_Amount_kg_per_hectare  3.038e-02  1.229e-03   24.71   <2e-16
## data$Sunny_Days                      7.328e-02  6.134e-03   11.95   <2e-16
## data$Rainfall_mm                    -1.827e-02  7.852e-04  -23.26   <2e-16
## data$Irrigation_Schedule             1.843e+00  5.870e-02   31.40   <2e-16
##
## (Intercept)                         ***
## data$Soil_Quality                   ***
## data$Seed_Variety                   ***
## data$Fertilizer_Amount_kg_per_hectare ***
## data$Sunny_Days                     ***
## data$Rainfall_mm                    ***
## data$Irrigation_Schedule            ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7834.2  on 15999  degrees of freedom
## Residual deviance: 2165.7  on 15993  degrees of freedom
## AIC: 2179.7
##
## Number of Fisher Scoring iterations: 11
```

Apparentely, all predictors are relevant.

```
exp(coef(logitmod)[1])
```

```
##  (Intercept)
## 1.695514e-16
```

```
exp(coef(logitmod)[2])
```

```
## data$Soil_Quality
##          1.060196
```

```
exp(coef(logitmod)[3])
```

```
## data$Seed_Variety
##          144579.6
```

```r
exp(coef(logitmod)[4])
```

```
## data$Fertilizer_Amount_kg_per_hectare
##                              1.030847
```

```r
exp(coef(logitmod)[5])
```

```
## data$Sunny_Days
##        1.076027
```

```r
exp(coef(logitmod)[6])
```

```
## data$Rainfall_mm
##        0.9818998
```

```r
exp(coef(logitmod)[7])
```

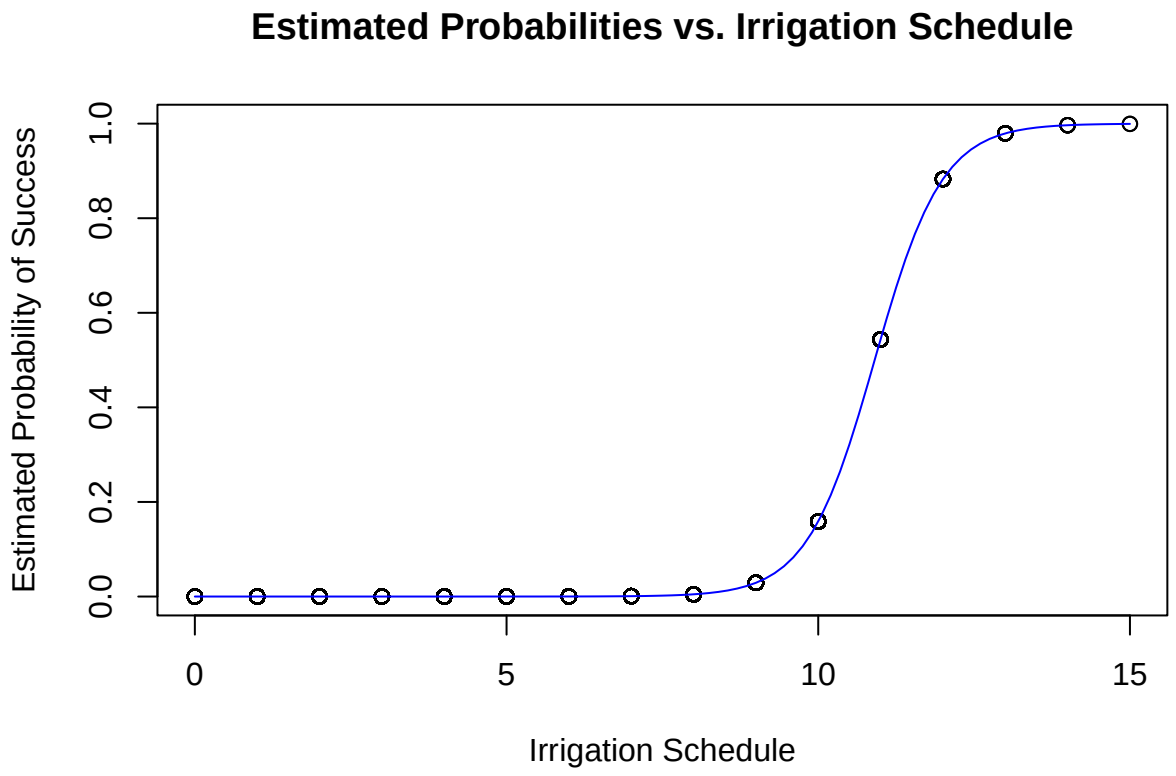
```
## data$Irrigation_Schedule
##                  6.31727
```

Since a binomial regression was performed, the estimates in the glm output (above) cannot be interpreted directly as it happens in the linear regression. The estimates are to be interpreted as logs-odd. Therefore, in order to obtain the log odds, it is needed to compute the exponential of the estimate in order to find the ratio between probability of success and probability of failure (odds). The exponential results of the odds of single regressors are showed above. They can be interpreted as it follows: -Soil_Quality : it is positive and greater than 1. It means that, for each additional unit, it is 1.06 times more likely to improve the Yield rather than not to improve it. -Seed_Variety : it is positive and way greater than 1 (144579.6). It is a categorical variable with values 0 and 1. The probability of a "high-yield observation" is very high once the cathegories change from no variety to variety in the fields. The value of odds is very high since the probability of success (changing the cathegory) tends to 0: the denominator of odds tends to 0 and it determines a substanital increase of probability of success. Basically if the cathegory is changed, the probability of success is almost certain. -Fertilizer_Amount_kg_per_hectare : it is positive and grater than 1. It means that, for each additional unit, it is 1.03 times more likely to improve the Yield rather than not to improve it. -Sunny_days : It means that, for each additional unit, it is 1.08 times more likely to improve the Yield rather than not to improve it. -Rainfall_mm : it is positive but lower than 1. It means that for each additional unit, the odds of success (improving the yield) decrease. It is 0.98 less likely to improve the yield rather than not to improve it.

(c) Provide the estimated probabilities of success and plot them versus the most relevant predictor, holding the other predictors constant to their mean values.

```r
invlogit = function(x) {
  exp(x) / (1 + exp(x))
}
mean_values = sapply(data[, c("Soil_Quality", "Seed_Variety", "Fertilizer_Amount_kg_per_hectare", "Sunny
"Rainfall_mm", "Irrigation_Schedule")], mean)
coefficients = coef(logitmod)
eta = coefficients[1] +
```

```
coefficients[2] * mean_values["Soil_Quality"] +
coefficients[3] * mean_values["Seed_Variety"] +
coefficients[4] * mean_values["Fertilizer_Amount_kg_per_hectare"] +
coefficients[5] * mean_values["Sunny_Days"] +
coefficients[6] * mean_values["Rainfall_mm"] +
coefficients[7] * data$Irrigation_Schedule
estimated_probabilities = invlogit(eta)
plot(data$Irrigation_Schedule, estimated_probabilities,
ylab = "Estimated Probability of Success",
xlab = "Irrigation Schedule",
main = "Estimated Probabilities vs. Irrigation Schedule",
ylim = c(0, 1))
curve(invlogit(coefficients[1] +
coefficients[2] * mean_values["Soil_Quality"] +
coefficients[3] * mean_values["Seed_Variety"] +
coefficients[4] * mean_values["Fertilizer_Amount_kg_per_hectare"] +
coefficients[5] * mean_values["Sunny_Days"] +
coefficients[6] * mean_values["Rainfall_mm"] +
coefficients[7] * x),
add = TRUE, col = "blue", lwd = 1)
```



(d) Test the null hypothesis of constant probability of success, versus the hypothesis that probability depends on your set of regressors. Test also a group of regressors (motivate your choice); explicitly state and discuss the hypotheses tested and the conclusions.

```
D_l=logitmod$deviance;
D_s=logitmod$null.deviance;
l = logitmod$df.null
s = logitmod$df.residual
1-pchisq(D_s-D_l,l-s)
```

```
## [1] 0
```

```
D_null = glm(data$Yield_kg_per_hectare~1, data =data, family = binomial(link =logit) )
anova(logitmod,D_null,test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: data$Yield_kg_per_hectare ~ data$Soil_Quality + data$Seed_Variety +
##     data$Fertilizer_Amount_kg_per_hectare + data$Sunny_Days +
##     data$Rainfall_mm + data$Irrigation_Schedule
## Model 2: data$Yield_kg_per_hectare ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     15993     2165.7
## 2     15999     7834.2 -6  -5668.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The hypothesis tested above (the results are the same, the second one has more details) are the following: null hypothesis : the complete model is not adequate. It implies that the deviance of null model is equal to the complete model's one. The deviance is a measure that tells how close the model is to the "perfect one", obtained by using likelihood maximization procedures. If the deviance of the null model is equal to the complete one, it can be said that there are no improvements with the inclusion of the regressors, therefore the model tested is not adequate; alternative hypothesis: the complete model is adequate. The test implies the fact that deviance is distributed as a chi squared with n - number of parameters degrees of freedom. As usual, if the statistic is greater than the observed value with a very low probability (p-value), it can be concluded that the null hypothesis is rejected. It happens in this case: as it can be seen in the output above, the deviance of the complete model is 2165.7 and the null model's one is 7834.2; therefore the p.value is very low and we reject the null hypothesis that states that the model is not adequate.

testing a group of regressors

```
logitmod2 = glm(data$Yield_kg_per_hectare ~
data$Seed_Variety +
data$Irrigation_Schedule +
data$Soil_Quality +
data$Fertilizer_Amount_kg_per_hectare+
data$Sunny_Days,
family = binomial(link = logit), data = data)
```

```r
D_complete = logitmod$deviance
D_reduced = logitmod2$deviance
c = logitmod$df.residual
r = logitmod2$df.residual
1-pchisq(D_reduced-D_complete,r-c)
```

```
## [1] 0
```

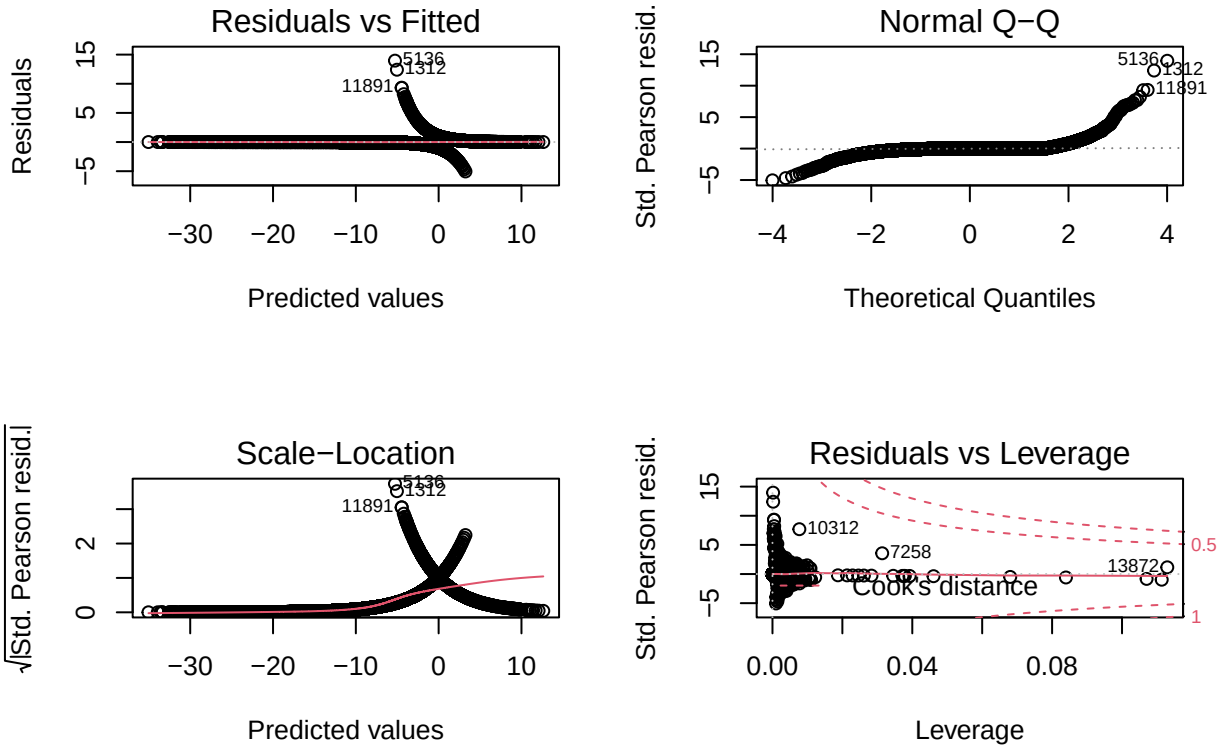
```r
anova(logitmod, logitmod2, test ="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: data$Yield_kg_per_hectare ~ data$Soil_Quality + data$Seed_Variety +
##     data$Fertilizer_Amount_kg_per_hectare + data$Sunny_Days +
##     data$Rainfall_mm + data$Irrigation_Schedule
## Model 2: data$Yield_kg_per_hectare ~ data$Seed_Variety + data$Irrigation_Schedule +
##     data$Soil_Quality + data$Fertilizer_Amount_kg_per_hectare +
##     data$Sunny_Days
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     15993     2165.7
## 2     15994     3054.6 -1  -888.91 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The reduced model is equale to the complete model excluded the variable "Rainfall_mm". In the first points it was showed that this variable was the only one that did not lead to an increase in odds, so it can be tested in order to establish whether it can be excluded from the model or not. With the same rationale explained above, the two following hypothesis are tested. Null hypothesis: the complete and the reduced model are equal. Alternative model: the complete model is better (with a lower deviance). The results show that the null hypothesis can be rejected.

(e) Perform residual checks, discuss the results and propose potential remedies for addressing model issues.

```r
par(mfrow=c(2,2))
plot(logitmod)
```



(f) Think about a potential competing model (motivate your choice) and use the AIC to compare your current model against the competing one.

```
logitmod3 = glm(data$Yield_kg_per_hectare ~
data$Seed_Variety +
data$Fertilizer_Amount_kg_per_hectare +
data$Irrigation_Schedule +
data$Soil_Quality+
data$Sunny_Days,
family = binomial(link = logit), data = data)
```

```
BS= stepAIC(logitmod,scope=list(lower=logitmod3,upper=logitmod),
 direction='backward',k=1,
 steps=16000,
 trace=1,data=data)
```

```
## Start:  AIC=2172.73
## data$Yield_kg_per_hectare ~ data$Soil_Quality + data$Seed_Variety +
##     data$Fertilizer_Amount_kg_per_hectare + data$Sunny_Days +
##     data$Rainfall_mm + data$Irrigation_Schedule
##
##                       Df Deviance    AIC
## <none>                     2165.7 2172.7
## - data$Rainfall_mm  1   3054.6 3060.6
```

With the same rationale of the points before, it can be reasonable to test whether the AIC of the model without "Rainfall_mm" is lower with respect to the complete one's, to see if that variable could be excluded. The results show that the AIC of the complete model is lower, i.e. the variable "Rainfall_mm" should be included.

(g) Use the cross-validation to compare your model against the competing model via the error rate.

```
set.seed(123)
train.idx <- sample(nrow(data), 0.8*nrow(data))
train <- data[train.idx, ]
test <- data[-train.idx, ]

logitmod_rainfall_mm_train =  glm(Yield_kg_per_hectare ~ Soil_Quality +
Seed_Variety +
Fertilizer_Amount_kg_per_hectare +
Sunny_Days +
Irrigation_Schedule, family = binomial(link = logit), data = data, subset = train.idx)

predicted_rainfall_mm = predict(logitmod_rainfall_mm_train, train, type = "response")
alpha = 0.5
yhat_int = ifelse(predicted_rainfall_mm >= alpha, 1, 0)
tab_int = table(train$Yield_kg_per_hectare, yhat_int)
tab_int
```

```
##    yhat_int
##         0     1
##   0 11796   154
##   1   356   494
```

```
#accuracy on the training set
sum(diag(tab_int))/sum(tab_int) #accuracy
```

```
## [1] 0.9601563
```

```
tab_int[1,1]/(tab_int[1,1]+tab_int[1,2]) #specificity
```

```
## [1] 0.987113
```

```
tab_int[2,2]/(tab_int[2,1]+tab_int[2,2]) #sensitivity
```

```
## [1] 0.5811765
```

```
predicted_rainfall_test = predict(logitmod_rainfall_mm_train, test, type = "response")
alpha = 0.5
yhat_int = ifelse(predicted_rainfall_test >= alpha, 1, 0)
tab_int = table(test$Yield_kg_per_hectare, yhat_int)
tab_int
```

```
##    yhat_int
##        0    1
##   0 2943   41
##   1   90  126
```

```
#accuracy on the training set
sum(diag(tab_int))/sum(tab_int) #accuracy
```

```
## [1] 0.9590625
```

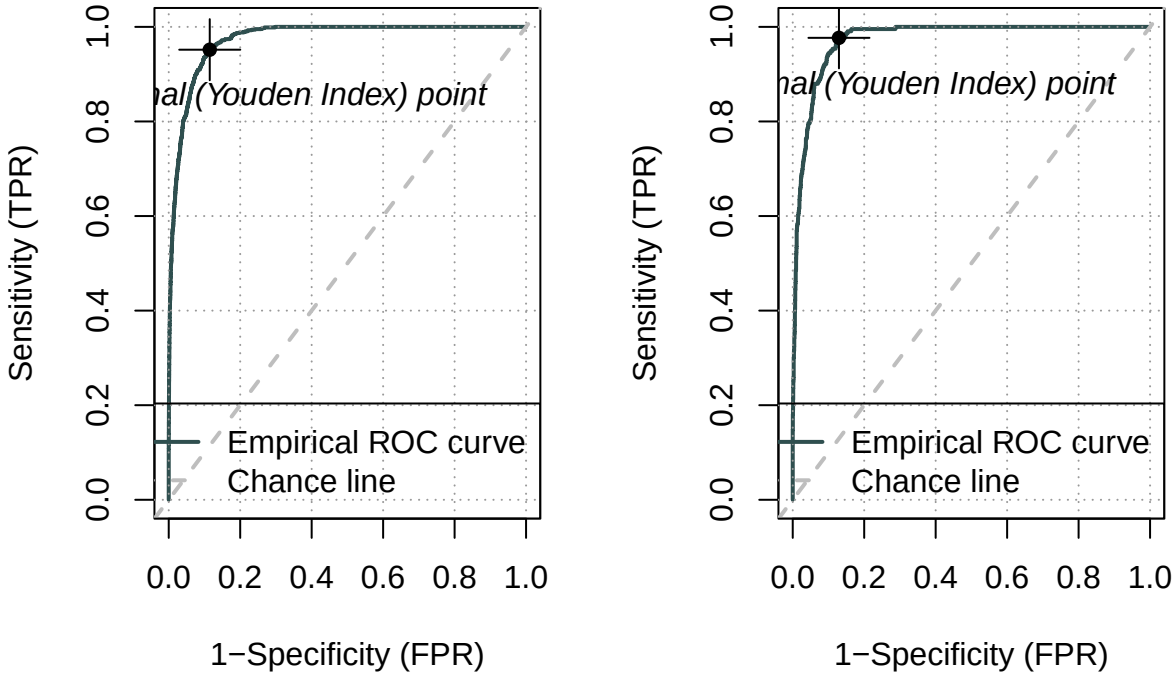
```
tab_int[1,1]/(tab_int[1,1]+tab_int[1,2]) #specificity
```

```
## [1] 0.9862601
```

```
tab_int[2,2]/(tab_int[2,1]+tab_int[2,2]) #sensitivity
```

```
## [1] 0.5833333
```

```
par(mfrow=c(1,2))
plot(rocit(predicted_rainfall_mm,train$Yield_kg_per_hectare))
plot(rocit(predicted_rainfall_test,test$Yield_kg_per_hectare))
```



```r
set.seed(12)
train.idx <- sample(nrow(data), 0.8*nrow(data))
train <- data[train.idx, ]
test <- data[-train.idx, ]

logitmod_train = glm(Yield_kg_per_hectare ~ Soil_Quality +
Seed_Variety +
Fertilizer_Amount_kg_per_hectare +
Sunny_Days +
Irrigation_Schedule + Rainfall_mm, family = binomial(link = logit), data = data, subset = train.idx)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
predicted_logitmod = predict(logitmod_train, train, type = "response")
alpha = 0.5
yhat_int = ifelse(predicted_logitmod >= alpha, 1, 0)
tab_int = table(train$Yield_kg_per_hectare, yhat_int)
tab_int
```

```
##    yhat_int
##         0     1
##   0 11799   136
##   1   230   635
```

```r
#accuracy on the training set
sum(diag(tab_int))/sum(tab_int) #accuracy
```

```
## [1] 0.9714063
```

```r
tab_int[1,1]/(tab_int[1,1]+tab_int[1,2]) #specificity
```

```
## [1] 0.9886049
```

```r
tab_int[2,2]/(tab_int[2,1]+tab_int[2,2]) #sensitivity
```

```
## [1] 0.734104
```

```r
predicted_logitmod_test = predict(logitmod_train, test, type = "response")
alpha = 0.5
yhat_int = ifelse(predicted_logitmod_test >= alpha, 1, 0)
tab_int = table(test$Yield_kg_per_hectare, yhat_int)
tab_int
```

```
##    yhat_int
##        0    1
##   0 2968   31
##   1   57  144
```

```r
#accuracy on the training set
sum(diag(tab_int))/sum(tab_int) #accuracy
```

```
## [1] 0.9725
```

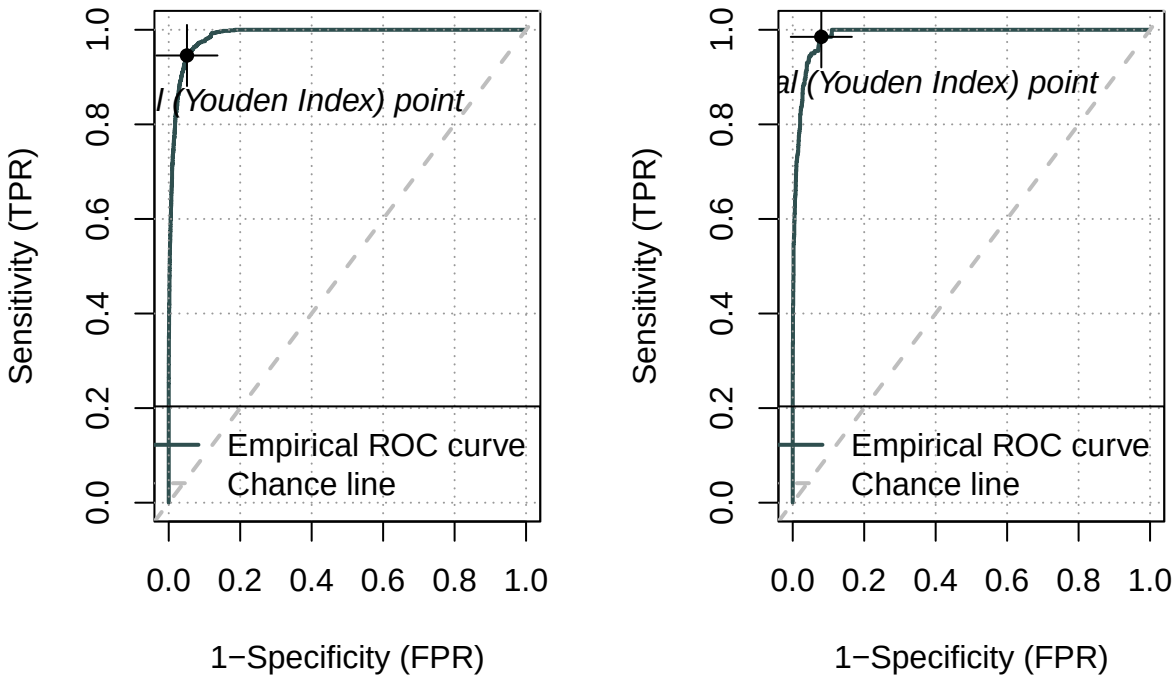
```r
tab_int[1,1]/(tab_int[1,1]+tab_int[1,2]) #specificity
```

```
## [1] 0.9896632
```

```r
tab_int[2,2]/(tab_int[2,1]+tab_int[2,2]) #sensitivity
```

```
## [1] 0.7164179
```

```r
par(mfrow=c(1,2))
plot(rocit(predicted_logitmod,train$Yield_kg_per_hectare))
plot(rocit(predicted_logitmod_test,test$Yield_kg_per_hectare))
```



The results show that are no substantial differences in training and test between the complete model and the reduced one