Filière ingénieure - Parcous MI

Fouille de Données T.P. N^o 5

Régression Logistique

Etude pratique avec R

17 novembre 2023

Etude du dataset PimaIndiansDiabetes2

1. **Import des packages** :
   (a) **dplyr** est une extension facilitant le traitement et la manipulation de données
   (voir `https://larmarange.github.io/analyse-R/manipuler-les-donnees-avec-d` `html`)
   (b) **funModeling** permettra d'utiliser la fonction ggplot

   Par ailleurs, les fonctions de dplyr sont en général plus rapides que leur équivalent sous R de base, elles permettent donc de traiter des données de grande dimension.

```
library(dplyr)
library(funModeling)
```

2. **Suppression des données manquantes** :

```
data("PimaIndiansDiabetes2", package = "mlbench")
PimaIndiansDiabetes2 <- na.omit(PimaIndiansDiabetes2)
```

3. **Devision en train et test**

```
set.seed(123)
training.samples <- sample(x = nrow(PimaIndiansDiabetes2),
 size = nrow(PimaIndiansDiabetes2)*0.8)
train.data <- PimaIndiansDiabetes2[training.samples,]
test.data <- PimaIndiansDiabetes2[-training.samples,]
```

Exercice 1 Inspectez les données !

4. **Modèle de regression logistique avec une seule variable**

```
model <- glm( diabetes ~ glucose, data = train.data, family = binomial)
summary(model)$coef
```

Voilà la signification des colonnes :

Estimate : the intercept (b0) and the beta coefficient estimates associated to each predictor variable

Std.Error : the standard error of the coefficient estimates. This represents the accuracy of the coefficients. The larger the standard error, the less confident we are about the estimate.

z value : the z-statistic, which is the coefficient estimate (column 2) divided by the standard error of the estimate (column 3)

Pr(>|z|) : the p-value corresponding to the z-statistic. The smaller the p-value, the more significant the estimate is.

Exercice 2 Quelle sont les fonctions Logit et Logistique ?

5. **Visualisation du modèle**

```
train.data %>%
  mutate(prob = ifelse(diabetes == "pos", 1, 0)) %>%
  ggplot(aes(glucose, prob)) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  labs(
    title = "Logistic Regression Model",
    x = "Plasma Glucose Concentration",
    y = "Probability of being diabete-pos"
  )
```

6. **Prédiction de la classe à partir de deux valeurs de glucose**

```
newdata <- data.frame(glucose = c(20,  180))
logit <-predict(model, newdata)
logit
probabilities <- predict(model, newdata, type = "response")
probabilities
predicted.classes <- ifelse(probabilities > 0.5, "pos", "neg")
predicted.classes
```

Par défaut la fonction predict retourne la valeur de la fonction logit. Pour avoir les probabilités il faut ajouter l'option type = "response"

Exercice 3 Ajuster maintenant un modèle de regression logistique en prenant en compte toutes les variables, observer les coefficients.

Exercice 4 Trier les coefficients selon les valeurs de p-value, pour pouvoir choisir un sous-ensemble nous allons appliquer la méthode de stepwise décrite dans `https://www.statology.org/stepwise-regression-r/`

La fonction step avec direction 'forward' permet d'ajouter à chaque étape la variable qui améliore au mieux le modèle selon le critère d'information d'Akaike.
Le critère d'information d'Akaike, (en anglais Akaike information criterion ou AIC) est une mesure de la qualité d'un modèle statistique. Ce critère repose donc sur un compromis entre la qualité de l'ajustement et la complexité du modèle, en pénalisant les modèles ayant un grand nombre de paramètres, ce qui limite les effets de sur-ajustement. Ce critère est donné par $AIC = 2K - 2ln(L)$, k étant le nombre de paramètres et L est le maximum de la fonction de vraisemblance du modèle.

7. **Selection des variables**

```
#Ajuster un modèle avec la biais seulement
intercept_only<-glm( diabetes ~1, data = train.data, family = binomial)
intercept_only$coef

# Appliquer la méthode step en direction forward
forward <- step(intercept_only, direction='forward',
 scope=formula(model), trace=0)

#Afficher les résultats
forward$anova
```

```
#Afficher les coefficients retenus
forward$coefficients
```

8. **Visualisation de l'évolution de l'AIC**

```
plot(0:(nrow(forward$anova)-1),forward$anova[,"AIC"],type="b",
xlab="# de var. introdutes",ylab="AIC",main="Sélection forward (AIC)")
```

Exercice 5 Ajuster maintenant un modèle avec les coefficients retenus.

Exercice 6 Donner la matrice de confusion sur l'ensemble de test.