Régression sur données non-linéaires - Atelier 1

Martin Chevalier *(INSEE)*

Cette page comporte l'ensemble des cas pratiques du module "Régression sur données non-linéaire" des 6 et 7 février et de l'atelier du 8 février 2017.

Le support de présentation est téléchargable ici et les données nécessaires aux cas pratiques ici. Les données sont issues de l'*European social survey* 2012 (*Round 6*) librement accessibles à cette adresse.

Cette page a été réalisée avec Rstudio sous Rmarkdown et compilée le 2017-03-26.

Partie 1 : Estimer un modèle linéaire dichotomique

L'*European social survey* est une enquête récurrente sur les opinions des habitants des pays européens sur un très grand nombre de sujet : politique, confiance dans les institutions, valeurs, bien-être, etc. Elle comporte à ce titre un grand nombre de variables qualitatives qui gagnent à être modélisées avec le modèle linéaire général.

Le support de ces cas pratiques est une étude sur la participation au dernier scrutin national, avec à l'appui plusieurs variables socio-démographiques classiques : âge, sexe, nombre d'années d'étude, catégorie sociale, position sur le marché du travail.

Les deux premières questions doivent vous permettre de vous approprier ces données, d'effectuer quelques manipulations élémentaires avant d'estimer un premier modèle de régression logistique dichotomique.

Question 1 Importer et explorer les données

Le fichier de données réduit est le fichier `ess_da.rds`.

a. Ouvrez **RStudio** et créez un nouvel script. Utilisez la fonction `setwd()` pour modifier le répertoire par défaut du logiciel et le faire pointer vers le répertoire contenant le fichier `ess_da.rds`. Si vous ne connaissez pas cette fonction, n'hésitez pas à chercher de l'aide à son sujet en tapant `? setwd`.

Proposition de solution

```
# La fonction setwd() permet d'indiquer au logiciel
# où chercher par défaut des données externes
setwd("//EMPLACEMENT/réseau/")
# BIEN PENSER A UTILISER DES / et NON DES \
```

b. Utilisez la fonction `readRDS()` pour lire le fichier`ess_da.rds` (appelez l'objet créé dans l'opération `ess`).

Proposition de solution

```
# La fonction readRDS() permet de lire
# des fichiers en format R natif
ess <- readRDS("ess_da.rds")
```

c. Combien cette table comporte-t-elle d'observations et de variables ? Dans combien de pays l'enquête a-t-elle été conduite et combien d'individus ont-ils été interrogés en France ?

Proposition de solution

```
# Nombre d'observations et de variables
nrow(ess)
## [1] 54673
ncol(ess)
## [1] 9
dim(ess)
## [1] 54673     9

# La variable cntry code le pays d'enquête
head(ess$cntry)
## [1] "AL" "AL" "AL" "AL" "AL" "AL"
table(ess$cntry)
##
##   AL   BE   BG   CH   CY   CZ   DE   DK   EE   ES   FI   FR   GB   HU   IE
## 1201 1869 2260 1493 1116 2009 2958 1650 2380 1889 2197 1968 2286 2014 2628
##   IL   IS   IT   LT   NL   NO   PL   PT   RU   SE   SI   SK   UA   XK
## 2508  752  960 2109 1845 1624 1898 2151 2484 1847 1257 1847 2178 1295
# 1 968 individus ont été interrogés en France.
```

Question 2 Statistiques descriptives

a. Quelle est la participation électorale moyenne selon l'enquête dans l'ensemble des pays interrogés ? Utilisez la fonction `weighted.mean()` avec le poids de sondage `weight` pour corriger les sur- et sous-représentations.

Proposition de solution

```r
# Moyenne non-pondérée : fonction mean()
mean(ess$vote)
## [1] NA
# Il semblerait qu'il y ait des valeurs manquantes
# dans la variable : on utilise l'option
# na.rm pour les écarter
mean(ess$vote, na.rm = TRUE)
## [1] 0.759883
# La participation moyenne non-pondérée est d'environ
# 76 %.

# Moyenne pondérée : fonction weighted.mean()
weighted.mean(ess$vote, ess$weight, na.rm = TRUE)
## [1] 0.7483744
# La participation moyenne non-pondérée est d'environ
# 75 %. Il y a peu de sur- ou sous-représentations
# dans l'ESS.
```

b. Quel est le taux de participation moyen aux dernières élections nationales selon les pays ? Pour répondre à cette question, vous pouvez soit approfondir la démarche de la question précédente par exemple avec un `tapply()`, soit utiliser la fonction `table()`. Comment feriez-vous pour pondérer ces résultats par le poids de sondage ?

Proposition de solution

```r
# Première solution : table()
vote_cntry <- table(ess$vote, ess$cntry)
prop.table(vote_cntry, 2)[2, ]
##        AL        BE        BG        CH        CY        CZ        DE
## 0.8461538 0.8933577 0.7383220 0.6634146 0.8031038 0.6481069 0.8203957
##        DK        EE        ES        FI        FR        GB        HU
## 0.9391711 0.6820388 0.7639047 0.8499016 0.7888514 0.7173713 0.7287582
##        IE        IL        IS        IT        LT        NL        NO
## 0.7437577 0.7772232 0.8695652 0.8056460 0.5603363 0.8366086 0.8665720
##        PL        PT        RU        SE        SI        SK        UA
## 0.6910112 0.6912210 0.6802548 0.9050179 0.7319763 0.7627778 0.7694497
##        XK
## 0.6525631

# Deuxième solution : tapply()
tapply(ess$vote, ess$cntry, mean, na.rm = TRUE)
##        AL        BE        BG        CH        CY        CZ        DE
## 0.8461538 0.8933577 0.7383220 0.6634146 0.8031038 0.6481069 0.8203957
##        DK        EE        ES        FI        FR        GB        HU
## 0.9391711 0.6820388 0.7639047 0.8499016 0.7888514 0.7173713 0.7287582
```

```
##        IE        IL        IS        IT        LT        NL        NO
## 0.7437577 0.7772232 0.8695652 0.8056460 0.5603363 0.8366086 0.8665720
##        PL        PT        RU        SE        SI        SK        UA
## 0.6910112 0.6912210 0.6802548 0.9050179 0.7319763 0.7627778 0.7694497
##        XK
## 0.6525631

# Pour pondérer, le plus simple est encore
# d'utiliser le package descr et sa fonction
# crosstab()

# INSTALLATION DU PACKAGE (à ne faire qu'une fois)
# install.packages("descr")

# CHARGEMENT DU PACKAGE (à faire à chaque session)
library(descr)
crosstab(
  ess$cntry, ess$vote, weight = ess$weight
  , prop.r = TRUE, chisq = TRUE
)
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
##
## ================================================
##              ess$vote
## ess$cntry            0           1       Total
## ------------------------------------------------
## AL              251601     1633074     1884675
##                   13.3%       86.7%        0.4%
## ------------------------------------------------
## BE              915480     7130975     8046455
##                   11.4%       88.6%        1.5%
## ------------------------------------------------
## BG             1675634     4384083     6059717
##                   27.7%       72.3%        1.2%
## ------------------------------------------------
## CH             1855911     3722958     5578869
##                   33.3%       66.7%        1.1%
## ------------------------------------------------
## CY              134788      507912      642700
##                   21.0%       79.0%        0.1%
## ------------------------------------------------
```

```
## CZ             2782404     5232855     8015259
##                   34.7%       65.3%        1.5%
## ---------------------------------------------
## DE            10977078    52028178    63005256
##                   17.4%       82.6%       12.1%
## ---------------------------------------------
## DK              247147     4004574     4251721
##                    5.8%       94.2%        0.8%
## ---------------------------------------------
## EE              302745      677779      980524
##                   30.9%       69.1%        0.2%
## ---------------------------------------------
## ES             8131543    28720055    36851598
##                   22.1%       77.9%        7.1%
## ---------------------------------------------
## FI              626649     3479508     4106157
##                   15.3%       84.7%        0.8%
## ---------------------------------------------
## FR             9934089    35942446    45876535
##                   21.7%       78.3%        8.8%
## ---------------------------------------------
## GB            14531104    33964634    48495738
##                   30.0%       70.0%        9.3%
## ---------------------------------------------
## HU             2107439     5591420     7698859
##                   27.4%       72.6%        1.5%
## ---------------------------------------------
## IE              841216     2430549     3271765
##                   25.7%       74.3%        0.6%
## ---------------------------------------------
## IL             1199655     3592959     4792614
##                   25.0%       75.0%        0.9%
## ---------------------------------------------
## IS               30018      194212      224230
##                   13.4%       86.6%        0.0%
## ---------------------------------------------
## IT            10302426    39682632    49985058
##                   20.6%       79.4%        9.6%
## ---------------------------------------------
## LT             1109684     1325755     2435439
##                   45.6%       54.4%        0.5%
## ---------------------------------------------
## NL             2050401    11078474    13128875
##                   15.6%       84.4%        2.5%
## ---------------------------------------------
```

```
## NO              507948     2946951     3454899
##                   14.7%       85.3%        0.7%
## ---------------------------------------------
## PL             9471567    21336361    30807928
##                   30.7%       69.3%        5.9%
## ---------------------------------------------
## PT             2387559     5755248     8142807
##                   29.3%       70.7%        1.6%
## ---------------------------------------------
## RU            36548067    75771261   112319328
##                   32.5%       67.5%       21.5%
## ---------------------------------------------
## SE              677187     6376705     7053892
##                    9.6%       90.4%        1.4%
## ---------------------------------------------
## SI              437187     1221889     1659076
##                   26.4%       73.6%        0.3%
## ---------------------------------------------
## SK             1166749     3220779     4387528
##                   26.6%       73.4%        0.8%
## ---------------------------------------------
## UA             9730268    27550885    37281153
##                   26.1%       73.9%        7.1%
## ---------------------------------------------
## XK              329255      891189     1220444
##                   27.0%       73.0%        0.2%
## ---------------------------------------------
## Total       131262799   390396300   521659099
## =============================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 12446456      d.f. = 28      p <2e-16
##
##         Minimum expected frequency: 56422.01
```

```r
# Une autre solution consiste à utiliser la fonction weighted.mean()
# séparément pour chaque pays composant le fichier. Pour ce faire :
# 1) On éclate le fichier avec la fonction split() par cntry
ess_cntry <- split(ess, ess$cntry)
# 2) On utilise sapply() pour appliquer systématiquement une
# fonction à ess_cntry (ici par exemple le nombre de ligne)
sapply(ess_cntry, nrow)
```

```
##   AL   BE   BG   CH   CY   CZ   DE   DK   EE   ES   FI   FR   GB   HU   IE
## 1201 1869 2260 1493 1116 2009 2958 1650 2380 1889 2197 1968 2286 2014 2628
##   IL   IS   IT   LT   NL   NO   PL   PT   RU   SE   SI   SK   UA   XK
## 2508  752  960 2109 1845 1624 1898 2151 2484 1847 1257 1847 2178 1295
# 3) On définit à la volée la fonction qui calcule la moyenne
# pondérée dans le sapply()
sapply(ess_cntry, function(x) weighted.mean(x$vote, x$weight, na.rm = TRUE))
##        AL        BE        BG        CH        CY        CZ        DE
## 0.8665015 0.8862257 0.7234798 0.6673320 0.7902784 0.6528616 0.8257752
##        DK        EE        ES        FI        FR        GB        HU
## 0.9418712 0.6912417 0.7793436 0.8473879 0.7834603 0.7003633 0.7262660
##        IE        IL        IS        IT        LT        NL        NO
## 0.7428862 0.7496868 0.8661278 0.7938899 0.5443598 0.8438251 0.8529775
##        PL        PT        RU        SE        SI        SK        UA
## 0.6925607 0.7067891 0.6746057 0.9039981 0.7364878 0.7340759 0.7390030
##        XK
## 0.7302170
```

c. Représenter le taux de participation aux dernières élections nationales par pays, par exemple en utilisant la fonction `barplot()`.

Proposition de solution

```
# On repart du résultat de la question précédente
txpart_cntry <- tapply(ess$vote, ess$cntry, mean, na.rm = TRUE)

# On trie par taux de participation croissant
txpart_cntry <- sort(txpart_cntry)

# On utilise la fonction barplot()
barplot(txpart_cntry)

# Pour faire apparaître la France
barplot(txpart_cntry, col = names(txpart_cntry) == "FR", las = 2)
```

d. Effectuez d'autres analyses descriptives sur les autres variables du fichier. En particulier, pensez à identifier les cas de valeurs manquantes.

Proposition de solution

```
# Par exemple sur le sexe
table(ess$gndr)
##
##     1     2
```

```
## 24929 29727

# L'option useNA = "always" permet d'afficher
# les valeurs manquantes
table(ess$gndr, useNA = "always")
##
##     1     2  <NA>
## 24929 29727    17
```

Question 3 Estimation d'un modèle logistique simple

Pour l'ensemble des estimations logistiques, on se restreint à la France (`cntry == "FR"`).

a. Utilisez la fonction `glm()` pour estimer sur la France le modèle logistique simple suivant :

$$vote_i = \beta_0 + \beta_1 chômage + \varepsilon_i$$

Proposition de solution

```
# On crée le fichier ess_fr pour les données
# correspondant à la France uniquement
ess_fr <- ess[ess$cntry == "FR", ]

# Pour estimer un modèle logistique avec glm()
# il faut spécifier la famille binomiale :
glm(vote ~ cho, data = ess_fr, family = binomial)
##
## Call:  glm(formula = vote ~ cho, family = binomial, data = ess_fr)
##
## Coefficients:
## (Intercept)          cho
##      1.3984      -0.9581
##
## Degrees of Freedom: 1775 Total (i.e. Null);  1774 Residual
##   (192 observations deleted due to missingness)
## Null Deviance:       1831
## Residual Deviance: 1809  AIC: 1813

# Sinon, le modèle n'est pas du tout le même !
glm(vote ~ cho, data = ess_fr)
##
```

```
## Call:  glm(formula = vote ~ cho, data = ess_fr)
##
## Coefficients:
## (Intercept)          cho
##      0.8019      -0.1936
##
## Degrees of Freedom: 1775 Total (i.e. Null);  1774 Residual
##   (192 observations deleted due to missingness)
## Null Deviance:       295.8
## Residual Deviance: 291.6     AIC: 1837

# En revanche, l'option link = "logit" dans
# la fonction binomial() est optionnelle
# (car définie ainsi par défaut)
glm(vote ~ cho, data = ess_fr, family = binomial(link = "logit"))
##
## Call:  glm(formula = vote ~ cho, family = binomial(link = "logit"),
##     data = ess_fr)
##
## Coefficients:
## (Intercept)          cho
##      1.3984      -0.9581
##
## Degrees of Freedom: 1775 Total (i.e. Null);  1774 Residual
##   (192 observations deleted due to missingness)
## Null Deviance:       1831
## Residual Deviance: 1809  AIC: 1813

# Pour estimer un modèle probit (cf. compléments), il suffit de
# changer la fonction de lien
glm(vote ~ cho, data = ess_fr, family = binomial(link = "probit"))
##
## Call:  glm(formula = vote ~ cho, family = binomial(link = "probit"),
##     data = ess_fr)
##
## Coefficients:
## (Intercept)          cho
##      0.8485      -0.5736
##
## Degrees of Freedom: 1775 Total (i.e. Null);  1774 Residual
##   (192 observations deleted due to missingness)
## Null Deviance:       1831
## Residual Deviance: 1809  AIC: 1813
```

b. Stockez le résultat de l'estimation dans un objet et affichez davantage de détails sur le modèle que vous avez estimé avec la fonction `summary()`. Comment interprétez-vous le coefficient associé à la variable `cho` ?

Proposition de solution

```
# On stocke le réséultat de l'estimation dans
# l'objet logit1
logit1 <- glm(vote ~ cho, data = ess_fr, family = binomial)

# Fonction summary()
summary(logit1)
##
## Call:
## glm(formula = vote ~ cho, family = binomial, data = ess_fr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7995   0.6644   0.6644   0.6644   0.9970
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.39842    0.06166  22.680  < 2e-16 ***
## cho         -0.95810    0.19692  -4.865 1.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1831.0  on 1775  degrees of freedom
## Residual deviance: 1809.1  on 1774  degrees of freedom
##   (192 observations deleted due to missingness)
## AIC: 1813.1
##
## Number of Fisher Scoring iterations: 4

# Le coefficient associé à la variable cho
# est négatif : le fait d'être au chômage semble
# négativement associé à la probabilité d'être
# allé voter aux dernières élections nationales

# Par ailleurs la p-valeur du test de significativité
# du coefficient est inférieure à 0,01 : on peut
# donc rejeter l'hypothèse nulle de nullité du
# coefficient au seuil de 1 %.
```

```
# Le fait d'être au chômage est donc négativement
# et statistiquement associé à l'abstention déclarée.
# au seuil de 1 %. On ne peut pas interpréter la valeur
# du coefficient directement (cf. partie 2).
```

c. Pour chaque individu, déterminez la probabilité prédite par le modèle. Menez à bien cette opération d'abord en utilisant la fonction `predict()` (avec l'option `type = "response"`) puis manuellement (référez-vous au support du module si besoin).

Proposition de solution

```
# La fonction predict() permet de récupérer
# facilement les probabilités prédites par le modèle
pre <- predict(logit1, type = "response")
summary(pre)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6083  0.8019  0.8019  0.7889  0.8019  0.8019
table(pre)
## pre
## 0.608333333333336 0.801932367149736
##               120              1656

# Le modèle n'intégrant qu'une seule variable
# explicative dichotomique, il n'y a que deux
# probabilités prédites distinctes (une pour
# les chômeurs, une pour les autres).

# À la main, il suffit de travailler avec les éléments
# de logit1 pour recalculer les valeurs
names(logit1)
##  [1] "coefficients"      "residuals"         "fitted.values"
##  [4] "effects"           "R"                 "rank"
##  [7] "qr"                "family"            "linear.predictors"
## [10] "deviance"          "aic"               "null.deviance"
## [13] "iter"              "weights"           "prior.weights"
## [16] "df.residual"       "df.null"           "y"
## [19] "converged"         "boundary"          "model"
## [22] "na.action"         "call"              "formula"
## [25] "terms"             "data"              "offset"
## [28] "control"           "method"            "contrasts"
## [31] "xlevels"

# model.matrix(logit1) reconstitue la matrice complète
# utilisée par le modèle pour l'estimation
```

```
str(model.matrix(logit1))
##  num [1:1776, 1:2] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:1776] "21023" "21024" "21025" "21026" ...
##   ..$ : chr [1:2] "(Intercept)" "cho"
##  - attr(*, "assign")= int [1:2] 0 1

# coef(logit1) correspond aux coefficients
coef(logit1)
## (Intercept)         cho
##   1.3984157  -0.9581039

# Pour obtenir la valeur X*beta prédite, il suffit de
# multiplier ces deux composantes (avec le produit
# matriciel)
Xbeta <- model.matrix(logit1) %*% coef(logit1)
summary(Xbeta)
##        V1
##  Min.   :0.4403
##  1st Qu.:1.3984
##  Median :1.3984
##  Mean   :1.3337
##  3rd Qu.:1.3984
##  Max.   :1.3984

# Xbeta n'est pas la probabilité mais
# ce qui permet de l'obtenir en passant
# par l'inverse du logit : exp(x) / (1 + exp())
pre2 <- as.numeric(exp(Xbeta) / (1 + exp(Xbeta)))

# On compare pre et pre2
all.equal(pre, pre2)
## [1] "names for target but not for current"
# Aucune différence sinon la présence de noms
# dans l'objet pre et pas dans pre2
```

Question 4 Estimation d'un modèle logistique multiple

a. Estimez désormais un modèle plus complet intégrant en plus du chômage l'âge, le sexe et la catégorie sociale. Pensez-bien à adapter les variables catégorielles pour les intégrer au modèle.

Proposition de solution

```
# Plusieurs solutions sont disponibles pour intégrer
# des variables catégorielles dans un modèle de régression

# 1) Recoder manuellement les variables sous forme d'indicatrices
ess_fr <- within(ess_fr, {
  femme <- (gndr == 2) * 1
  agri <- (catsoc == "Agriculture et agro-alimentaire") * 1
  cadre <- (catsoc == "Cadres et prof intellectuelles") * 1
  empl <- (catsoc == "Employés") * 1
  ouv <- (catsoc == "Ouvriers") * 1
  tech <- (catsoc == "Techniciens") * 1
})

# Vérification
summary(ess_fr)
##     cntry                idno          weight            gndr
##  Length:1968        Min.   :   4   Min.   :  4226   Min.   :1.000
##  Class :character   1st Qu.:1077   1st Qu.: 17376   1st Qu.:1.000
##  Mode  :character   Median :2035   Median : 21130   Median :2.000
##                     Mean   :2068   Mean   : 27027   Mean   :1.552
##                     3rd Qu.:3069   3rd Qu.: 35199   3rd Qu.:2.000
##                     Max.   :4198   Max.   :108305   Max.   :2.000
##
##       agea            eduyrs         catsoc               cho
##  Min.   : 15.00   Min.   : 0.00   Length:1968        Min.   :0.00000
##  1st Qu.: 37.75   1st Qu.:10.00   Class :character   1st Qu.:0.00000
##  Median : 52.00   Median :12.00   Mode  :character   Median :0.00000
##  Mean   : 51.82   Mean   :12.36                      Mean   :0.07266
##  3rd Qu.: 66.00   3rd Qu.:15.00                      3rd Qu.:0.00000
##  Max.   :101.00   Max.   :30.00                      Max.   :1.00000
##                   NA's   :7
##       vote             tech             ouv              empl
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.7889   Mean   :0.2215   Mean   :0.2434   Mean   :0.2348
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :192
##      cadre             agri             femme
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :1.0000
```

```
##  Mean   :0.1905   Mean   :0.1098   Mean   :0.5523
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
```

```r
# 2) Transformer les variables en facteur
# et les intégrer tels quels à la régression
ess_fr$gndr_fac <- factor(ess_fr$gndr, labels = c("Homme", "Femme"))
ess_fr$gndr_fac <- relevel(ess_fr$gndr_fac, ref = "Homme")
ess_fr$catsoc_fac <- factor(ess_fr$catsoc)
ess_fr$catsoc_fac <- relevel(ess_fr$catsoc_fac, ref = "Techniciens")


# Estimation du modèle
logit2 <- glm(
  vote ~ cho + agea + femme + agri + cadre + empl + ouv
  , data = ess_fr, family = binomial
)
logit2b <- glm(
  vote ~ cho + agea + gndr_fac + catsoc_fac
  , data = ess_fr, family = binomial
)

# Comparaison des deux formulation du modèle
logLik(logit2)
## 'log Lik.' -861.3093 (df=8)
logLik(logit2b)
## 'log Lik.' -861.3093 (df=8)
coef(logit2)
## (Intercept)         cho        agea       femme        agri       cadre
##   0.3772656  -0.6289442   0.0264202  -0.1632706  -0.7021386   0.2041402
##        empl         ouv
##  -0.1971839  -0.6965609
coef(logit2b)
##                                 (Intercept)
##                                   0.3772656
##                                         cho
##                                  -0.6289442
##                                        agea
##                                   0.0264202
##                                gndr_facFemme
##                                  -0.1632706
## catsoc_facAgriculture et agro-alimentaire
##                                  -0.7021386
##  catsoc_facCadres et prof intellectuelles
```

```
##                                 0.2041402
##                       catsoc_facEmployés
##                                -0.1971839
##                       catsoc_facOuvriers
##                                -0.6965609

# Les modèles sont bien identiques.
```

b. Affichez et interprétez les résultats de ce modèle.

Proposition de solution

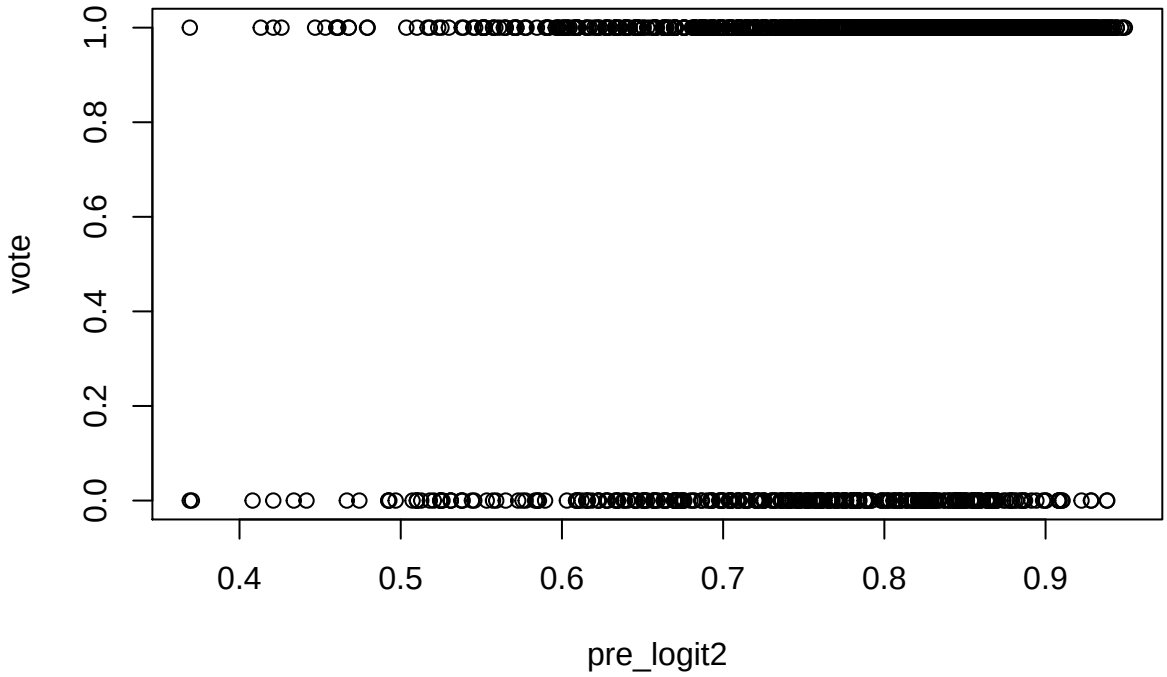
```
# Affichage des résultats avec summary()
summary(logit2)
##
## Call:
## glm(formula = vote ~ cho + agea + femme + agri + cadre + empl +
##     ouv, family = binomial, data = ess_fr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3601   0.4092   0.5719   0.7131   1.4118
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.377266   0.227235   1.660  0.09687 .
## cho         -0.628944   0.204607  -3.074  0.00211 **
## agea         0.026420   0.003568   7.405 1.31e-13 ***
## femme       -0.163271   0.126200  -1.294  0.19575
## agri        -0.702139   0.230980  -3.040  0.00237 **
## cadre        0.204140   0.202718   1.007  0.31393
## empl        -0.197184   0.181778  -1.085  0.27803
## ouv         -0.696561   0.173612  -4.012 6.02e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1831.0  on 1775  degrees of freedom
## Residual deviance: 1722.6  on 1768  degrees of freedom
##   (192 observations deleted due to missingness)
## AIC: 1738.6
##
## Number of Fisher Scoring iterations: 4
```

```
# Il se dégage de ce modèle deux ensembles de variables :
# - le fait d'être au chômage, ouvrier ou d'exercer une
# activité agricole est négativement et statistiquement
# associé avec la probabilité d'être aller voter aux
# dernières élections nationales.
# - l'âge est la seule variable qui soit positivement
# et (très) significativement associée à la participation
# au dernier scrutin national.
```

c. Utilisez la fonction `predict()` pour calculer les probabilités prédites par ce modèle. Les probabilités les plus élevées sont-elles associées aux individus ayant effectivement été voté aux dernières élections nationales ? Que pensez-vous du caractère prédictif de ce modèle ?

Proposition de solution

```
# A nouveau on utilise predict() avec l'option
# type = "response"
pre_logit2 <- predict(logit2, type = "response")

# On calcule la corrélation avec la vraie valeur
# de vote
cor(ess_fr$vote, pre_logit2)
## Error in cor(ess_fr$vote, pre_logit2): dimensions incompatibles

# En raison des non-réponses dans le modèle,
# pre_logit2 et ess_fr n'ont pas la même taille.
# On utilise les noms de pre_logit2 pour restreindre
# ess_fr
vote <- ess_fr[names(pre_logit2), "vote"]
cor(vote, pre_logit2)
## [1] 0.2461675
# La corrélation est positive mais pas très élevée

# On représente une variable en fonction de l'autre :
plot(pre_logit2, vote)
```




```
# Il n'est pas claire que les probabilités soient plus
# élevées pour les individus s'étant effectivement
# déplacés aux dernières élections nationales.

# Si l'on cherchait à prédire à partir de pre_logit2
# le fait d'être allé voter ou non, on utiliserait
# la probabilité prédite en fixant un seuil
# au-dessus duquel on "prédirait" que les individus
# se sont déplacés pour voter aux dernières élections
# nationales.

# On peut par exemple le fixer au niveau de la fréquence
# des votants dans l'échantillon. On définit ainsi
vote_predict <- (pre_logit2 > mean(vote)) * 1

# On croise vote et vote_predict pour avoir
# une idée de la qualité prédictive du modèle
table(vote, vote_predict)
##     vote_predict
## vote   0   1
##    0 235 140
##    1 516 885

# Beaucoup d'individus (environ 1/3) ne sont pas sur
# la diagonale : le modèle est peu prédictif.
```

Partie 2 : Interpréter un modèle linéaire dichotomique

Question 5 Indicateurs de qualité du modèle

a. Comparez la log-vraisemblance, l'AIC et le BIC des deux modèles estimés jusqu'à présent (le modèle avec uniquement le chômage comme variable explicative d'une part, le modèle complet d'autre part). Vérifiez que vous pouvez bien recalculer manuellement l'AIC et le BIC à partir des objets `logit1` et `logit2`.

Proposition de solution

```
# Uilisation des fonctions d'extraction de R
logLik(logit1)
## 'log Lik.' -904.5494 (df=2)
logLik(logit2)
## 'log Lik.' -861.3093 (df=8)
AIC(logit1)
## [1] 1813.099
AIC(logit2)
## [1] 1738.619
BIC(logit1)
## [1] 1824.063
BIC(logit2)
## [1] 1782.476

# Clairement logit2 est meilleur que logit1 :
# log-vraisemblance plus élevée, AIC et BIC plus faibles

# Recalculs manuels à partir de la deviance dans logit2
logit2$deviance * (-0.5)
## [1] -861.3093
logit2$deviance + 2*logit2$rank
## [1] 1738.619
logit2$deviance + log(NROW(logit2$model))*logit2$rank
## [1] 1782.476
```

b. (Optionnel) Utilisez un `sapply()` pour automatiser l'application des fonctions `logLik()`, `AIC()` et `BIC()` à autant de modèles que nécessaire. Créez la fonction `indic_perf_ll()` qui produit automatiquement le tableau d'indicateurs comparant la performance d'un nombre quelconque de modèles.

Proposition de solution

```r
# INDICATION : commencez par placer tous les modèles
# dans une même liste puis appliquez à chacun de ses
# éléments la fonction logLik() par exemple.
# Définissez une fonction à la volée dans le sapply()
# pour calculer également l'AIC et le BIC.

# CORRECTION
sapply(list(logit1, logit2), logLik)
## [1] -904.5494 -861.3093
sapply(list(logit1, logit2), function(m) {
  matrix(  c(logLik(m), AIC(m), BIC(m)), ncol = 1)
})
##           [,1]      [,2]
## [1,] -904.5494 -861.3093
## [2,] 1813.0989 1738.6186
## [3,] 1824.0631 1782.4755

# Fonction qui fait tout toute seule
indic_perf_ll <- function(...){
  result <- sapply(list(...), function(m) {
matrix(  c(logLik(m), AIC(m), BIC(m)), ncol = 1)
  })
  # Difficile : on récupère le nom des objets
  # passés dans ...
  noms <- as.character(eval(substitute(alist(...))))
  # cf. http://adv-r.had.co.nz/Computing-on-the-language.html#substitute
  dimnames(result) <- list(c("logLik", "AIC", "BIC"), noms)
  return(result)
}
indic_perf_ll(logit1, logit2)
##           logit1    logit2
## logLik -904.5494 -861.3093
## AIC    1813.0989 1738.6186
## BIC    1824.0631 1782.4755

# Uilisation des fonctions d'extraction de R
logLik(logit1)
## 'log Lik.' -904.5494 (df=2)
logLik(logit2)
## 'log Lik.' -861.3093 (df=8)
AIC(logit1)
## [1] 1813.099
AIC(logit2)
## [1] 1738.619
```

```
BIC(logit1)
## [1] 1824.063
BIC(logit2)
## [1] 1782.476
```

c. Utilisez la fonction `lrtest()` du *package* `lmtest` (à installer) pour mener à bien le test de significativité globale du modèle complet par le rapport de vraisemblance. Vérifiez que vous pouvez bien le mener "à la main" également.

Proposition de solution

```
# Installation du package lmtest (à ne faire qu'une fois)
# install.packages("lmtest")

# Chargement du package lmtest (à faire à chaque session)
library(lmtest)
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

# Utilisation de la fonction lrtest()
lrtest(logit2)
## Likelihood ratio test
##
## Model 1: vote ~ cho + agea + femme + agri + cadre + empl + ouv
## Model 2: vote ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -861.31
## 2   1 -915.48 -7 108.35  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# On rejette très largement l'hypothèse nulle.

# À la main
# 1) Calcul de la statistique de test
LR <- logit2$null.deviance - logit2$deviance
# 2) Récupération des degrés de liberté
p <- logit2$rank - 1
# 3) Calcul de la p-valeur
1 - pchisq(LR, p)
## [1] 0
```

d. Le code de la fonction `conc()` mentionné dans le support de formation figure ci-dessous. Copiez-collez le dans **R** et soumettez-le pour pouvoir utiliser la fonction.

```
conc <- function(m){
  un <- m$fitted.values[m$y == 1]
  zero <- m$fitted.values[m$y == 0]
  t <-rowSums(sapply(un, function(i){
c(sum(i > zero), sum(i < zero), sum(i == zero))
  }))
  return(c(
"Pct concordant" = t[1] * 100 / sum(t)
, "Pct discordant" = t[2] * 100 / sum(t)
, "Pct ex-aequo"= t[3] * 100 / sum(t)
  ))
}
```

Analysez le code fourni (par exemple en décomposant l'étape centrale) pour percevoir le lien avec la définition des paires discordantes et concordantes. Appliquez-la aux deux modèles estimés.

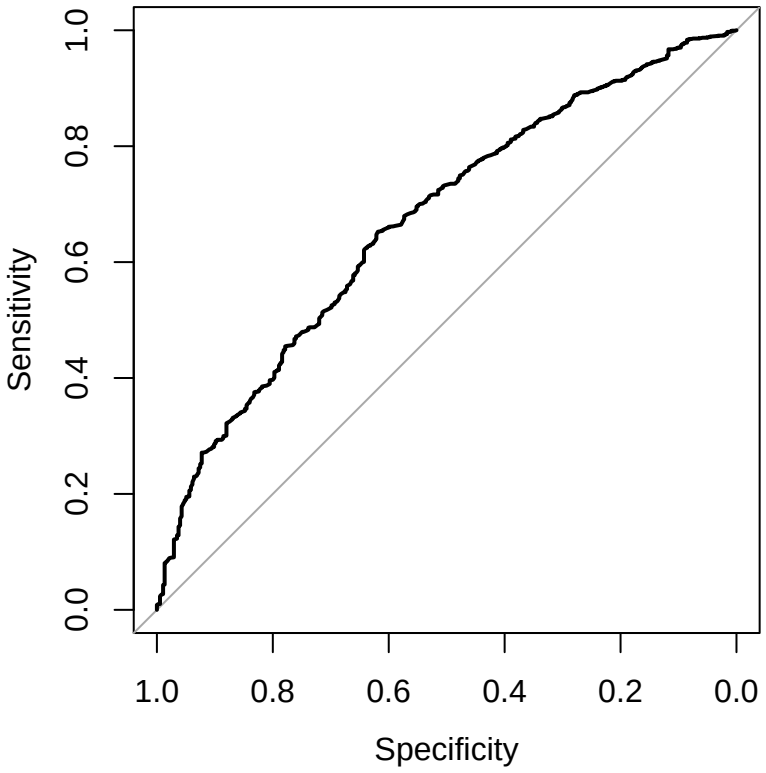
Proposition de solution

```
# un et zero sont deux vecteurs contenant respectivement
# les valeurs prédites des cas positifs (y = 1) et les
# valeurs prédites des cas négatifs (y = 0).
# L'étape centrale correspond peu ou proue à une
# double-boucle sur ces deux vecteurs : pour chaque
# élément de un, on calcule le nombre d'éléments
# qui sont concordants, discordants et ex-aequo avec lui.
# La fonction rowSums() fait enfin la somme sur tous
# les éléments de un.

# Application à logit1 et logit2
sapply(list(logit1, logit2), conc)
##                      [,1]       [,2]
## Pct concordant 11.880276 66.9674042
## Pct discordant  4.557507 32.8833690
## Pct ex-aequo   83.562217  0.1492267
# logit2 est bien plus prédictif que logit1
# du fait principalement des ex-aequo
# beaucoup moins nombreux.
```

e. Utilisez la fonction `roc()` du *package* `pROC` (à installer) pour dessiner la courbe ROC du modèle complet.

Proposition de solution

```r
# Installation du package pROC (à ne faire qu'une fois)
# install.packages("pROC")

# Chargement du package pROC (à faire à chaque session)
library(pROC)
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

# Calcul et affichage de la courbe ROC de logit2
par(pty="s") # Pour avoir une zone de tracé carrée
plot(roc(logit2$y ~ logit2$fitted.values))
```




```r
##
## Call:
## roc.formula(formula = logit2$y ~ logit2$fitted.values)
##
## Data: logit2$fitted.values in 375 controls (logit2$y 0) < 1401 cases (logit2$y 1
## Area under the curve: 0.6704
# L'aire sous la courbe apparaît dans la console : 0.6704
# Modèle intéressant, mais pas très prédictif en l'état.
```

Question 6 Odds-ratio et effets marginaux

a. Calculez les odds-ratio du modèle complet assorti de son intervalle de confiance. Comment déterminer dans ce contexte si la relation estimée par le modèle est statistiquement significative ? Interprétez la valeur de l'odds-ratio associé à la variable `ouv`.

Proposition de solution

```
# Calcul des odds-ratio
cbind(
  OR=exp(coef(logit2))
  , exp(confint.default(logit2))
)
##                      OR     2.5 %    97.5 %
## (Intercept) 1.4582915 0.9341603 2.2764981
## cho         0.5331544 0.3570192 0.7961859
## agea        1.0267723 1.0196174 1.0339774
## femme       0.8493613 0.6632399 1.0877131
## agri        0.4955244 0.3151043 0.7792482
## cadre       1.2264701 0.8243332 1.8247827
## empl        0.8210396 0.5749557 1.1724487
## ouv         0.4982961 0.3545757 0.7002707

# Comme exp(0) = 1, quand on interprète les odds-ratio
# c'est la présence de la valeur 1 dans l'intervalle
# de confiance qui traduit l'absence de significativité
# de l'association entre les variables au niveau choisi
# pour construire l'intervalle de confiance (95 % par défaut).

# Ici OR(ouv) = 0,50 et 1 n'est pas dans son IC à 95 % :
# la relation entre le fait d'appartenir à la catégorie
# sociale "ouvrier" et le vote aux dernières élections
# nationales est négative. On dira qu'en matière de vote
# aux dernières élections nationales la cote des ouvriers
# est environ 2 fois inférieure à celle des techniciens
# (la modalité de référence).
```

b. Le code de la fonction `margins()` mentionné dans le support de formation figure ci-dessous. Copiez-collez le dans **R** et soumettez-le pour pouvoir utiliser la fonction.

```
margins <- function(fit, factor = NULL){
  # fit <- m2; factor <- NULL
```

```
  if(as.character(fit$call[1]) != "glm")
stop("Le modèle doit avoir été estimé avec la fonction glm().", call. = FALSE)
  if(!(fit$family$link %in% c("identity", "log", "inverse", "logit")))
stop("La fonction de lien doit être \"identity\", \"log\", \"inverse\" ou \"logit\"
  x1 <- model.matrix(fit)
  be <- as.matrix(coef(fit))
  vcv <- vcov(fit)
  disch <- if(is.null(factor)){
temp1 <-apply(x1,2,function(x)length(table(x))==2)
names(temp1[temp1])
  }else factor
  invlink <- switch(fit$family$link
, logit = plogis, log = exp, identity = identity, inverse = (function(x) 1/x)
  )
  invlink_derivative <- switch(fit$family$link
, "logit" = dlogis, "log" = exp, "identity" = identity, inverse = (function(x) -1/x
  )
  lapply(setNames(disch, disch), function(var){
# var <- "femme"
disx0 = disx1 = x1
disx1[, var] = max(x1[, var])
disx0[, var] = min(x1[, var])
est1 <- mean(invlink(disx1 %*% be))
est0 <- mean(invlink(disx0 %*% be))
est <- est1 - est0
gr = as.numeric(invlink_derivative(disx1 %*% be)) * disx1 - as.numeric(invlink_deri
avegr = as.matrix(colMeans(gr))
se <- sqrt(t(avegr) %*% vcv %*% avegr)
matrix(c(
  est0, est1, est1 - est0
  , NA, NA,  se
  , NA, NA, est/se
  , NA, NA, 2 * pt(-abs(est/se), df = Inf)
), nrow = 3, dimnames = list(
  c(paste(var, "=", c(0, 1)), "Diff")
  , c("Average MFX", "Std. Error", "z value", "P>|z|")
))
  })
}
```

Il reprend celui des fonctions du *package* `mfx` (par exemple `logitmfx`) mais peut porter sur davantage de modèles `glm` (Gamma par exemple). En contrepartie, il ne prend en charge que les variables qualitatives du modèle (comme dans Stata).

Calculez les effets marginaux moyens dans le modèle ne comportant que la variable `chomage`. Comparez-les aux proportions relatives de votants parmi les chômeurs et les

non-chômeurs. Que remarquez-vous ?

Proposition de solution

```
# Calcul des effets marginaux avec margins()
margins(logit1)
## $cho
##         Average MFX Std. Error   z value        P>|z|
## cho = 0   0.8019324         NA        NA           NA
## cho = 1   0.6083333         NA        NA           NA
## Diff     -0.1935990 0.04562289 -4.243463 2.200971e-05

# Proportion de votant selon le chômage
tapply(ess_fr$vote, ess_fr$cho, mean, na.rm = TRUE)
##         0         1
## 0.8019324 0.6083333
# Ce sont exactement les mêmes résulats :  c'est logique
# dans la mesure où le chômage est la seule variable
# explicative du premier modèle.
```

c. Calculez les effets marginaux moyens du modèle complet. Interprétez la valeur de l'effet marginal moyen associé à la variable `ouv`.

Proposition de solution

```
# Calcul des effets marginaux moyens avec margins()
margins(logit2)
## $cho
##         Average MFX Std. Error   z value       P>|z|
## cho = 0   0.7976842         NA        NA          NA
## cho = 1   0.6855357         NA        NA          NA
## Diff     -0.1121486 0.04066601 -2.757797 0.005819237
##
## $femme
##           Average MFX Std. Error   z value     P>|z|
## femme = 0   0.8027151         NA        NA        NA
## femme = 1   0.7773778         NA        NA        NA
## Diff       -0.0253373 0.01946175 -1.301902 0.1929498
##
## $agri
##          Average MFX Std. Error   z value      P>|z|
## agri = 0   0.7983081         NA        NA         NA
## agri = 1   0.6729916         NA        NA         NA
## Diff      -0.1253165 0.04563077 -2.746316 0.00602687
##
```

```
## $cadre
##           Average MFX Std. Error  z value     P>|z|
## cadre = 0  0.78363382         NA       NA        NA
## cadre = 1  0.81447727         NA       NA        NA
## Diff       0.03084345 0.02960031 1.041997 0.2974129
##
## $empl
##          Average MFX Std. Error   z value     P>|z|
## empl = 0  0.79581785         NA        NA        NA
## empl = 1  0.76427341         NA        NA        NA
## Diff     -0.03154443 0.02976406 -1.059816 0.2892283
##
## $ouv
##         Average MFX Std. Error   z value        P>|z|
## ouv = 0   0.8165879         NA        NA           NA
## ouv = 1   0.6979453         NA        NA           NA
## Diff     -0.1186426  0.0315649 -3.758688 0.0001708069

# En moyenne dans l'échantillon, le fait d'appartenir
# à la catégorie sociale "ouvrier" plutôt que "technicien"
# est associé à une diminution de l'ordre de 12 % de la
# probabilité d'être allé voter aux dernières élections
# nationales.

# Les erreurs standard semblent assez élevées par rapport
# à ce que l'on constate pour les coefficients.
# On vérifie à partir du package mfx qu'il n'y a pas
# un problème dans le recodage de la fonction.

# Installation du package mfx (à ne faire qu'une fois)
# install.packages("mfx")

# Chargement du package mfx (à faire à chaque session)
library(mfx)
## Loading required package: sandwich
## Loading required package: MASS
## Loading required package: betareg

# Estimation avec la fonction logitmfx()
logitmfx(vote ~ cho + agea + femme + agri + cadre + empl + ouv, data = ess_fr)
## Call:
## logitmfx(formula = vote ~ cho + agea + femme + agri + cadre +
##     empl + ouv, data = ess_fr)
##
## Marginal Effects:
```

```
##              dF/dx   Std. Err.       z     P>|z|
## cho   -0.11466905  0.04232155 -2.7095 0.0067390 **
## agea   0.00413581  0.00054125  7.6413 2.151e-14 ***
## femme -0.02541660  0.01951489 -1.3024 0.1927724
## agri  -0.12911633  0.04828194 -2.6742 0.0074904 **
## cadre  0.03077063  0.02935604  1.0482 0.2945523
## empl  -0.03184728  0.03024547 -1.0530 0.2923591
## ouv   -0.12100013  0.03278502 -3.6907 0.0002236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "cho"   "femme" "agri"  "cadre" "empl"  "ouv"

# Les résultats diffèrent et sont plus crédibles !
# Il y a donc manifestement un problème dans le calcul
# des erreurs standard par margins() pour l'instant :
# cela sera corrigé au plus vite !
```

Question 7 Tests d'hypothèses complexes

a. On veut tout d'abord valider rigoureusement le fait que le modèle complet est statistiquement meilleur que le premier. Menez (par exemple avec `lrtest()`) le test du ratio de vraisemblance correspondant.

Proposition de solution

```
# Il suffit de mener le test du ratio de vraisemblance
# sur les modèles logit1 et logit2.
lrtest(logit1, logit2)
## Likelihood ratio test
##
## Model 1: vote ~ cho
## Model 2: vote ~ cho + agea + femme + agri + cadre + empl + ouv
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   2 -904.55
## 2   8 -861.31  6 86.48  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# Rappel : lrtest() est une fonction du package lmtest().
```

```
# Il faut qu'il soit chargé pour l'utiliser.

# La p-valeur est extrêmement faible autrement dit
# on peut sans risque rejeter l'hypothèse nulle d'égalité
# des vraisemblances des deux modèles. Le modèle complet
# est donc bien davantage explicatif que le modèle simple.
```

b. On cherche désormais à tester si l'augmentation de la relation croissante entre participation électorale et âge est valable autant chez les chômeurs et chez les non-chômeurs. Estimez un nouveau modèle dans lequel les variables `cho` et `agea` sont croisées et utilisez-le pour mener le test du ratio de vraisemblance correspondant à cette question.

Proposition de solution

```
# L'opérateur : permet de croiser des variables facilement.
logit3 <-  glm(
  vote ~ cho + agea + cho:agea + femme + agri + cadre + empl + ouv
  , data = ess_fr, family = binomial
)
summary(logit3)
##
## Call:
## glm(formula = vote ~ cho + agea + cho:agea + femme + agri + cadre +
##     empl + ouv, family = binomial, data = ess_fr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3426   0.4160   0.5734   0.7071   1.6676
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.452704   0.230773   1.962   0.0498 *
## cho         -1.972119   0.745981  -2.644   0.0082 **
## agea         0.024891   0.003642   6.834 8.23e-12 ***
## femme       -0.162423   0.126450  -1.284   0.1990
## agri        -0.681590   0.231960  -2.938   0.0033 **
## cadre        0.208576   0.203133   1.027   0.3045
## empl        -0.198832   0.182134  -1.092   0.2750
## ouv         -0.702845   0.173790  -4.044 5.25e-05 ***
## cho:agea     0.032350   0.017390   1.860   0.0628 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1831  on 1775  degrees of freedom
## Residual deviance: 1719  on 1767  degrees of freedom
##   (192 observations deleted due to missingness)
## AIC: 1737
##
## Number of Fisher Scoring iterations: 4

# Le coefficient de cho:agea autorise la pente de la droite
# selon l'âge à varier en fonction du chômage.

# On reformule alors l'hypothèse à tester :
# H_0 : beta(agea) = beta(cho:agea) contre H_1 : beta(agea) <> beta(cho:agea)
# Tester cette hypothèse revient à vérifier que le modèle
# logit3 qui autorise la variation de l'effet de l'âge
# selon la position sur le marché du travail est suffisament
# plus explicatif que le modèle logit2. On mène donc
# le test du ratio de vraisemblance de logit3 et logit2.
lrtest(logit2, logit3)
## Likelihood ratio test
##
## Model 1: vote ~ cho + agea + femme + agri + cadre + empl + ouv
## Model 2: vote ~ cho + agea + cho:agea + femme + agri + cadre + empl +
##     ouv
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -861.31
## 2   9 -859.49  1 3.6378    0.05648 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# La p-valeur est  supérieure à 0.05 donc on ne peut
# pas rejeter l'hypothèse d'égalité des coefficients
# à 5 %.
```

Atelier : Étude à partir de l'enquête Pisa 2012

Les objectifs de cet atelier sont de trois ordres :

- mobiliser en pratique les **méthodes et outils** introduits lors des sessions précédentes du certificat *Data Analyst* ;
- acquérir ou renforcer les compétences en **manipulations courantes de données avec R** : import de données depuis plusieurs formats, modification de la structure des données, fusions de table ;
- approfondir pour ceux et celles qui le souhaitent les questions relatives à la **production des graphiques** (notamment avec `ggplot2()`) et l'**automatisation de certaines**

opérations de *reporting* (avec Rmarkdown).

Concrètement, cet atelier est **organisé autour des étapes du travail sur de nouvelles données** et de la production d'un rapport statistique sur une question précise :

1. Présentation du contexte et définition de variables d'intérêt
2. Import et préparation des fichiers
3. Statistiques descriptives
4. Modélisation
5. Présentation des résultats

Bien que les données utilisées présentent nécessairement des particularités (*cf. infra*), **ces étapes sont communes à tout travail sur des données statistiques** : plus que ces particularités, c'est la **démarche générale**, les outils pratiques à mettre en oeuvre et leur interprétation qui constituent le coeur de cet atelier.

Les questions des deux premières sous-parties qui appellent une correction sont intégralement corrigées. Des indications sont fournies pour les questions des parties suivantes.

Présentation de l'enquête Pisa 2012

L'enquête Pisa (*Program for International Student Assessment*) est une enquête réalisée **tous les trois ans** par l'Organisation de coopération et de développement économique (OCDE) dans une soixantaine de pays auprès des **élèves de 15 ans** (quelle que soit leur classe au moment de l'enquête).

Elle vise à mesurer les **acquis des élèves de 15 ans dans trois disciplines** : mathématiques, compréhension de l'écrit (ou *littéracie*) et sciences. En plus des scores aux **tests standardisés** de mathématiques, compréhension de l'écrit et sciences, cette enquête comporte de très nombreuses informations sur l'origine sociale des élèves, leurs conditions d'enseignement ainsi que leur rapport aux enseignants et à l'école.

Il s'agit d'une **enquête par sondage** : à ce titre, chaque fichier dispose d'un poids à utiliser pour **généraliser les résultats** de l'échantillon d'élèves interrogés à l'ensemble de la population qu'ils représentent.

Organisation des fichiers *Les fichiers de l'enquête Pisa 2012 et leur documentation sont librement téléchargeables sur le site de l'OCDE. Seuls **deux des nombreux fichiers de données** qui constituent l'enquête seront utilisés et **ont été restreints à la France** uniquement :

— le **fichier élève** `int_stu12` comporte la plupart des variables d'intérêt de l'enquête ;
— le **fichier établissement** `int_scq12` comporte quelques données de contexte intéressantes.

Le fichier établissement est présent en plusieurs formats informatiques (`.txt`, `.csv` et `.sas7bdat`) pour illustrer les différentes méthodes d'importation des données. Les tableaux suivants recensent les principales variables d'intérêt de ces deux fichiers.

Fichier élèves (`int_stu12`)

Variable	Description
cnt	Pays
stidstd	Identifiant de l'élève
schoolid	Identifiant de l'établissement
w_fstuwt	Poids de sondage final de l'élève
st01q01	Classe en nombre d'années depuis l'entrée en primaire : la $10^{ème}$ classe correspond à la seconde en France.
st04q01	Sexe : (1) Femme (2) Homme
st05q01	A suivi une scolarité pré-primaire (1) Non (2) Oui, un an ou moins (3) Oui, plus d'un an
st07q01 st07q02 st07q03	A redoublé à un moment de sa scolarité : (1) Non (2-3) Oui, une ou plusieurs fois
st08q01	Est arrivé en retard au cours des deux semaines précédant l'enquête
st09q01	A séché les cours au cours des deux semaines précédant l'enquête
anxmat	Score synthétique d'anxiété en mathématiques
disclima	Score synthétique de climat de discipline dans la classe
escs	Indicateur synthétique de statut économique, social et culturel
immig	Immigration : (1) Né en France (2) Immigré de deuxième génération (3) Immigré de première génération
hisced	Niveau d'étude le plus élevé des parents (nomenclature CITE)
pv1math	Score synthétique à l'évaluation de mathématiques
pv1read	Score synthétique à l'évaluation de compréhension de l'écrit
pv1scie	Score synthétique à l'évaluation de sciences

Fichier établissements (`int_scq12`)

Variable	Description
cnt	Pays
schoolid	Identifiant de l'établissement
senwgt_scq	Poids de sondage (la somme vaut 1 000 dans chaque pays)
sc01q01	Statut public ou privé (1) public (2) privé

Variable	Description
sc03q01	Taille de la commune de l'établissement : (1) *Village* (2) *Small town* (3) *Town* (4) *City* (5) *Large city*
sc05q01	Taille de la classe en cours de français : (01) 15 ou moins (02) 16-20 (03) 21-25 ... (08) 46-50 (09) Plus de 50 élèves

Documentaton Plusieurs éléments de documentation sont fournis :

- le questionnaire rempli par les élèves (`PISA12_ScQ_ENG.pdf`) et la description du fichier élèves (`M_stu_codebook.pdf`) *(en anglais)* ;
- le questionnaire rempli par les établissements (`PISA12_ScQ_ENG.pdf`) et la description du fichier établissements (`M_sch_codebook.pdf`) *(en anglais)* ;
- les principaux résultats de l'enquête en France (*PISA-2012-results-france.pdf*) **(en français)** et le rapport technique international (`PISA-2012-technical-report-final.pdf`) *(en anglais).*

Question 1 Découverte de l'enquête et de sa documentation

a. Vérifiez la présence et identifiez l'ensemble des fichiers mentionnés ci-dessus. Ouvrez en particulier le questionnaire élève et observez les premières questions posées (« Section A : About you »). Retrouvez les variables correspondantes dans le tableau ci-dessus ainsi que dans le fichier `.pdf` de description du fichier élève.

b. Ouvrez la note rédigée sur la France (`PISA-2012-results-france.pdf`) et prenez connaissance de ses principaux résultats. Quand il est fait référence à un résultat statistique précis (moyenne, pourcentage, etc.), recherchez dans les tableaux ci-dessus les variables susceptibles d'avoir été utilisées pour y aboutir.

Import et préparation des données

Question 2 Import des données

Plusieurs formats de fichiers sont fournis afin de **présenter différentes méthodes d'importation courantes**. Les fichiers de référence à importer et à utiliser dans le reste de l'atelier sont les fichiers SAS au format `.sas7bdat` (importation présentée à la sous-question d.).

Le fichier `INT_SCQ12_DEC03.txt` correspond au **format original des données** (sur le site de l'OCDE). Il s'agit d'un fichier de **données à largeur fixe** (*fixed-width format*) : contrairement

aux fichiers dont les colonnes sont séparées par un délimiteur (virgules, tabulations, etc.), ici à **chaque colonne correspond un nombre de caractères fixe** qui est utilisé pour **reconstituer les variables** du fichier dans le logiciel statistique.

a. Recherchez le programme Notepad++ sur l'ordinateur et utilisez-le (ou à défaut le bloc-note Windows) pour visualiser le contenu de `INT_SCQ12_DEC03.txt` (celui-ci étant un peu gros l'ouverture peut prendre un peu de temps). Vérifiez visuellement que les colonnes sont bien de largeur fixe.

b. Définissez comme répertoire de travail le dossier dans lequel se situent les données de l'enquête. La fonction `read.fwf()` permet de lire un fichier de données à largeur fixe pour autant qu'on lui fournisse la largeur correspondant à chaque variable. Comment comprenez-vous alors le code suivant (établi à partir des fichiers `.sas` mis en ligne par l'OCDE) ? Soumettez le code suivant et vérifiez que l'importation se déroule correctement.

```
sch_fwf <- read.fwf("INT_SCQ12_DEC03.txt", width = c(
  3, 7, 7, 1, 6, 7, 1, 9, 9, 9, 9, 1, 1, 2, 9, 9, 4, 4, 4, 4, 4, 4, 4
  , 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
  , 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
  , 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
  , 1, 1, 1, 1, 1, 1, 1, 1, 9, 1, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9
  , 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
  , 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
  , 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
  , 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
  , 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 9
  , 9, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
  , 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 9, 9, 1, 1, 2, 9, 1, 9, 9, 9, 9, 1
  , 1, 9, 9, 9, 9, 9, 9, 8, 8, 9, 1, 1, 8, 9, 9, 9, 9, 9, 9, 9, 9, 9
  , 9, 9, 9, 7
), stringsAsFactors = FALSE)
```

Proposition de solution

```
# INDICATION : commencez par afficher l'aide de la fonction read.fwf()
# avec ? read.fwf

# Définition du répertoire de travail
setwd("//emplacement/reseau")
# NOTE : pensez bien à remplacer les \ par des /

# L'argument width = de la fonction read.fwf() permet d'indiquer
# la longeur des variables dans l'ordre du fichier. Par exemple ici,
# la première variable correspond aux trois premières colonnes, la
# deuxième aux sept colonnes suivantes, etc.
```

```
# L'importation réussit après un certain temps, mais le fichier ne
# comporte aucun nom de variables. Dans le cas d'un fichier à largeur
# fixe en général le nom des variables est contenu dans un fichier
# externe à utiliser pour renommer les variables par la suite.
```

c. Bien souvent les fichiers à importer ne sont pas des fichiers à largeur fixe mais des fichiers séparés par un délimiteur (virgule, tabulation, etc.). Rercherchez de l'aide sur la fonction `read.csv()` et utilisez-la pour importer le fichier `INT_SCQ12_DEC03.csv`. Le *package* de manipulation de données `data.table` comporte une fonction d'import optimisé pour les données plates séparées par des délimiteurs `fread()` dont la syntaxe est proche de celle de `read.csv()` : installez ce *package*, recherchez de l'aide sur `fread()` et comparez (soit "à vue d'oeil", soit en mesurant le temps d'exécution) la vitesse de `fread()` par rapport à `read.csv()`.

Proposition de solution

```
# Import du fichier .csv avec read.csv()
sch_csv_read <- read.csv("INT_SCQ12_DEC03.csv", stringsAsFactors = FALSE)

# Installation du package data.table (A NE FAIRE QU'UNE FOIS)
# install.packages("data.table")

# Chargement du package data.table (A FAIRE A CHAQUE SESSION)
library(data.table)

# Import du fichier .csv avec fread()
sch_csv_fread <- fread("INT_SCQ12_DEC03.csv", data.table = FALSE)
# Note : l'option data.table = FALSE permet d'obtenir en sortie
# un objet de type data.frame, comme avec read.csv()

# Comparaison des temps d'exécution
system.time(read.csv("INT_SCQ12_DEC03.csv", stringsAsFactors = FALSE))
system.time(fread("INT_SCQ12_DEC03.csv", data.table = FALSE))
```

d. Sur le site de l'OCDE pour l'heure tout est fait pour privilégier un import avec les logiciels SAS ou SPSS. Les fichiers `int_stu12.sas7bdat` et `int_scq12.sas7bdat` correspondent aux fichiers importés par SAS au format natif SAS `.sas7bdat` et **restreints à la France**. Utilisez la fonction `read_sas()` pour lire ces fichiers (c'est sur ceux-ci que l'on travaillera désormais).

Proposition de solution

```
# Installation du package haven (A NE FAIRE QU'UNE FOIS)
# install.packages("haven")
```

```r
# Chargement du package haven (A FAIRE A CHAQUE SESSION)
library(haven)

# Lecture des fichiers .sas7bdat
stu <- read_sas("int_stu12.sas7bdat")
sch <- read_sas("int_scq12.sas7bdat")
```

Question 3 Exploration et mise en forme des données

a. Utilisez la fonction `names()` pour afficher les noms de variables des deux tables. Afin de faciliter les exploitations futures, passez tous ces noms en minuscules.

Proposition de solution

```r
# Affichage des noms
names(stu)[1:10]
##  [1] "CNT"      "SUBNATIO" "STRATUM"  "OECD"     "NC"       "SCHOOLID"
##  [7] "StIDStd"  "ST01Q01"  "ST02Q01"  "ST03Q01"
names(sch)[1:10]
##  [1] "CNT"      "SUBNATIO" "STRATUM"  "OECD"     "NC"       "SCHOOLID"
##  [7] "SC01Q01"  "SC02Q01"  "SC02Q02"  "SC02Q03"
# Note : je me limite aux 10 premiers pour ne pas surcharger la correction.

# Passage des noms en minuscules avec la fonction tolower()
names(stu) <- tolower(names(stu))
names(sch) <- tolower(names(sch))
```

b. Affichez les principales caractéristiques des deux objets importés (structure, dimension, nom des variables, etc.). Vérifiez qu'ils ne comportent que des informations relatives à la France.

Proposition de solution

```r
# Fonctions classiques de description d'objets
# str(stu)
# str(sch)
# Note : commenté pour ne pas surcharger la correction
dim(stu)
## [1] 4613  634
dim(sch)
```

```
## [1] 226 291
# Note : tbl_df est le format par défaut dans lequel
# importe read_sas(). Il s'agit d'une amélioration
# de data.frame portée par l'équipe de RStudio (dplyr, etc.)

# La variable cnt code le code du pays (FRA) pour la France.
# Plusieurs solutions :
# 1) fonction table()
table(stu$cnt)
##
##  FRA
## 4613
table(sch$cnt)
##
## FRA
## 226
# 1) fonction unique()
unique(stu$cnt)
## [1] "FRA"
unique(sch$cnt)
## [1] "FRA"
```

c. Afin d'alléger les traitements à venir (pas impératif ici mais toujours utile en pratique), on décide de restreindre les variables d'intérêt à celles identifiées dans les tableaux ci-dessus. Supprimez toutes les autres variables des fichiers.

Proposition de solution

```
# Restriction aux variables mentionnées des tableaux
# ci-dessus avec `[`
stu <- stu[, c("cnt", "stidstd", "schoolid", "w_fstuwt", "st01q01", "st04q01", "st0
sch <- sch[, c("cnt", "schoolid", "senwgt_scq", "sc01q01", "sc03q01", "sc05q01")]

# Affichage des caractéristiques des tables réduites
str(stu)
## Classes 'tbl_df', 'tbl' and 'data.frame':    4613 obs. of  20 variables:
##  $ cnt     : chr  "FRA" "FRA" "FRA" "FRA" ...
##  $ stidstd : chr  "00001" "00002" "00003" "00004" ...
##  $ schoolid: chr  "0000001" "0000001" "0000001" "0000001" ...
##  $ w_fstuwt: num  136 149 136 152 136 ...
##  $ st01q01 : num  10 11 10 10 10 10 11 10 10 10 ...
##  $ st04q01 : num  2 2 2 1 2 1 2 1 2 2 ...
##  $ st05q01 : num  3 3 3 3 3 3 3 3 NA 3 ...
##  $ st07q01 : num  1 1 1 1 1 1 1 1 NA 1 ...
##  $ st07q02 : num  1 1 1 1 1 1 1 1 NA 1 ...
```

```
##  $ st07q03 : num  1 1 1 1 1 1 1 1 NA 1 ...
##  $ st08q01 : num  1 1 1 1 1 1 1 1 NA 1 ...
##  $ st09q01 : num  1 1 1 1 1 1 1 1 NA 1 ...
##  $ anxmat  : num  NA NA 0.79 NA -0.2 NA -1.3 -0.2 NA 0.56 ...
##  $ disclima: num  NA NA 0.49 NA -0.53 NA -0.08 -1.2 NA -0.53 ...
##  $ escs    : num  0.18 0.18 -0.53 0.8 0.57 1.07 0.15 1.12 NA 1.18 ...
##  $ immig   : num  1 1 1 1 1 1 1 1 NA NA ...
##  $ hisced  : num  5 4 3 6 4 6 5 6 NA 6 ...
##  $ pv1math : num  579 479 600 494 598 ...
##  $ pv1read : num  582 493 526 505 604 ...
##  $ pv1scie : num  568 478 584 528 532 ...
str(sch)
## Classes 'tbl_df', 'tbl' and 'data.frame':    226 obs. of  6 variables:
##  $ cnt       : chr  "FRA" "FRA" "FRA" "FRA" ...
##  $ schoolid  : chr  "0000001" "0000002" "0000003" "0000004" ...
##  $ senwgt_scq: num  1.37 2.41 12.6 3.48 1.17 ...
##  $ sc01q01   : num  1 2 1 NA 1 2 2 2 1 1 ...
##  $ sc03q01   : num  3 3 1 NA 3 3 2 3 1 1 ...
##  $ sc05q01   : chr  "04" "03" "04" "97" ...
```

Question 4 Fusion de tables

On souhaite utiliser les informations au niveau de l'établissement dans des exploitations au niveau des élèves. Pour ce faire, il convient de fusionner les tables `stu` et `sch` sur la base de la variable `schoolid` présente dans les deux tables :

— dans la table élèves, il s'agit de l'identifiant de l'établissement auquel appartient l'élève (les élèves provenant du même établissement ont le même `schoolid`) ;
— dans la table établissement, il s'agit de l'identifiant unique de l'établissement.

a. Utilisez les fonctions `unique()`, `intersect()` et `setdiff()` pour vérifier que l'identifiant `schoolid` prend bien les mêmes valeurs dans les deux tables.

Proposition de solution

```
# INDICATION : commencez par chercher de l'aide sur ces fonctions
# puis testeé-les avec des petits vecteurs simples

# Exemple
a <- rep(1:10, 2)
a
##  [1]  1  2  3  4  5  6  7  8  9 10  1  2  3  4  5  6  7  8  9 10
```

```
b <- rep(6:15, 2)
b
##  [1]  6  7  8  9 10 11 12 13 14 15  6  7  8  9 10 11 12 13 14 15
unique(a)
##  [1]  1  2  3  4  5  6  7  8  9 10
unique(b)
##  [1]  6  7  8  9 10 11 12 13 14 15
unique(c(a, b))
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
intersect(a, b)
## [1]  6  7  8  9 10
setdiff(a, b)
## [1] 1 2 3 4 5
setdiff(b, a)
## [1] 11 12 13 14 15
# Note : setdiff() est la différence ensembliste :
# - setdiff(a, b) renvoie les valeurs dans a MAIS pas dans b
# - setdiff(b, a) renvoie les valeurs dans b MAIS pas dans a

# On peut dès lors utiliser ces opérations pour vérifier que
# schoolid prend bien les mêmes valeurs de part et d'autre
length(intersect(stu$schoolid, sch$schoolid))
## [1] 226
length(unique(stu$schoolid))
## [1] 226
length(unique(sch$schoolid))
## [1] 226

# Cela fonctionne aussi plus finement avec setdiff()
setdiff(stu$schoolid, sch$schoolid)
## character(0)
# Aucune valeur de stu$schoolid n'est pas dans sch$schoolid
setdiff(sch$schoolid, stu$schoolid)
## character(0)
# Aucune valeur de sch$schoolid n'est pas dans stu$schoolid
```

b. Vérifiez que la variable `schoolid` est un identifiant pour la table `sch`, à savoir : (1) qu'elle est renseignée pour chaque ligne (2) qu'elle prend une valeur distincte pour chaque ligne.

Proposition de solution

```
# Pour vérifier que schoolid identifie sch il suffit
# de comparer le nombre de valeurs distinctes au
# nombre de lignes
```

```
length(unique(sch$schoolid))
## [1] 226
nrow(sch)
## [1] 226
# Tout va bien !
```

c. Utilisez la fonction `merge()` pour fusionner `stu` et `sch` par `schoolid` dans l'objet `stusch`. Vérifiez que ses propriétés sont cohérentes avec le résultat des questions précédentes : même nombre de lignes que `stu`, nombre de colonnes égal à celui de `stu` et de `sch` moins 1.

Proposition de solution

```
# INDICATION : commencez par regarder l'aide de la fonction merge()

# Fusion des deux tables
stusch <- merge(stu, sch, by = c("cnt", "schoolid"))
# Note : le nom de la variable de fusion doit être entre guillemets

# Vérifications
nrow(stusch) == nrow(stu)
## [1] TRUE
ncol(stusch) == ncol(stu) + ncol(sch) - 1
## [1] FALSE
```

Statistiques uni- et bivariée

Inspirez-vous de la note rédigée sur la France pour choisir un sujet d'étude. Par exemple :

- relation entre caractéristiques socio-démographiques et score synthétique en mathématiques, compréhension de l'écrit ou sciences ou probabilité d'être redoublant ;
- relation entre scolarité pré-primaire et score synthétique en mathématiques, compréhension de l'écrit ou sciences ou probabilité d'être redoublant ;
- relation entre rapport à l'institution scolaire (angoisse, fait de sécher les cours, etc.) et score synthétique en mathématiques ou probabilité d'être redoublant ;
- relation entre le contexte d'enseignement de l'élève (statut de l'établissement, climat de discipline, etc.) et score synthétique en mathématiques, compréhension de l'écrit ou sciences ou probabilité d'être redoublant.

Vous pouvez en particulier chercher à recalculer certaines statistiques de la note rédigée sur la France.

Proposition de solution

```r
# Recodage de variables

# Transformation de la variable de sexe en facteur avec
# un libellé plus clair
stusch$sexe <- factor(stusch$st04q01, labels = c("Femme",  "Homme"))

# L'élève a-t-il redoublé à un moment ou à un autre ?
# Les variables st07q01 st07q02 st07q03 codent le fait
# d'avoir redoublé une ou plusieurs fois en primaire,
# au collège ou au lycée : Un élève a redoublé dès lors
# qu'une de ces variables prend la valeur 2 ou 3.
stusch$redoublant <- (
  stusch$st07q01 %in% c(2, 3) |
    stusch$st07q02 %in% c(2, 3) |
    stusch$st07q03 %in% c(2, 3)
) * 1
table(stusch$redoublant)
##
##    0    1
## 3356 1257
# Environ 25 % des élèves interrogés ont redoublé, ce qui paraît raisonnable.

# Indicateur synthétique de statut économique, social et culturel
# La variable escs est de nature quantitative, ce qui n'est pas très naturel
summary(stusch$escs)
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -3.46000 -0.60000  0.04000 -0.02122  0.60000  2.20000      124
# On la recode en quintiles pour facilier les interprétations
stusch$escsq <-  cut(
  stusch$escs
  , quantile(stusch$escs, probs = c(0, 0.20, 0.40, 0.60, 0.80, 1), na.rm = TRUE)
)
table(stusch$escsq)
##
## (-3.46,-0.74] (-0.74,-0.21]  (-0.21,0.25]   (0.25,0.72]    (0.72,2.2]
##           902           896           896           905           889

# Recodage en deux modalités de la variable immig
stusch$immig2 <- (stusch$immig %in% c(2, 3)) * 1
table(stusch$immig2)
##
##    0    1
## 3960  653

# Recodage en deux modalités de la variable de scolarité pré-primaire
```

```
stusch$paspreprim <- ( stusch$st05q01 %in% c(1, 2) ) * 1
table(stusch$paspreprim)
##
##    0    1
## 4252  361


# Création d'un vecteur de poids normalisé
stusch$poids <- stusch$w_fstuwt / mean(stusch$w_fstuwt)
```

Question 5 Statistique descriptive

Utilisez les fonctions vues lors de la première session du certificat pour mener l'analyse uni- et bi-variée des variables pertinentes pour le sujet d'étude que vous avez choisi. En particulier :

— construisez au moins un tri croisé entre deux variables qualitatives et interprétez les sur- ou sous-représentations dans la distribution de la première variable ventilée selon les modalités de la seconde ;

Proposition de solution

```
# Sexe des élèves et redoublement
# Utilisation du package descr pour simplifier l'analyse de tri croisé
library(descr)
crosstab(
  stusch$sexe, stusch$redoublant, weight = stusch$poids
  , prop.r = TRUE, chisq = TRUE, prop.c = TRUE
)
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |          Column Percent |
## |-------------------------|
##
## ====================================
##                stusch$redoublant
## stusch$sexe        0       1   Total
## ------------------------------------
## Femme           1776     595    2371
##                 74.9%   25.1%   51.4%
##                 53.4%   46.3%
## ------------------------------------
## Homme           1550     691    2241
```

```
##                69.2%   30.8%   48.6%
##                46.6%   53.7%
## -----------------------------------
## Total          3326    1286    4612
##                72.1%   27.9%
## ===================================
##
## Statistics for All Table Factors
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 = 18.87363      d.f. = 1      p = 1.4e-05
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 = 18.58929      d.f. = 1      p = 1.62e-05
##         Minimum expected frequency: 624.8755
```

Dans l'ensemble de la population, 27,9 % des élèves de 15 ans ont redoublé à un moment ou à un autre de leur scolarité. Ils sont 30,8 % parmi les hommes et 25,1 % parmi les femmes : autrement dit, les élèves ayant deroublé à un moment ou à un autre de leur scolarité sont surreprésentés parmi les hommes.
Le test d'indépendance du χ^2 permet de confirmer cette analyse : sa p-valeur est inférieure à 0,01 aussi il est possible de rejeter l'hypothèse nulle d'indépendance entre les variables de sexe et de reoublement au seuil de 1 %.

— calculez le coefficient de corrélation entre deux variables quantitatives et menez le test de nullité de ce coefficient (avec la fonction `cor.test()`) ;

Proposition de solution

```
# Notes en mathématiques et en sciences
cor(stusch$pv1math, stusch$pv1scie)
## [1] 0.9019808
cor.test(stusch$pv1math, stusch$pv1scie)
##
##  Pearson's product-moment correlation
##
## data:  stusch$pv1math and stusch$pv1scie
## t = 141.85, df = 4611, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8964568 0.9072244
## sample estimates:
##       cor
## 0.9019808
```

Le coefficient de corrélation entre note en mathématiques et note en sciences est positif et extrêmement élevé (0,9020). La p-valeur du test de nullité de ce coefficient est inférieure à 0,01 aussi il est possible de rejeter au seuil de 1 % l'hypothèse que ces deux scores ne soient pas corrélés.

— calculez la moyenne d'une variable quantitative ventilée selon les modaliités d'une variable qualitative et effectuez l'analyse de la variance à un seul facteur correspondante pour déterminer si la relation entre ces variables est significative (avec la fonction `anova()`).

Proposition de solution

```
# Score en mathématiques selon le statut économique, social et culturel
tapply(stusch$pv1math, stusch$escsq, mean)
## (-3.46,-0.74] (-0.74,-0.21]  (-0.21,0.25]   (0.25,0.72]    (0.72,2.2]
##      441.6501      468.6728      499.9407      532.5550      567.2039
anova(lm(pv1math ~ escsq, data = stusch))
## Analysis of Variance Table
##
## Response: pv1math
##             Df   Sum Sq Mean Sq F value    Pr(>F)
## escsq        4  8908656 2227164  304.51 < 2.2e-16 ***
## Residuals 4483 32788549    7314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La moyenne du score en mathématiques est d'autant plus élevée que le statut économique, social et culturel est favorisé. Le test de Fisher associé à l'analyse de variance a une p-valeur inférieure à 0,01 : cela confirme que l'association entre statut économique, social et culturel et résultats en mathématiques est statistiquement significative.

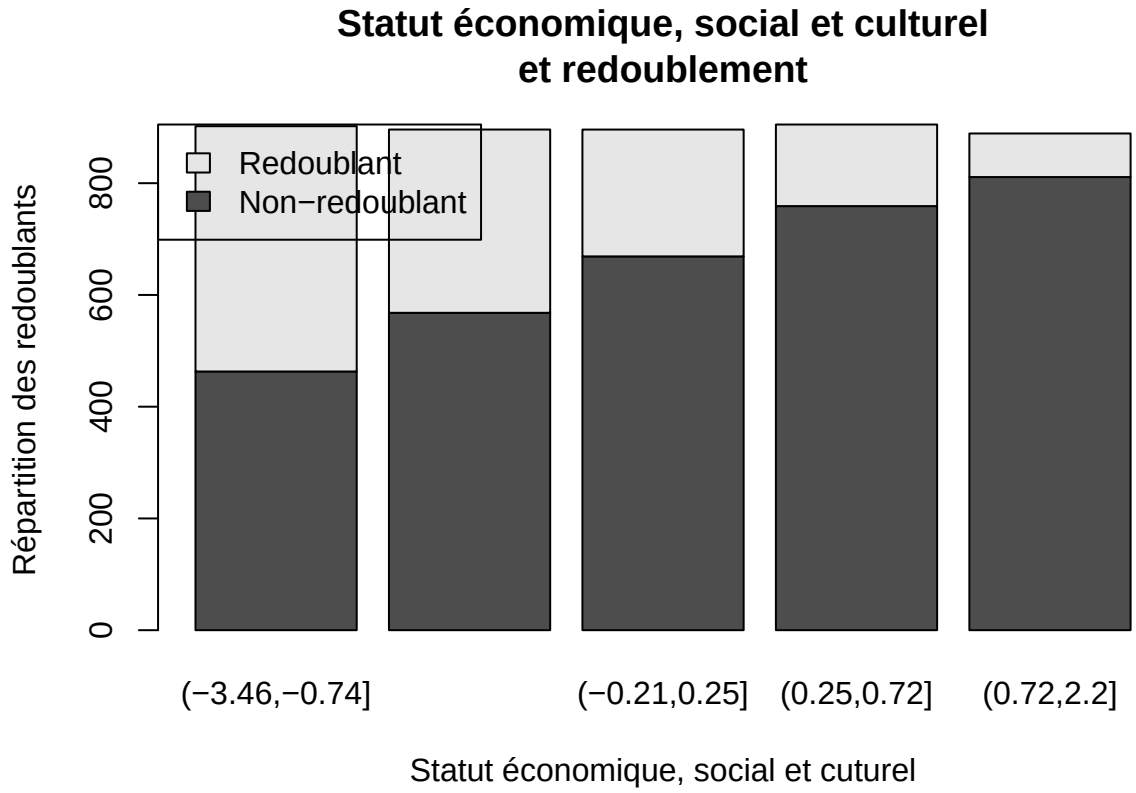
Question 6 Réalisation de graphiques avec base R

Proposez des représentations graphiques pertinentes pour synthétiser les relations mises en évidence à la question précédente. Vous pouvez consulter à ce sujet la première partie du support complémentaire pour l'atelier. En particulier :

— construisez au moins un diagramme en bâton ou circulaire à l'aide des fonctions `barplot()` ou `pie()` et utilisez les options de mise en forme pour améliorer sa présentation (ajouter un titre avec `main()`, modifiez les titres des axes avec `xlab()` et `ylab()`, etc.) ;

Proposition de solution

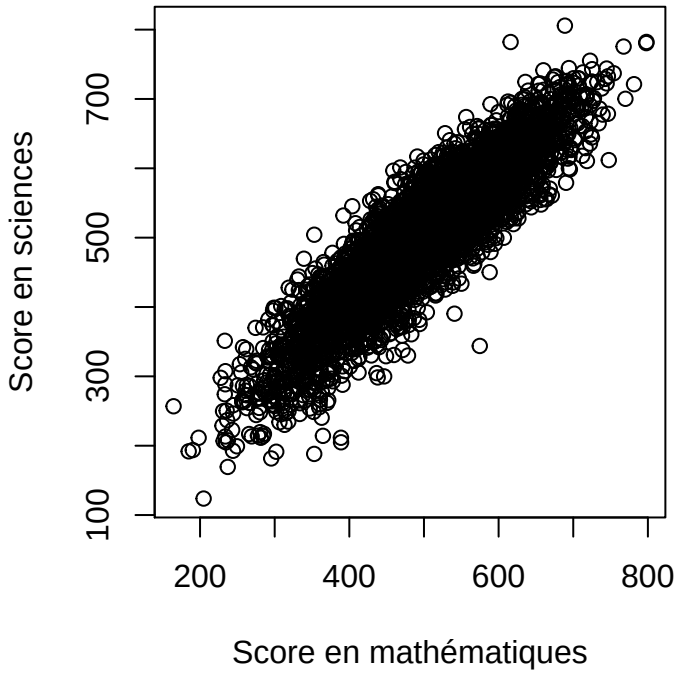
```r
# Statut économique, social et culturel et redoublement
barplot(
  table(stusch$redoublant, stusch$escsq)
  , legend.text = c("Non-redoublant", "Redoublant"), args.legend = list(x = "toplef
  , main = "Statut économique, social et culturel\net redoublement"
  , xlab = "Statut économique, social et cuturel"
  , ylab = "Répartition des redoublants"
)
```



— construisez au moins un nuage de points à l'aide de la fonction `plot()` ;

Proposition de solution

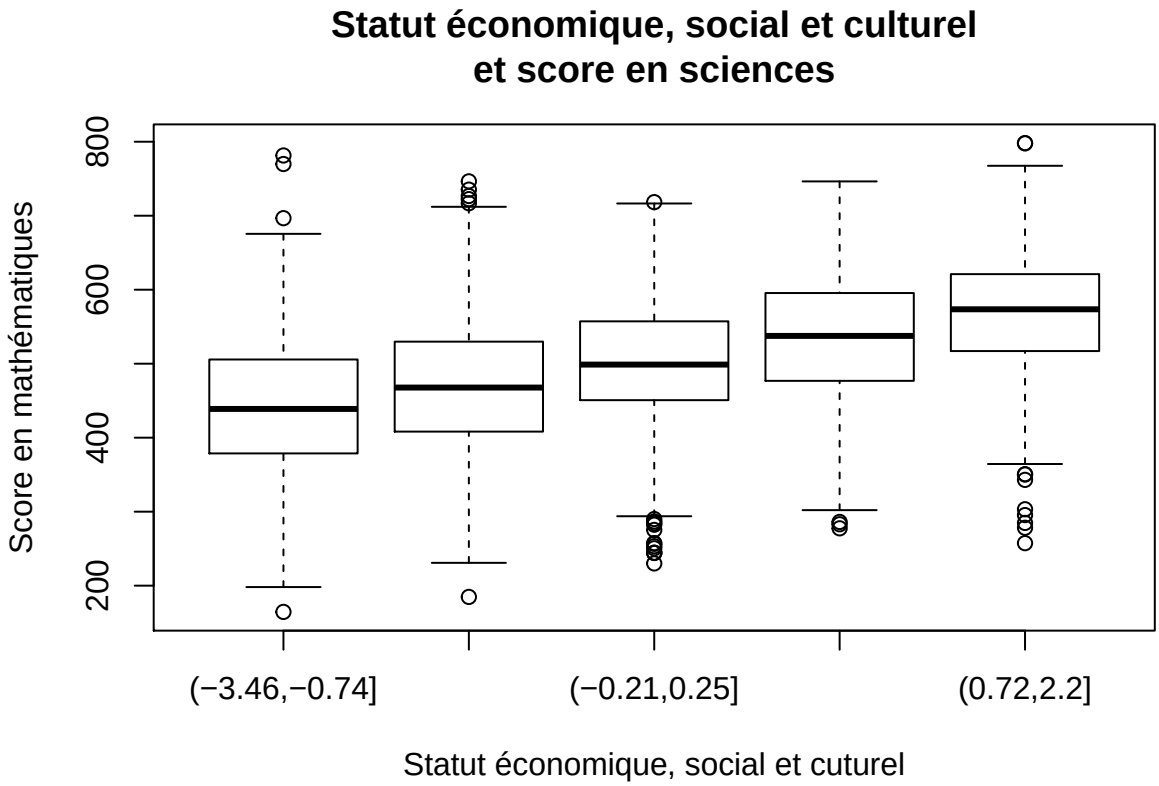
```r
# Score en mathématiques et en sciences
par(pty="s")
plot(
  stusch$pv1math, stusch$pv1scie
  , xlab = "Score en mathématiques"
  , ylab = "Score en sciences"
)
```



— construisez au moins une série de boîtes à moustache à l'aide de la fonction `boxplot()`.

Proposition de solution

```
# Statut économique, social et culturel et score en mathématiques
boxplot(
  stusch$pv1math ~ stusch$escsq
  , main = "Statut économique, social et culturel\net score en sciences"
  , xlab = "Statut économique, social et cuturel"
  , ylab = "Score en mathématiques"
)
```



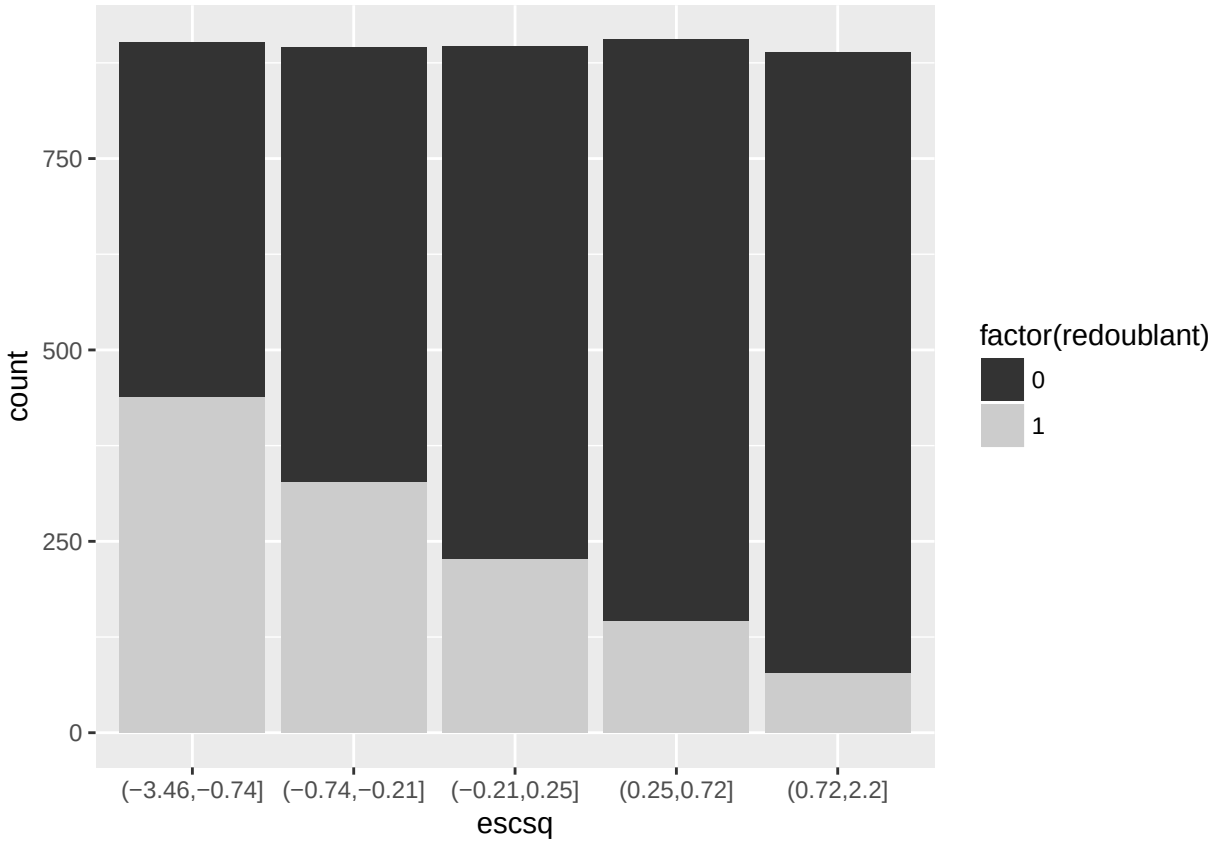
— utilisez les fonctionnalités de RStudio (menus déroulants de la fenêtre de graphiques) pour sauvegarder ces graphiques dans la qualité et le format que vous préférez.

Question 7 (Optionnel) Réalisation de graphiques avec ggplot2

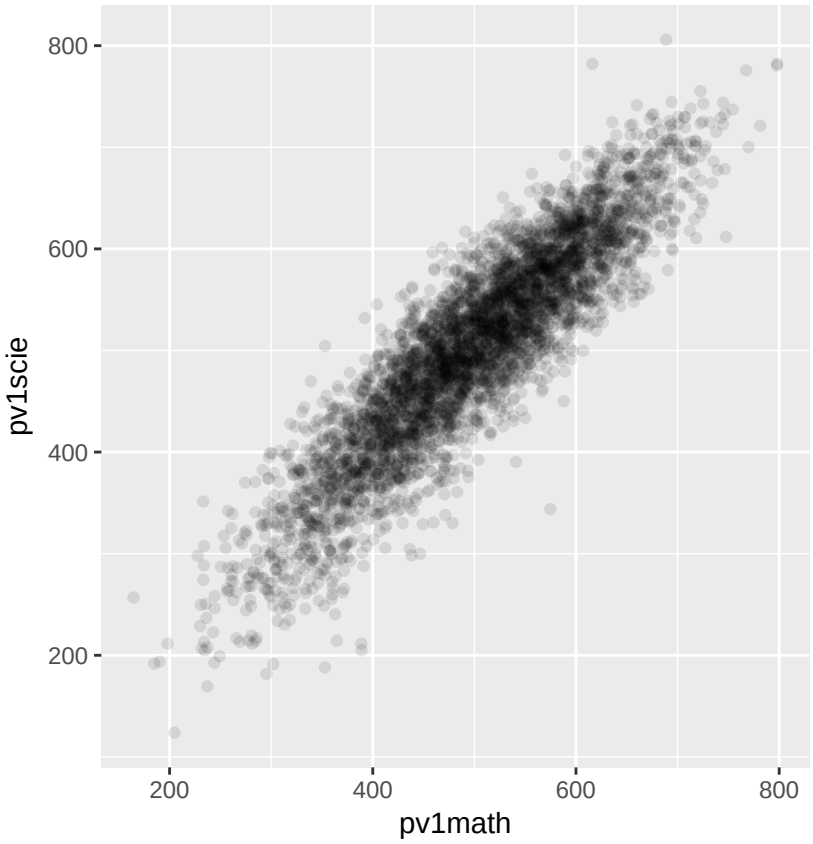
Consultez attentivement la deuxième partie du support complémentaire pour l'atelier ainsi éventuellement que le chapitre 2 de l'ouvrage de référence d'Hadley Wickham sur le sujet puis cherchez à reproduire les graphiques de la question précédente avec `ggplot2`.

Proposition de solution

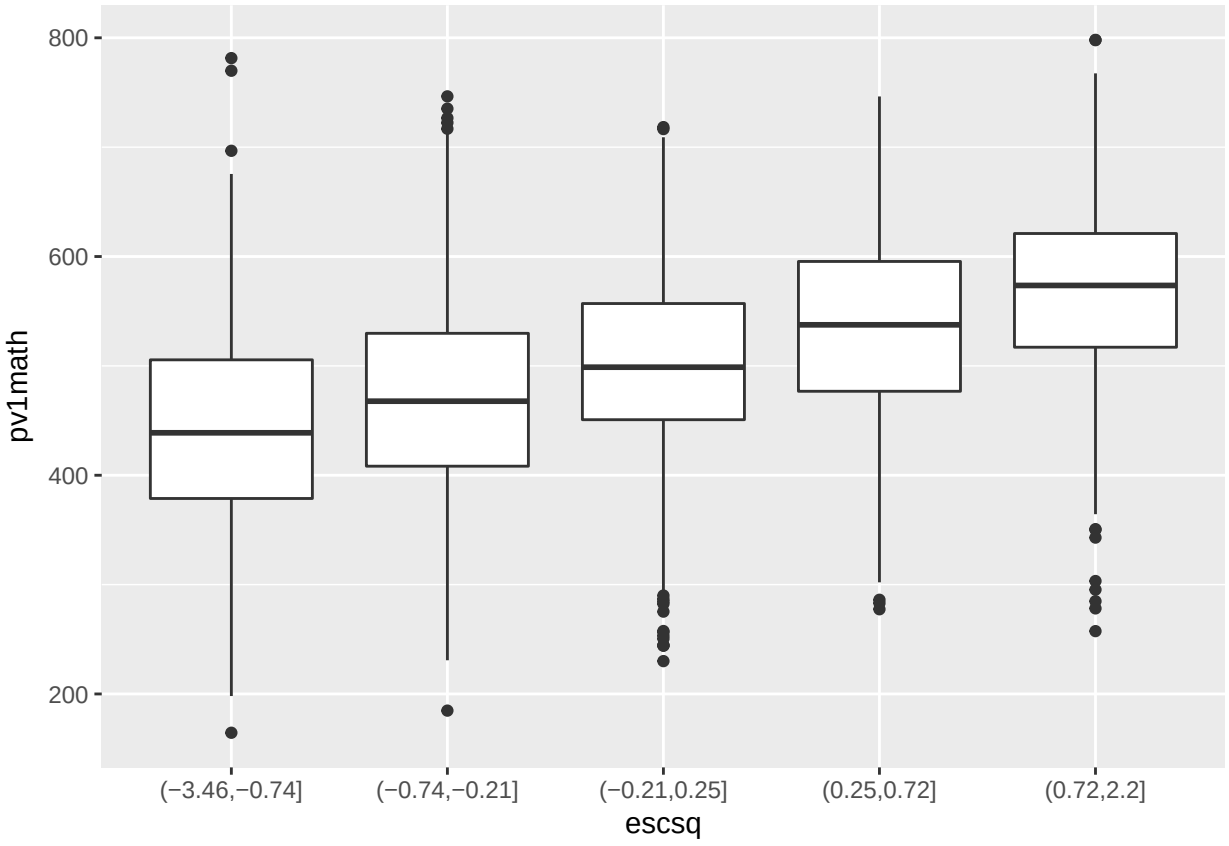
```
# Installation du package ggplot2 (A NE FAIRE QU'UNE FOIS)
# install.packages("ggplot2")

# Chargement du package ggplot2
library(ggplot2)

# Statut économique, social et culturel et redoublement
ggplot(data = stusch[!is.na(stusch$escsq), ], aes(escsq, fill = factor(redoublant))) +
  geom_bar() + scale_fill_grey()
```




```
# Score en mathématiques et en sciences
ggplot(data = stusch, aes(pv1math, pv1scie)) + geom_point(alpha = 0.1) +
  coord_fixed()
```




```
# Statut économique, social et culturel et score en mathématiques
ggplot(data = stusch[!is.na(stusch$escsq), ], aes(escsq, pv1math)) +
  geom_boxplot()
```



Modélisations

Sur la mise en oeuvre pratique des méthodes de régressions avec **R**, ne jamais hésiter à consulter les tutoriels très bien faits de UCLA.

Question 8 Régression linéaire

Inspirez-vous de certaines analyses de la note sur la France (par exemple p. 12, sur les performances des élèves issus de l'immigration après contrôle des caractéristiques socio-démographiques) pour proposer des modèles de régression linéaire cohérents avec votre étude. Pensez bien en particulier :

- à vérifier le codage des non-réponses dans les données pour que celles-ci ne viennent pas perturber l'analyse (passez-les en `NA` si ce n'est pas déjà le cas) ;
- à évaluer la qualité de vos modèles à l'aide des indicateurs classiques : R^2, $R^2 ajusté$, test de significativité globale de Fisher ;
- à étudier les résidus et leur distribution ;
- à interpréter la valeur des coefficients en termes de relation (toutes les autres variables du modèle égales par ailleurs) entre la variable explicative considérée et la variable expliquée ;

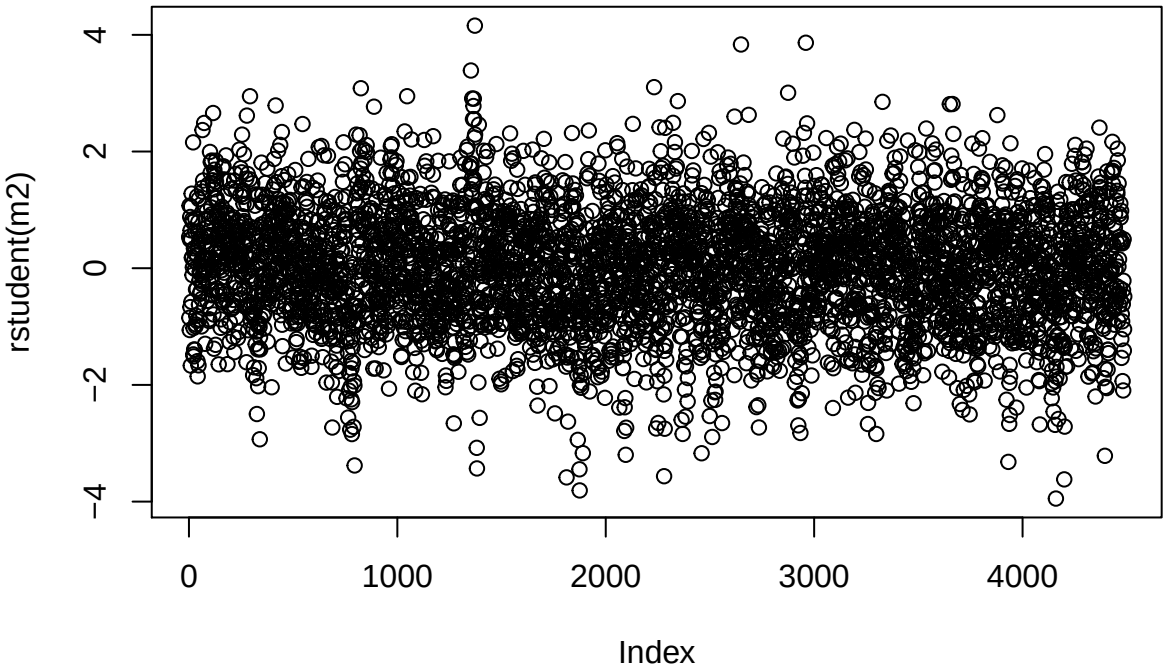
— à interpréter la significativité des coefficients à l'aide soit de la statistique de test, soit de la p-valeur, soit encore de l'intervalle de confiance au niveau souhaité (à construire avec `confint.default()`) ;
— à utiliser le modèle pour prédire la valeur d'individus avec des profils particuliers avec la fonction `predict()` et à les représenter.

Proposition de solution

Comme suggéré dans l'énoncé, on choisit d'analyser les relations entre performances en mathématiques et origine migratoire des élèves, en ajoutant quelques variables contrôle.

```
# Analyse bivariée : test d'égalité des moyennes
tapply(stusch$pv1math, stusch$immig2, mean)
##        0        1
## 508.3966 445.3395
t.test(stusch$pv1math ~ stusch$immig2)
##
##  Welch Two Sample t-test
##
## data:  stusch$pv1math by stusch$immig2
## t = 15.596, df = 872.06, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  55.12156 70.99259
## sample estimates:
## mean in group 0 mean in group 1
##        508.3966        445.3395

# Analyse multivariée : contrôle par le statut socio-économique brut
m1 <- lm(pv1math ~ immig2 + escs, data = stusch)
summary(m1)
##
## Call:
## lm(formula = pv1math ~ immig2 + escs, data = stusch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -297.96  -56.56    1.71   58.44  357.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  507.819      1.359 373.793   <2e-16 ***
## immig2       -34.065      3.718  -9.162   <2e-16 ***
## escs          53.465      1.623  32.952   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.09 on 4486 degrees of freedom
##   (124 observations deleted due to missingness)
## Multiple R-squared:  0.2396, Adjusted R-squared:  0.2393
## F-statistic: 706.8 on 2 and 4486 DF,  p-value: < 2.2e-16
# On retrouve presque le chiffre avancé dans le rapport p. 12 :
# "Même après contrôle du milieu socio-économique, en France, les élèves
# issus de l'immigration accusent des scores en mathématiques inférieurs
# de 37 points à ceux des élèves autochtones"

# Analyse multivariée : recodage du statut économique, social et culturel
# et ajout d'autres variables de contrôle
stusch$escsq <- relevel(stusch$escsq, ref = "(-0.21,0.25]")
stusch$sexe <- relevel(stusch$sexe, ref = "Homme")
m2 <- lm(pv1math ~ immig2 + escsq + sexe + redoublant + paspreprim, data = stusch)
summary(m2)
##
## Call:
## lm(formula = pv1math ~ immig2 + escsq + sexe + redoublant + paspreprim,
##     data = stusch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -283.86  -48.19    2.04   48.09  298.66
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         540.450      2.786 194.005  < 2e-16 ***
## immig2              -29.839      3.192  -9.349  < 2e-16 ***
## escsq(-3.46,-0.74]  -27.838      3.491  -7.974 1.94e-15 ***
## escsq(-0.74,-0.21]  -17.286      3.423  -5.050 4.59e-07 ***
## escsq(0.25,0.72]     21.978      3.406   6.453 1.21e-10 ***
## escsq(0.72,2.2]      48.318      3.442  14.040  < 2e-16 ***
## sexeFemme           -18.428      2.162  -8.523  < 2e-16 ***
## redoublant          -99.641      2.578 -38.655  < 2e-16 ***
## paspreprim          -34.342      4.098  -8.380  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.07 on 4479 degrees of freedom
##   (125 observations deleted due to missingness)
## Multiple R-squared:  0.442,  Adjusted R-squared:  0.441
## F-statistic: 443.5 on 8 and 4479 DF,  p-value: < 2.2e-16
plot(rstudent(m2))
```



Analyse des résultats du modèle `m2` :

- R^2 et R^2 ajusté de l'ordre de 44 % : faible quand le modèle est utilisé à des fins prédictives, mais suffisant quand le modèle est utilisé à des fins explicatives ;
- p-valeur du test de significativité globale de Fisher de 0,01 : on peut sans problème rejeter l'hypothèse que le modèle n'apprend rien sur les variations de la variable expliquée.
- la valeur du coefficient associé à la variable `redoublant` vaut -99,64 : en moyenne dans l'échantillon et à sexe, origine migratoire, statut économique, social et culturel et scolarité pré-primaire égaux par ailleurs, le fait d'avoir redooublé est associé à un score en mathématiques de -99,64 points inférieur.
- significativité du coefficient associé à `redoublant` :
  - la statistique de test vaut -38,66 et est supérieure en valeur absolue au quantile) 97,5 % d'une loi normale centrée réduite : on peut rejeter l'hypothèse de nullité du coefficient au seuil de 5 % ;
  - la p-valeur du test de nullité du coefficient est inférieure à 0,01 : on peut même rejeter l'hypothèse de nullité du coefficient au seuil de 1 % ;
  - la fonction `confint.default()` permet de calculer les intervalles de confiance des coefficients :

```
confint.default(m2)
##                           2.5 %    97.5 %
## (Intercept)          534.99032 545.91030
## immig2               -36.09381 -23.58326
## escsq(-3.46,-0.74]   -34.68068 -20.99516
## escsq(-0.74,-0.21]   -23.99409 -10.57742
## escsq(0.25,0.72]      15.30310  28.65344
## escsq(0.72,2.2]       41.57291  55.06373
```

```
## sexeFemme           -22.66567 -14.19007
## redoublant         -104.69320 -94.58881
## paspreprim          -42.37370 -26.30951
```

0 n'appartient pas à l'intervalle de confiance à 95 % du paramètre associé à `redoublant`, ce qui est une dernière méthode (exactement équivalente aux précédentes) pour conclure que l'on peut rejeter l'hypothèse au seuil de 5 %.

Question 9 Régression logistique dichotomique

Utilisez les variables socio-démographiques disponibles pour modéliser la probabilité d'être en retard scolaire au moment de l'enquête. Pensez bien en particulier :

- à évaluer la qualité de vos modèles à l'aide des indicateurs classiques : log-vraisemblance, AIC, BIC, test du ratio de vraisemblance de significativité globale, pourcentage de concordance, courbe ROC ;
- à interpréter le signe, l'amplitude et la significativité des coefficients ;
- à calculer les *odds-ratio* correspondants et leur intervalle de confiance ;
- à calculer les effets marginaux moyens, à les interpréter ainsi que leur test de significativité.

Proposition de solution

```
# Modélisation
m3 <- glm(
  redoublant ~ sexe + escsq + immig2 + paspreprim
  , data = stusch, family = binomial(link = "logit")
)
```

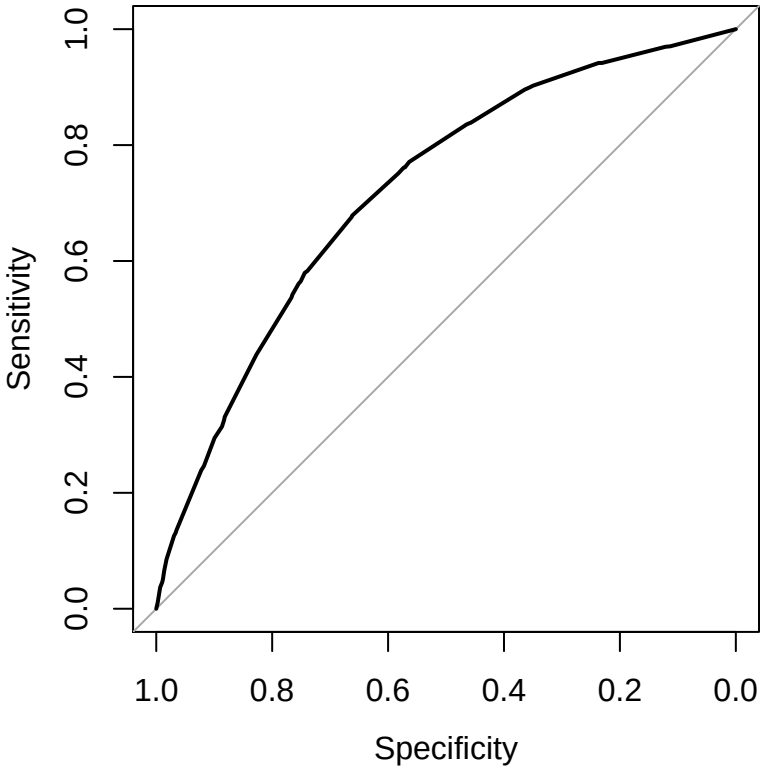
Indicateurs de qualité

```
# - Log-vraisemblance, AIC, BIC
logLik(m3)
## 'log Lik.' -2348.076 (df=8)
AIC(m3)
## [1] 4712.151
BIC(m3)
## [1] 4763.424

# - test de significativité globale par le ratio de vraisemblance
library(lmtest)
lrtest(m3)
## Likelihood ratio test
##
## Model 1: redoublant ~ sexe + escsq + immig2 + paspreprim
```

```
## Model 2: redoublant ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -2348.1
## 2   1 -2623.8 -7 551.55  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# - pourcentage de concordance
conc(m3)
## Pct concordant Pct discordant   Pct ex-aequo
##      69.384488      25.147482       5.468031
```

```
# - courbe ROC
library(pROC)
par(pty="s")
plot(roc(m3$y ~ m3$fitted.values))
```




```
##
## Call:
## roc.formula(formula = m3$y ~ m3$fitted.values)
##
## Data: m3$fitted.values in 3270 controls (m3$y 0) < 1218 cases (m3$y 1).
## Area under the curve: 0.7212
```

— Ici la log-vraisemblance, l'AIC et le BIC ne sont que de peu d'intérêt dans la mesure

où on ne compare pas plusieurs modèles les uns par rapport aux autres.
— La p-valeur du test de significativité globale par le ratio de vraisemblance est inférieure à 0,01 : on peut largement rejeter l'hypothèse que le modèle n'apprend rien sur la variable modélisée.
— Avec environ 70 %, le pourcentage de concordance n'est pas très élevé (on préfère une valeur supérieureà 75 %).
— De façon relativement cohérente, la courbe ROC et l'aire associée (0,7212) ne rendent pas compte d'une très bonne qualité du modèle en matière de prédiction.

Pour conclure : modèle identifiant correctement certaines variables bien associées au retard scolaire (finalité explicative) mais ne présentant pas un pouvoir prédictif très élevé (finalité prédictive).

Interprétation des coefficients

```
summary(m3)
##
## Call:
## glm(formula = redoublant ~ sexe + escsq + immig2 + paspreprim,
##     family = binomial(link = "logit"), data = stusch)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6067  -0.7881  -0.5319   1.0785   2.3014
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.01012    0.08641 -11.689  < 2e-16 ***
## sexeFemme          -0.31738    0.07179  -4.421 9.83e-06 ***
## escsq(-3.46,-0.74]  0.95511    0.10387   9.195  < 2e-16 ***
## escsq(-0.74,-0.21]  0.49486    0.10466   4.728 2.27e-06 ***
## escsq(0.25,0.72]   -0.55680    0.11948  -4.660 3.16e-06 ***
## escsq(0.72,2.2]    -1.24734    0.14205  -8.781  < 2e-16 ***
## immig2              0.27157    0.09643   2.816  0.00486 **
## paspreprim          0.75246    0.12135   6.201 5.63e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5247.7  on 4487  degrees of freedom
## Residual deviance: 4696.2  on 4480  degrees of freedom
##   (125 observations deleted due to missingness)
## AIC: 4712.2
##
```

```
## Number of Fisher Scoring iterations: 5
```

La p-valeur de tous les coefficients est inférieure à 0,01 : quel que soit le coefficient, on peut rejeter au seuil de 1 % l'hypothèse qu'il est nul. Autrement dit, toutes les variables intégrées dans le modèle sont statistiquement liées à la variable expliquée.

Les coefficients dont l'amplitude est la plus forte sont ceux associés au statut économique, social et culturel : à sexe, origine migratoire et scolarité pré-primaire égaux par ailleurs, les élèves les plus favorisés en termes de statut économique, social et culturel ont une probabilité beaucoup plus faible (coefficient -1,25) d'avoir redoublé au cours de leur scolarité par rapport aux élèves de statut économique, social et culturel intermédiaire ; inversement, les élèves les plus défavorisés ont une probabilités beaucoup plus élevée (coefficient 0,96) d'avoir redoublé au cours de leur scolarité. Le fait de ne pas avoir de scolarisation pré-primaire complète est également fortement associé à une probabilité plus élevée d'avoir redoublé à 15 ans, avec un coefficient de 0,75.

```
# Calcul des odds-ratio et de leur IC à 95 %
exp(cbind(or = m3$coefficients, confint.default(m3)))
##                             or     2.5 %    97.5 %
## (Intercept)          0.3641737 0.3074348 0.4313841
## sexeFemme            0.7280558 0.6324949 0.8380547
## escsq(-3.46,-0.74] 2.5989440 2.1202103 3.1857736
## escsq(-0.74,-0.21] 1.6402707 1.3360650 2.0137403
## escsq(0.25,0.72]   0.5730409 0.4534043 0.7242451
## escsq(0.72,2.2]    0.2872686 0.2174558 0.3794945
## immig2               1.3120220 1.0860821 1.5849646
## paspreprim           2.1222070 1.6729822 2.6920565
```

En termes d'*odds-ratio*, on dira par exemple que les élèves les plus défavorisés ont 2,6 plus de chances d'avoir redoublé à un moment ou à un autre de leur scolarité par rapport aux élèves de statut économique, social et culturel intermédiaire, à sexe, origine migratoire et scolarité pré-primaire égaux par ailleurs.

```
# Calcul des effets marginaux moyens
margins(m3)
## $sexeFemme
##                Average MFX Std. Error   z value        P>|z|
## sexeFemme = 0  0.30024454         NA        NA           NA
## sexeFemme = 1  0.24478718         NA        NA           NA
## Diff          -0.05545735 0.01252721 -4.426952 9.557393e-06
##
## $`escsq(-3.46,-0.74]`
##                          Average MFX Std. Error z value        P>|z|
## escsq(-3.46,-0.74] = 0   0.2283160         NA      NA           NA
## escsq(-3.46,-0.74] = 1   0.4152643         NA      NA           NA
## Diff                     0.1869483 0.02179163  8.5789 9.578442e-18
##
```

```
## $`escsq(-0.74,-0.21]`
##                          Average MFX Std. Error  z value        P>|z|
## escsq(-0.74,-0.21] = 0  0.25071833         NA       NA           NA
## escsq(-0.74,-0.21] = 1  0.34082671         NA       NA           NA
## Diff                    0.09010838 0.01973147 4.566734 4.953826e-06
##
## $`escsq(0.25,0.72]`
##                        Average MFX Std. Error   z value        P>|z|
## escsq(0.25,0.72] = 0  0.28924861         NA        NA           NA
## escsq(0.25,0.72] = 1  0.19718906         NA        NA           NA
## Diff                 -0.09205955 0.01845464 -4.988423 6.087423e-07
##
## $`escsq(0.72,2.2]`
##                       Average MFX Std. Error   z value        P>|z|
## escsq(0.72,2.2] = 0   0.3032600         NA        NA           NA
## escsq(0.72,2.2] = 1   0.1192096         NA        NA           NA
## Diff                 -0.1840504 0.01639681 -11.22477 3.081726e-29
##
## $immig2
##            Average MFX Std. Error  z value       P>|z|
## immig2 = 0   0.2632762         NA       NA          NA
## immig2 = 1   0.3123842         NA       NA          NA
## Diff         0.0491080 0.01804905 2.720809 0.006512242
##
## $paspreprim
##                 Average MFX Std. Error z value        P>|z|
## paspreprim = 0   0.2589493         NA      NA           NA
## paspreprim = 1   0.4044567         NA      NA           NA
## Diff             0.1455074 0.02521696 5.77022 7.916814e-09
```

En termes d'effets marginaux moyens, on dira par exemple qu'à sexe, origine migratoire et scolarité pré-primaire égaux par ailleurs les élèves, les plus défavorisés ont 41,53 % de chances d'avoir redoublé, contre 22,83 % pour les autres élèves. Cet écart est significatif au seuil de 1 %.

Présentation des résultats

Question 10 Exportation et mise en forme des résultats

Exportez et mettez en forme vos résultats de façon à pouvoir construire un rapport statistique sur votre étude. En particulier :

— intégrez les résultats numériques dans un logiciel d'édition sous forme de tableaux et présentez-les correctement : numéro de tableau, titre, source, note de lecture, décimales homogènes.
— intégrez les graphiques dans un logiciel d'édition et présentez-les correctement : numéro de figure, titre, source, note de lecture.
— en particulier pour les modèles de régression : choisissez les indicateurs pertinents à représenter (quelques indicateurs d'ajustements, les coefficients, les *odds-ratio* ou les effets marginaux d'un modèle logistique dichotomique) et la manière de faire apparaître la significativité (par des étoiles - rappelez alors les seuils en note de lecture - avec les p-valeurs ou les erreurs standards).

Question 11 Génération automatique de rapport avec Rmarkdown

Consultez attentivement la dernière partie du support complémentaire pour l'atelier sur Rmarkdown puis créez un nouveau fichier Rmarkdown dans RStudio. Après installation des *packages* nécessaires, sauvegardez et compilez une première fois le fichier d'exemple. Modifiez petit à petit cette base pour intégrer les exploitations statistiques et la présentation des résultats de votre étude.