BE2- plans d'expériences et régression logistique

Tulio NAVARRO TUTUI, Filipe PENNA CERAVOLO SOARES

26 October, 2022

Exercice 01 - Régression logistique

1. On considère le modèle de régression logistique en l'absence de covariables

```
# tuyns = read.table(file = "./02. Segundo BE/02. Aula/tuyns.txt",header = TRUE)
tuyns = read.table(file = "tuyns.txt",header = TRUE)
head(tuyns)
```

```
##   Ncas NTem Strate TAB ALC cancer
## 1    0   43      1   1   1      0
## 2    0   21      1   1   2      0
## 3    0    6      1   1   3      0
## 4    0    2      1   1   4      0
## 5    0   18      1   2   1      0
## 6    0   15      1   2   2      0
```

À la base de données, on verifie que les lignes correspondent à patients et la variable d'interêt "cancer". Elle possède 200 valeurs "1" qu'on interprete comme des personnes que n'ont pas de cancer et 315 valeurs "0" qu'on interprete comme des personnes qu'ont cancer.

1.1. À quoi correspond le coefficient α ?

La fonction Logit est la fonction définie par : $p \to log(p/1-p) = \alpha$. Où p est la probabilité de success du modèle.

1.2. Comment s'interprète-t-il ?

```
logistic_model = glm(cancer ~ 1, family=binomial(link="logit"), data=tuyns)
summary(logistic_model)
```

```
##
## Call:
## glm(formula = cancer ~ 1, family = binomial(link = "logit"),
##     data = tuyns)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -0.9916  -0.9916  -0.9916   1.3754   1.3754
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.45426    0.09041  -5.024 5.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 688.04  on 514  degrees of freedom
## Residual deviance: 688.04  on 514  degrees of freedom
## AIC: 690.04
##
## Number of Fisher Scoring iterations: 4
```

```
predict = exp(predict(logistic_model, newdata = tuyns))/(1+exp(predict(logistic_model, newdata=tuyns)))
results = table(predict > 0.5, tuyns$cancer)
results
```

```
##
##           0   1
##   FALSE 315 200
```

```
exp(logistic_model$coefficients[1])/(1+exp(logistic_model$coefficients[1]))
```

```
## (Intercept)
##   0.3883495
```

Si on construit une modèle en l'absence de covariables, on fait un modèle 0R, cet-à-dire la prévision sera qu'une personne n'a pas de cancer, car la majorité de l'échantillon n'a pas de cancer.

2. Etudier l'effet de la variable TAB sur la survenue du cancer.

2.1 Donner l'équation du modèle et interpréter l'intercept dans ce cas.

Dans ce cas, l'équation du modèle est donné par:

$\pi = \frac{e^{\beta_0 + x\beta_1}}{1+e^{\beta_0 + x\beta_1}}$

2.2 Mettre en oeuvre le modèle sur R.

```
tuyns$TAB = as.factor(tuyns$TAB)
logistic_model_1 = glm(cancer ~ TAB, family=binomial(link="logit"), data=tuyns)
```

2.3 Interpréter le résultat de la routine Anova.

```r
library(car)
```

```
## Loading required package: carData
```

```r
anova(logistic_model_1,test="Chisq") # test du Chi carré
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cancer
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  514     688.04
## TAB   3   36.549       511     651.50 5.732e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(logistic_model_1,test.statistic = "LR", type= 'III') # test de maximum de vraisemblance
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cancer
##     LR Chisq Df Pr(>Chisq)
## TAB   36.549  3  5.732e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(logistic_model_1,test.statistic = "Wald", type= 'III') # test de Wald-Wolfowitz
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cancer
##             Df  Chisq Pr(>Chisq)
## (Intercept)  1 43.123  5.140e-11 ***
## TAB          3 32.299  4.526e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On peut observer que en utilisant n'importe quel des méthodes au-dessus (Chi-carré, maximum de vraisemblance ou Wald-Wolfowitz), la variable TAB possède une significance au modèle.

2.4 Interpréter le résultat du summary, on regardera si tous les niveaux d'exposition ont une influence significative.

```
summary(logistic_model_1)
```

```
##
## Call:
## glm(formula = cancer ~ TAB, family = binomial(link = "logit"),
##     data = tuyns)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2338  -1.0469  -0.6541   1.1221   1.8150
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.4333     0.2183  -6.567 5.14e-11 ***
## TAB2          1.1182     0.2823   3.962 7.45e-05 ***
## TAB3          1.5648     0.2774   5.641 1.69e-08 ***
## TAB4          1.0639     0.2892   3.679 0.000234 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 688.04  on 514  degrees of freedom
## Residual deviance: 651.50  on 511  degrees of freedom
## AIC: 659.5
##
## Number of Fisher Scoring iterations: 4
```

On verifie que tous les categories de TAB sont relevant pour le modèle, en prenant en compte $\alpha = 5$.

3. Recommencer les étapes précédentes en étudiant l'effet de la variable ALC sur la survenue du cancer. Quelle est la variable la plus influente ALC ou TAB ?

```
tuyns$ALC = as.factor(tuyns$ALC)
logistic_model_2 = glm(cancer ~ ALC, family=binomial(link="logit"), data=tuyns)
```

2.3 Interpréter le résultat de la routine Anova.

```
library(car)
anova(logistic_model_2,test="Chisq") # test du Chi carré
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
```

```
## Response: cancer
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                   514     688.04
## ALC   3   37.202       511     650.84 4.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(logistic_model_2,test.statistic = "LR", type= 'III') # test de maximum de vraisemblance
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cancer
##     LR Chisq Df Pr(>Chisq)
## ALC   37.202  3   4.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(logistic_model_2,test.statistic = "Wald", type= 'III') # test de Wald-Wolfowitz
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cancer
##             Df  Chisq Pr(>Chisq)
## (Intercept)  1 41.152  1.408e-10 ***
## ALC          3 33.688  2.306e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On peut observer que en utilisant n'importe quel des méthodes au-dessus (Chi-carré, maximum de vraisemblance ou Wald-Wolfowitz), la variable ALT possède une significance au modèle.

2.4 Interpréter le résultat du summary, on regardera si tous les niveaux d'exposition ont une influence significative.

```
summary(logistic_model_2)
```

```
##
## Call:
## glm(formula = cancer ~ ALC, family = binomial(link = "logit"),
##     data = tuyns)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2933  -1.0059  -0.6762   1.3357   1.7820
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3591     0.2119  -6.415 1.41e-10 ***
## ALC2          0.9414     0.2750   3.424 0.000617 ***
## ALC3          0.9945     0.2807   3.543 0.000395 ***
## ALC4          1.6274     0.2808   5.796 6.77e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 688.04  on 514  degrees of freedom
## Residual deviance: 650.84  on 511  degrees of freedom
## AIC: 658.84
##
## Number of Fisher Scoring iterations: 4
```

On verifie que tous les categories de ALC sont relevant pour le modèle, en prenant en compte $\alpha = 5$, surtout ALC4 qui a une relaction plus important que les autres categories.

4. Proposer un modèle complet avec interaction. Interpréter les résultats de ce modèle. Simplifier éventuellement le modèle et calculer l'Odds pour une population ALC = 1 TAB = 2. Donner la matrice de confusion de ce modèle

On propose une procédure AIC forward pour construire an ajoutant une paramètre à la fois. Celle qui est ajoutée est celle qui minimisera le critère AIC.

```
next_step <- step(logistic_model, direction="forward", scope=list(upper=~(TAB + ALC)), trace = TRUE)
```

```
## Start:  AIC=690.04
## cancer ~ 1
##
##        Df Deviance    AIC
## + ALC   3   650.84 658.84
## + TAB   3   651.50 659.50
## <none>      688.04 690.04
##
## Step:  AIC=658.84
## cancer ~ ALC
##
##        Df Deviance    AIC
## + TAB   3   611.84 625.84
## <none>      650.84 658.84
##
## Step:  AIC=625.84
## cancer ~ ALC + TAB
```

```
next_step$anova
```

```
##    Step Df Deviance Resid. Df Resid. Dev      AIC
## 1       NA       NA       514   688.0442 690.0442
## 2 + ALC -3 37.20228       511   650.8419 658.8419
## 3 + TAB -3 39.00549       508   611.8364 625.8364
```

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

```
## Analysis of Deviance Table (Type III tests)
##
## Response: cancer
##     LR Chisq Df Pr(>Chisq)
## ALC   39.659  3  1.259e-08 ***
## TAB   39.005  3  1.731e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cette procédure confirme ce qu'on a observé dans les exercices précedents, car les deux variables diminuent le critère AIC et sont donc interessantes d'êtres ajoutéesau modèle.

```
predict = exp(predict(next_step, newdata = tuyns))/(1+exp(predict(next_step, newdata=tuyns)))
table(predict > 0.5, tuyns$cancer)
```

```
##
##           0   1
##   FALSE 254 102
##   TRUE   61  98
```

Exercice 02 - Plan d'expériences - Surface de réponse

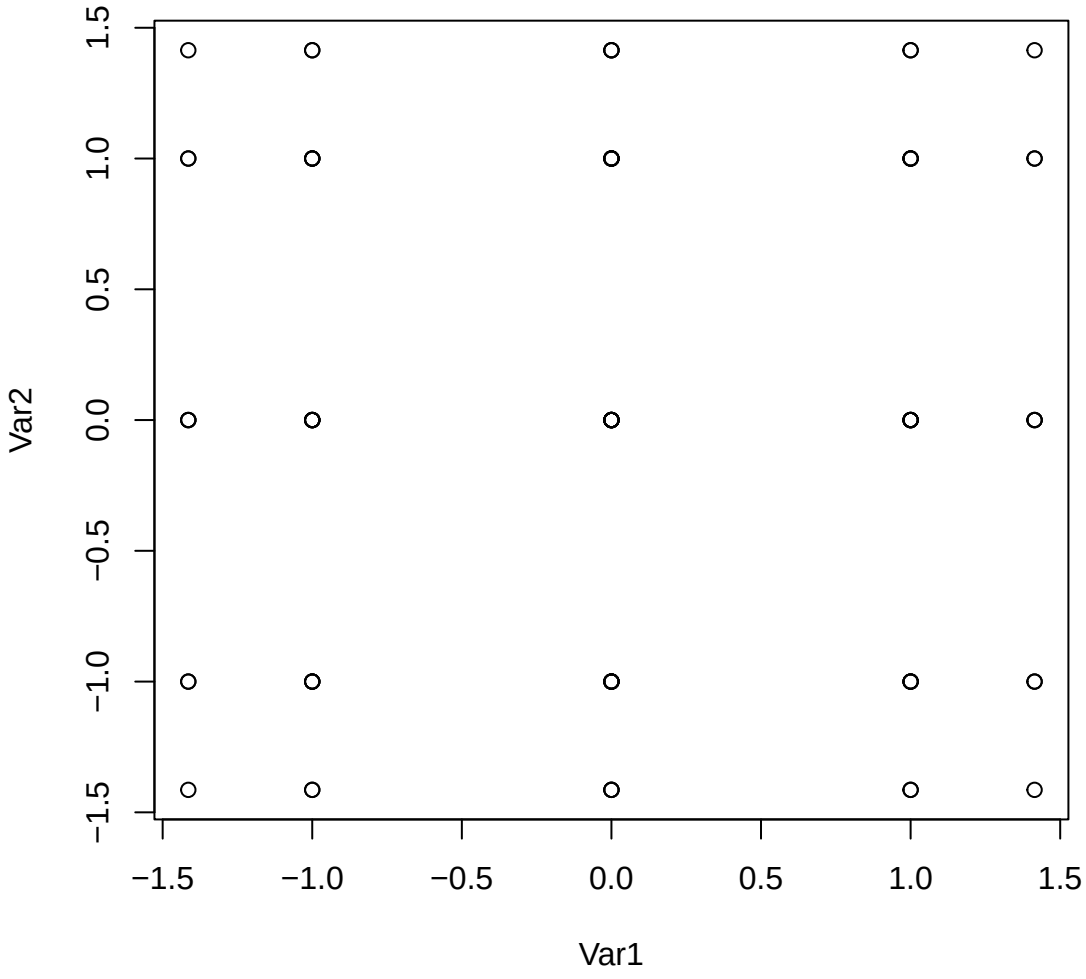
```
hormones = read.table(file = "hormones.txt", header = T, dec = ",")
# hormones = read.table(file = "hormones.txt", header = T, dec = ",")
head(hormones)
```

```
##      X1 X2    Y1    Y2
## 1  50.0 15  7.52  8.12
## 2 120.0 15 12.37 11.84
## 3  50.0 25 13.55 12.35
## 4 120.0 25 16.48 15.32
## 5  35.5 20  8.63  9.44
## 6 134.5 20 14.22 12.57
```

1. Représenter graphiquement le plan d'expériences en 2D. Il s'agit d'un plan composé des 4 sommets du carré, de 4 points sur les axes et du centre du domaine. On appelle ce plan un plan composite. Il permet d'estimer un modèle comprenant les facteurs principaux, les termes d'interactions et les termes carrés. Executer les commandes suivantes pour revenir aux variables adimentionnées.

```r
hormones$X1 = (hormones$X1- 85)/ 35
hormones$X2 = (hormones$X2- 20)/ 5

# plot(c(0, hormones$X1), c(hormones$X2, 0), col='blue')
plot(expand.grid(hormones$X1, hormones$X2))
```



2. Faire l'analyse sur le premier groupe d'enfant.

```r
# hormones = read.table(file = "./02. Segundo BE/02. Aula/hormones.txt", header = T, dec = ",")
mod1 = lm(Y1 ~ X1 + X2 + X1:X2 + I(X1^2) + I(X2^2), data = hormones)
summary(mod1)
```

```
##
```

```
## Call:
## lm(formula = Y1 ~ X1 + X2 + X1:X2 + I(X1^2) + I(X2^2), data = hormones)
##
## Residuals:
##         1         2         3         4         5         6         7
##  0.6019590  0.5706947  0.0995525  0.0682883 -0.3571954 -0.3129834 -0.6905335
##         8         9
##  0.0200838  0.0001341
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.7299     0.6844  22.982  0.00018 ***
## X1            1.9606     0.2420   8.102  0.00393 **
## X2            2.7862     0.2420  11.513  0.00141 **
## I(X1^2)      -1.9847     0.4012  -4.946  0.01585 *
## I(X2^2)      -1.6003     0.4014  -3.987  0.02825 *
## X1:X2        -0.4800     0.3422  -1.403  0.25531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6844 on 3 degrees of freedom
## Multiple R-squared:  0.9869, Adjusted R-squared:  0.965
## F-statistic: 45.16 on 5 and 3 DF,  p-value: 0.005037
```

On observe que les hormones isolés possèdent des p-values qui satisfaissent un $\alpha = 5$. Par contre, le modèle complét en considérant leur intéraction ne satisfait pas cette hypothèse.

4. Quel gain de taille peut on attendre avec le traitement suivant : X1 = 100 ppm et X2 = 20 ppm ? Donner un intervalle de prédiction.

```
mod2 = lm(Y1 ~ X1 + X2, data = hormones)
new_data = data.frame(X1 = 100, X2 = 20)

predict = predict(mod1, newdata = new_data, interval="prediction", level=0.95)
predict
```

```
##         fit       lwr       upr
## 1 -21179.46 -34458.12 -7900.806
```

bonus1. Les conclusions sont elles similaires sur le deuxième groupe d'enfants ?

```
# hormones = read.table(file = "./02. Segundo BE/02. Aula/hormones.txt", header = T, dec = ",")
mod3 = lm(Y2 ~ X1 + X2 + X1:X2 + I(X1^2) + I(X2^2), data = hormones)
summary(mod3)
```

```
##
## Call:
## lm(formula = Y2 ~ X1 + X2 + X1:X2 + I(X1^2) + I(X2^2), data = hormones)
##
```

```
## Residuals:
##          1          2          3          4          5          6          7
##  6.450e-01  1.211e+00 -9.881e-01 -4.221e-01  2.888e-01 -5.116e-01 -1.266e+00
##          8          9
##  1.043e+00  4.459e-05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.0000     1.4230  11.947  0.00126 **
## X1            1.3895     0.5031   2.762  0.07004 .
## X2            2.7440     0.5031   5.454  0.01211 *
## I(X1^2)      -2.9415     0.8342  -3.526  0.03875 *
## I(X2^2)      -2.2624     0.8345  -2.711  0.07310 .
## X1:X2        -0.1875     0.7115  -0.264  0.80921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.423 on 3 degrees of freedom
## Multiple R-squared:  0.9437, Adjusted R-squared:  0.8497
## F-statistic: 10.05 on 5 and 3 DF,  p-value: 0.04313
```

Pour ce groupe des enfants, on observe un résultat différent. L'hormone X2 ne possèdent pas un p-valeur qui satisfait un $\alpha = 5$. Par contre, le modèle en considérant leur intéraction satisfait cette hypothèse, en arrivant à une valeur encore moins important de p-valeur qui celle de X1.

bonus2. Y a-t-il un effet groupe ?

D'après les résultats, on peut constater qu'il y a un effet groupe