BE 2 - Régression Logistique - Exercice

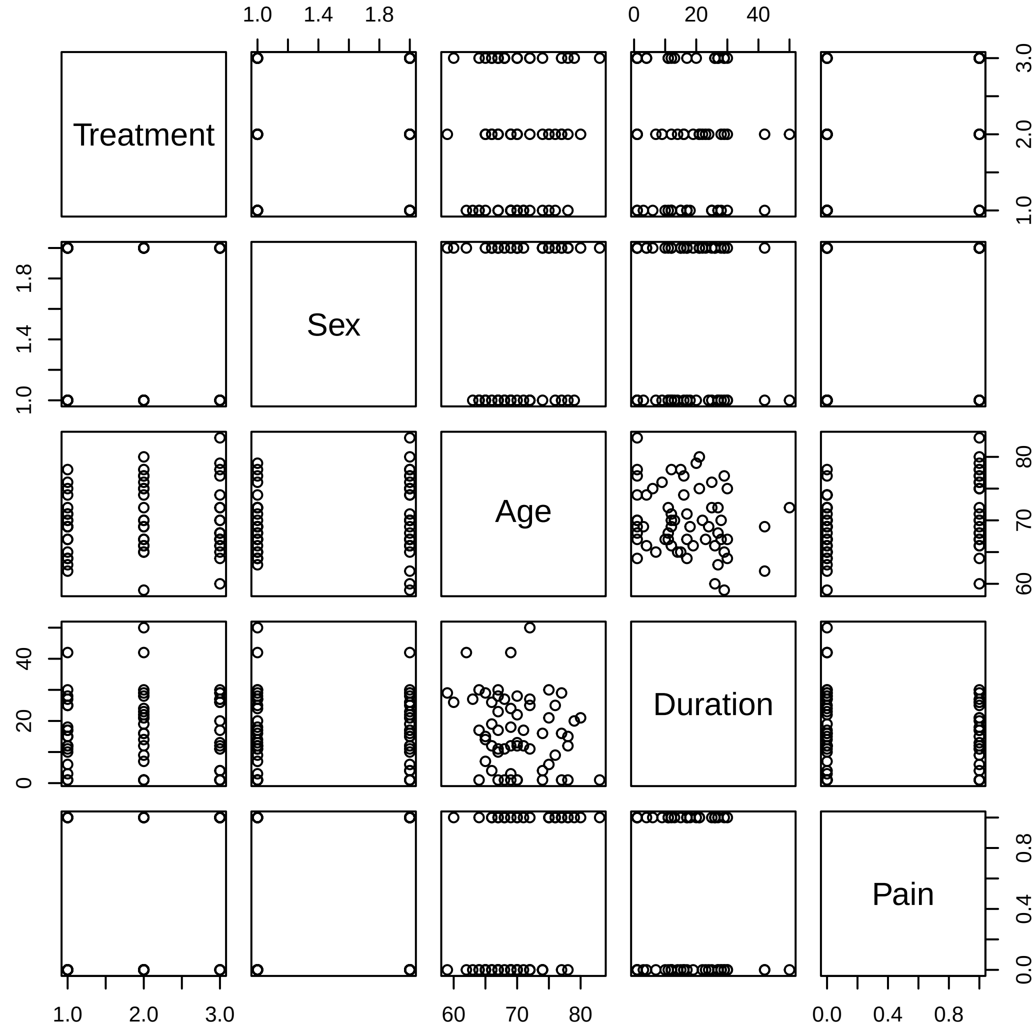
Exercice 2

Question 1

```r
Neuralgia <- read.table("neuralgia.txt", header = TRUE)
summary(Neuralgia)
```

```
##   Treatment              Sex                  Age           Duration    
##  Length:60          Length:60          Min.   :59.00   Min.   : 1.00  
##  Class :character   Class :character   1st Qu.:66.75   1st Qu.: 8.50  
##  Mode  :character   Mode  :character   Median :69.00   Median :16.00  
##                                        Mean   :70.05   Mean   :16.73  
##                                        3rd Qu.:74.00   3rd Qu.:26.00  
##                                        Max.   :83.00   Max.   :50.00  
##       Pain       
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.4167  
##  3rd Qu.:1.0000  
##  Max.   :1.0000
```

```r
plot(Neuralgia)
```



On dispose de 4 variables explicatives : Age et Duration sont des variables quantitatives, tandis que Sexe et Treatment sont des variables qualitatives. On veut expliquer la variable Pain qui vaut 0 ou 1.

Question 2 :

```
set.seed(123) # pour reproductibilité
u <- sample(1:60, 48)
NeuralgiaApp <- Neuralgia[u, ]
NeuralgiaTest <- Neuralgia[-u, ]
```

Question 3 :

```
reglNeuralgia <- glm(Pain ~ ., family = binomial(link = "logit"), data = NeuralgiaApp)
summary(reglNeuralgia)
```

```
##
## Call:
## glm(formula = Pain ~ ., family = binomial(link = "logit"), data = NeuralgiaApp)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -22.09711    8.16210  -2.707  0.00678 **
## TreatmentB   -0.84282    1.01571  -0.830  0.40666
## TreatmentP    2.74823    1.13765   2.416  0.01570 *
## SexM          1.51851    0.84189   1.804  0.07128 .
## Age           0.28625    0.11081   2.583  0.00979 **
## Duration      0.01366    0.03607   0.379  0.70500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 65.790  on 47  degrees of freedom
## Residual deviance: 40.972  on 42  degrees of freedom
## AIC: 52.972
##
## Number of Fisher Scoring iterations: 5
```

La régression logistique modélise la probabilité que Pain $= 1$ (présence de douleur) en fonction des variables explicatives. Le modèle Logit associé est :

$$\log\left(\frac{P(\text{Pain}=1)}{1-P(\text{Pain}=1)}\right) = \beta_0 + \beta_1\text{Treatment} + \beta_2\text{Sex} + \beta_3\text{Age} + \beta_4\text{Duration}$$

Question 4 :

```r
Anova(reglNeuralgia, type = "III", test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: Pain
##           LR Chisq Df Pr(>Chisq)
## Treatment  12.9378  2   0.001551 **
## Sex         3.6116  1   0.057378 .
## Age        10.3860  1   0.001270 **
## Duration    0.1417  1   0.706551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(reglNeuralgia, type = "III", test.statistic = "Wald")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: Pain
##             Df  Chisq Pr(>Chisq)
## (Intercept)  1 7.3294   0.006784 **
## Treatment    2 8.5486   0.013922 *
## Sex          1 3.2533   0.071278 .
```

```
## Age          1 6.6732   0.009787 **
## Duration     1 0.1433   0.704999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(reglNeuralgia, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Pain
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         47     65.790
## Treatment  2   8.2428        45     57.547 0.016222 *
## Sex        1   5.7087        44     51.839 0.016881 *
## Age        1  10.7246        43     41.114 0.001057 **
## Duration   1   0.1417        42     40.972 0.706551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 5

Principe du forward AIC : On commence avec le modèle le plus simple (constant), et on ajoute une à une les variables qui minimisent le critère d'AIC. Lorsque l'ajout d'une variable réduit le critère d'AIC, on arrête le programme.

```r
constant <- glm(Pain ~ 1, family = binomial, data = NeuralgiaApp)
forward <- step(constant, scope = list(lower = constant, upper = Pain ~ Treatment + Sex + Age + Duratio
```

```
## Start:  AIC=67.79
## Pain ~ 1
##
##             Df Deviance    AIC
## + Age        1   56.395 60.395
## + Treatment  2   57.547 63.547
## + Sex        1   61.576 65.576
## <none>           65.790 67.790
## + Duration   1   64.960 68.960
##
## Step:  AIC=60.39
## Pain ~ Age
##
##             Df Deviance    AIC
## + Treatment  2   44.902 52.902
## + Sex        1   54.020 60.020
## <none>           56.395 60.395
```

```
## + Duration   1   56.323 62.323
##
## Step:  AIC=52.9
## Pain ~ Age + Treatment
##
##             Df Deviance    AIC
## + Sex        1   41.114 51.114
## <none>           44.902 52.902
## + Duration  1   44.584 54.584
##
## Step:  AIC=51.11
## Pain ~ Age + Treatment + Sex
##
##             Df Deviance    AIC
## <none>           41.114 51.114
## + Duration  1   40.972 52.972
```

```r
# Vérification de l'ANOVA du modèle sélectionné

forward$anova
```

```
##          Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1              NA        NA        47   65.79016 67.79016
## 2       + Age -1  9.395630        46   56.39453 60.39453
## 3 + Treatment -2 11.492946        44   44.90159 52.90159
## 4       + Sex -1  3.787475        43   41.11411 51.11411
```

```r
Anova(forward, test.statistic = "LR", type = 'III')
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: Pain
##           LR Chisq Df Pr(>Chisq)
## Age        10.7246  1   0.001057 **
## Treatment  12.9063  2   0.001576 **
## Sex         3.7875  1   0.051637 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 6

```r
# Modèle avec Duration supprimé

reglNeuragliaTest <- glm(Pain ~ . - Duration, family = binomial(link = "logit"), data = NeuralgiaApp)
prediction <- predict(reglNeuragliaTest, newdata = NeuralgiaTest, type = 'response')

# Matrice de confusion

table(prediction > 0.5, NeuralgiaTest$Pain)
```

```
##
##         0 1
##   FALSE 7 0
##   TRUE  1 4
```

```r
# Modèle avec Duration et Sex supprimés

reglNeuragliaTestBis <- glm(Pain ~ . - Duration - Sex, family = binomial(link = "logit"), data = Neural
predictionBis <- predict(reglNeuragliaTestBis, newdata = NeuralgiaTest, type = 'response')

table(predictionBis > 0.5, NeuralgiaTest$Pain)
```

```
##
##         0 1
##   FALSE 7 0
##   TRUE  1 4
```

Question 7

```r
set.seed(123)
n_sim <- 50
erreurs_totales <- 0

for (i in 1:n_sim) {
u2 <- sample(1:60, 48)
NeuralgiaApp2 <- Neuralgia[u2, ]
NeuralgiaTest2 <- Neuralgia[-u2, ]

reglNeuragliaTest2 <- glm(Pain ~ . - Duration, family = binomial(link = "logit"), data = NeuralgiaApp2)
prediction2 <- predict(reglNeuragliaTest2, newdata = NeuralgiaTest2, type = 'response')

confusion <- table(prediction2 > 0.5, NeuralgiaTest2$Pain)

erreurs_totales <- erreurs_totales + confusion[2,1] + confusion[1,2]
}

# Moyenne des erreurs

moyenne_erreurs <- erreurs_totales / n_sim
moyenne_erreurs
```

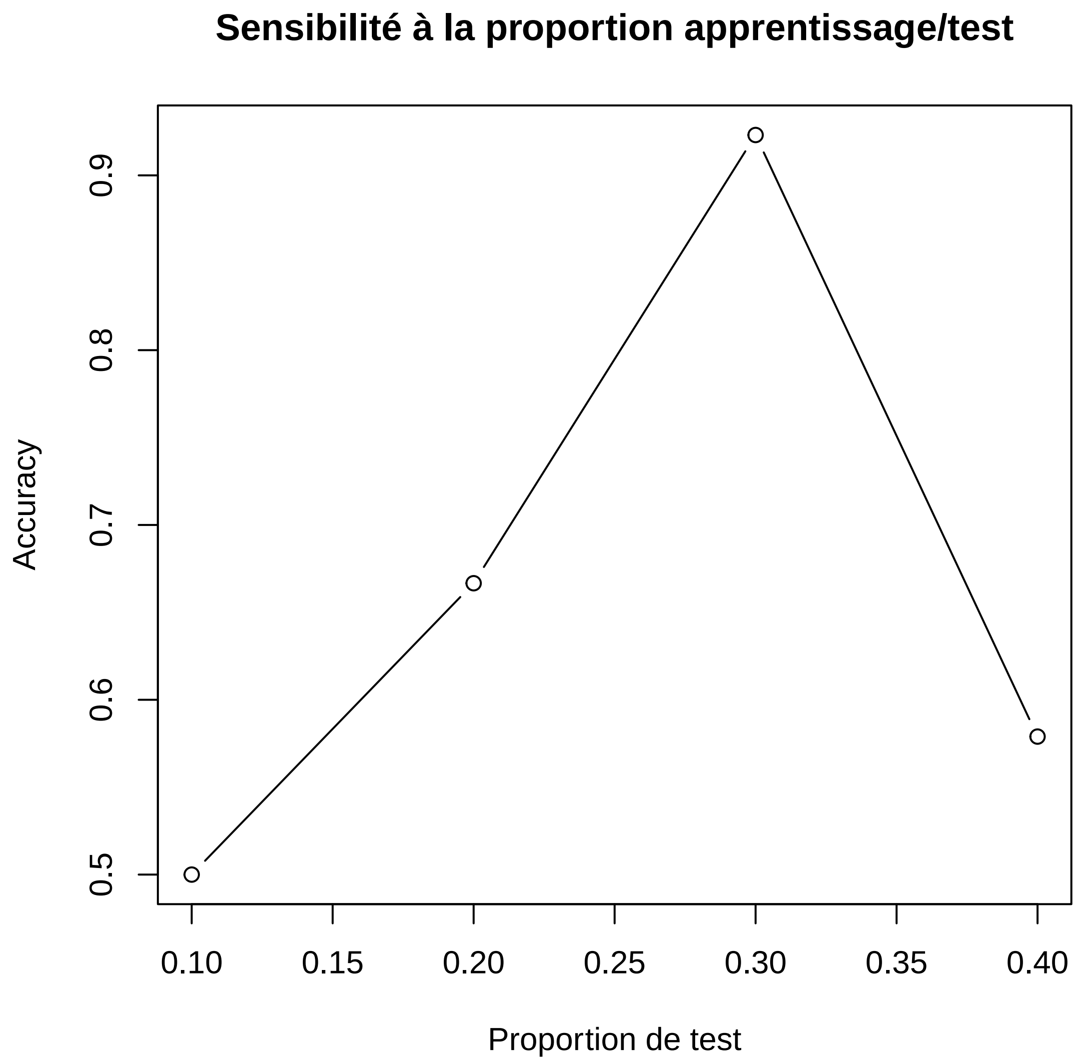
```
## [1] 2.6
```

```r
set.seed(123)
proportions <- c(0.1, 0.2, 0.3, 0.4)  # Proportions pour l'ensemble de test
results_prop <- data.frame(proportion = numeric(), accuracy = numeric())
n = 60
for (prop in proportions) {
  test_size <- round(prop * n)
  train_indices <- sample(1:n, n - test_size)
```

```r
  train <- NeuralgiaApp[train_indices, ]
  test <- NeuralgiaApp[-train_indices, ]

  # Entraînement du modèle
  model <- glm(Pain ~ Treatment + Age, data = train, family = binomial)

  # Prédictions et calcul de l'accuracy
  predictions <- predict(model, newdata = test, type = "response")
  predicted_class <- ifelse(predictions > 0.5, 1, 0)
  accuracy <- mean(predicted_class == test$Pain)

  # Stockage des résultats
  results_prop <- rbind(results_prop, data.frame(proportion = prop, accuracy = accuracy))
}

# Affichage des résultats
print(results_prop)
```

```
##   proportion  accuracy
## 1        0.1 0.5000000
## 2        0.2 0.6666667
## 3        0.3 0.9230769
## 4        0.4 0.5789474
```

```r
plot(results_prop$proportion, results_prop$accuracy, type = "b",
     xlab = "Proportion de test", ylab = "Accuracy",
     main = "Sensibilité à la proportion apprentissage/test")
```


Sensibilité à la proportion apprentissage/test
Accuracy
0.5
0.6
0.7
0.8
0.9
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Proportion de test