TP_économétrie_fish_dataset

Aymeric COUPRIE

08/01/2022

https://github.com/aymericCOUPRIE/Fish_econometrie.git
(https://github.com/aymericCOUPRIE/Fish_econometrie.git)

Peut-on prédire si les poissons appartiennent à l'espèce étudiée en fonction de leur mensuration ?

Choix des technologies

Pour faire ce sujet nous avions la possibilité de choisir le langage de programmation souhaité. J'ai donc choisis d'utiliser R afin de mettre en pratique ce que nous avons vu en cours.

Récupération des données

Ici, on a un jeu de données sous format csv, qu'on va lire et importer en R, en utilisant le délimiteur ";" et en gardant les en-tête des colonnes.

```
# Lecture du csv
fish_dataset <- read.table("Fish.csv", header = TRUE, sep = ";")
```

Représentation des données

Cette phase permet d'avoir quelques informations sur le dataset que l'on vient d'importer. On a donc le nombre de variables contenues dans le dataset. Dans un deuxième temps on affiche les n premières lignes (5 par défaut) du dataset afin d'avoir un aperçu générales des données contenues.

```
str(fish_dataset)
```

```
## 'data.frame':    111 obs. of  4 variables:
##  $ Species: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Weight : num  242 290 340 363 430 450 500 390 450 500 ...
##  $ Height : num  11.5 12.5 12.4 12.7 12.4 ...
##  $ Width  : num  4.02 4.31 4.7 4.46 5.13 ...
```

```
head(fish_dataset)
```

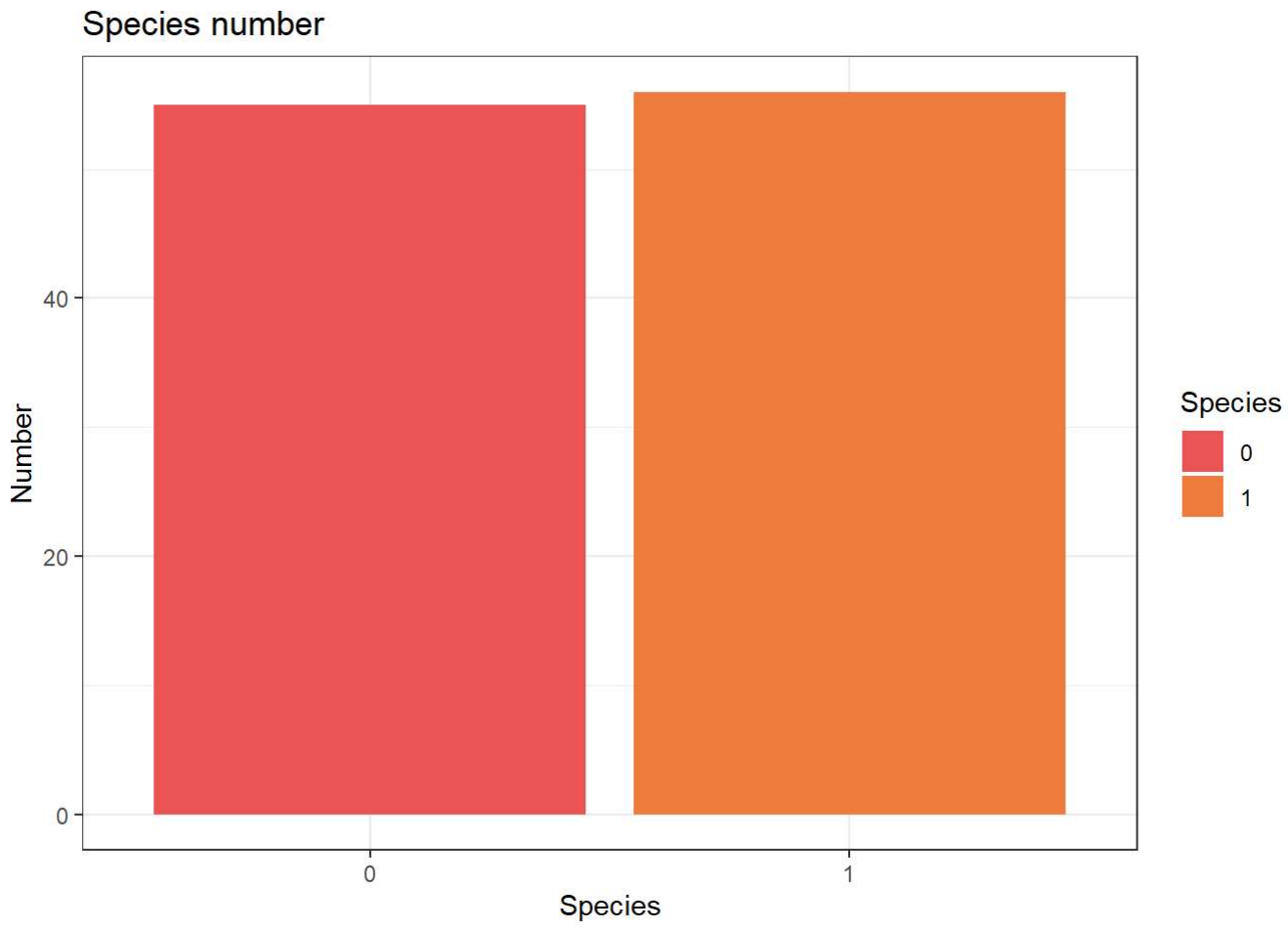
	Species <int>	Weight <dbl>	Height <dbl>	Width <dbl>
1	0	242	11.5200	4.0200

	Species <int>	Weight <dbl>	Height <dbl>	Width <dbl>
2	0	290	12.4800	4.3056
3	0	340	12.3778	4.6961
4	0	363	12.7300	4.4555
5	0	430	12.4440	5.1340
6	0	450	13.6024	4.9274

6 rows

Pour continuer la visualisation rapide des données on va afficher un graphe qui permet de séparer les 2 espèces et d'avoir le nombre de poissons pour chacune.

```
# Graphe répartion des effectifs des espèces
ggplot(fish_dataset, aes(x = as.factor(Species))) +
  geom_bar(aes(fill = as.factor(Species))) +
  scale_fill_manual(values = colors) +
  xlab("Species") +
  ylab("Number") +
  ggtitle("Species number") +
  labs(fill = "Species")
```



Séparation des données

Lors de cette phase on sépare le dataset en 2 : On décompose les données en un échantillon d'apprentissage utilisé pour apprendre le modèle contenant 70% des données et un échantillon de test tester les performances en prédiction du modèle (et sa capacité de généralisation) comprenant les 30% des données restantes.

```
# taille de l'échantillon
n <- nrow(fish_dataset)

train_index <- sample(x = 1:n, size = round(0.7 * n), replace = FALSE)

# Répartion du dataset de base
train_dataset <- fish_dataset[train_index,]
test_dataset <- fish_dataset[-train_index,]
```

Training du model

Pour la prochaine étape, on va essayer de prédire à quelle espèce le poisson étudiée appartient.

Régression backward

Ici on tente d'améliorer le modèle en partant du modèle complet puis en essayant de retirer des colonnes qui pourraient fausser la précision du modèle.

```
# Apprentissage
log_reg2 <- glm(Species ~ ., data = train_dataset, family="binomial")
log_reg2 <- step(log_reg2, direction="backward")
```

```
## Start:  AIC=44.32
## Species ~ Weight + Height + Width
##
##          Df Deviance     AIC
## - Width   1   36.992  42.992
## <none>        36.315  44.315
## - Weight  1   46.226  52.226
## - Height  1  104.077 110.077
##
## Step:  AIC=42.99
## Species ~ Weight + Height
##
##          Df Deviance     AIC
## <none>        36.992  42.992
## - Weight  1   90.190  94.190
## - Height  1  107.545 111.545
```

```
summary(log_reg2)
```

```
##
## Call:
## glm(formula = Species ~ Weight + Height, family = "binomial",
##     data = train_dataset)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.45605  -0.05353  -0.00017   0.26425   1.83670
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 20.37461    6.21790   3.277  0.00105 **
## Weight       0.05824    0.01905   3.057  0.00224 **
## Height      -4.50418    1.39249  -3.235  0.00122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 107.926  on 77  degrees of freedom
## Residual deviance:  36.992  on 75  degrees of freedom
## AIC: 42.992
##
## Number of Fisher Scoring iterations: 8
```

```
# Prédiction
hat_pi2 <- predict(log_reg2, newdata = test_dataset, type = "response")
hat_y2 <- as.integer(hat_pi2 > 0.5)
```

Régression forward

C'est le même principe que la sélection backward sauf que cette fois ci. On part du modèle vide et on ajoute les colonnes une à une afin de terminer avec le modèle complet et ensuite on regarde quel combinaison de colonnes offre le modèle le plus performant.

```
# Apprenstissage
log_reg3 <- glm(Species ~ 1, data = train_dataset, family="binomial")
log_reg3 <- step(log_reg3, direction="forward", scope=list(lower=log_reg3, upper=~Weight+Height+
Width))
```

```
## Start:  AIC=109.93
## Species ~ 1
##
##          Df Deviance    AIC
## + Height  1    90.19  94.19
## <none>        107.93 109.93
## + Weight  1   107.55 111.55
## + Width   1   107.92 111.92
##
## Step:  AIC=94.19
## Species ~ Height
##
##          Df Deviance    AIC
## + Weight  1   36.992 42.992
## + Width   1   46.226 52.226
## <none>        90.190 94.190
##
## Step:  AIC=42.99
## Species ~ Height + Weight
##
##         Df Deviance    AIC
## <none>       36.992 42.992
## + Width  1   36.315 44.315
```

```
summary(log_reg3)
```

```
##
## Call:
## glm(formula = Species ~ Height + Weight, family = "binomial",
##     data = train_dataset)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.45605  -0.05353  -0.00017   0.26425   1.83670
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 20.37461    6.21790   3.277  0.00105 **
## Height      -4.50418    1.39249  -3.235  0.00122 **
## Weight       0.05824    0.01905   3.057  0.00224 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 107.926  on 77  degrees of freedom
## Residual deviance:  36.992  on 75  degrees of freedom
## AIC: 42.992
##
## Number of Fisher Scoring iterations: 8
```

```
# Prédiction
hat_pi3 <- predict(log_reg3, newdata = test_dataset, type = "response")
hat_y3 <- as.integer(hat_pi3 > 0.5)
```

On peut voir que la sélection backward et forward donnent le même résultat. Afin d'obtenir un meilleur score il ne faut sélectionner que les colonnes Height & Weight.

Matrices de confusion

Matrice de confusion pour le système de régression avec sélection des bonnes colonnes. Ici on se sert du résultat de la sélection backward mais utiliser la sélection forward aurait abouti au même résultat

```
result <- table(hat_y2, test_dataset$Species)
result
```

```
##
## hat_y2  0  1
##      0 13  1
##      1  1 18
```

```
#Accuracy
accuracy <- round((result[1] + result[4]) / sum(result), 4)

# Matrice de confusion
confusionMatrix(data = as.factor(hat_y2), reference = as.factor(test_dataset$Species), positive
 = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 13  1
##          1  1 18
##
##                Accuracy : 0.9394
##                  95% CI : (0.7977, 0.9926)
##     No Information Rate : 0.5758
##     P-Value [Acc > NIR] : 3.82e-06
##
##                   Kappa : 0.8759
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9474
##             Specificity : 0.9286
##          Pos Pred Value : 0.9474
##          Neg Pred Value : 0.9286
##              Prevalence : 0.5758
##          Detection Rate : 0.5455
##    Detection Prevalence : 0.5758
##       Balanced Accuracy : 0.9380
##
##        'Positive' Class : 1
##
```

Grâce à ce modèle on trouve une accuracy de 0.9394.
- 18 : vrais positifs (espèce 1, classé 1)
- 13 : faux négatifs (espèce 0, classé 1)
- 1 : faux positifs (espèce 1, classé 0)
- 1 : vrais négatifs (espèce 0, classé 0)

On peut voir que le ratio de positifs est plutot bon, de meme qu'il y a peu de négatifs. Le modèle semble donc bien fonctionner

Remarque

En raison du nombre de données assez faible, on peut avoir des résultats qui varient selon la répartition des données entre le dataset de train et de test. En effet l'accuracy finale peut varier énormément en raison d'une approche parfois différente, selon la répartition des données, il peut notement s'avérer que la colonne Weight soit utile.