```r
rm(list=ls())
setwd("~/Desktop/regression_avancee")
library(faraway)

## ------------------------------------------------
## QUESTION 1

## Chargement des données
data(gavote)

## Affichage de la tête du tableau de données
head(gavote)

## Résumé statistique (catégorielle vs numérique)
summary(gavote)

## ------------------------------------------------
## QUESTION 2

## nombre de votes nuls = nombres de bulletins - nombre de votes validés
gavote$undercount <- (gavote$ballots - gavote$votes)/gavote$ballots

## Représentation de la dispertion
par(mfrow= c(2,2))
hist(gavote$undercount)
plot(density(gavote$undercount))
rug(gavote$undercount)
plot(ecdf(gavote$undercount))
qqnorm(gavote$undercount)
qqline(gavote$undercount)

## création de la variable pergore (pourcentage de votant pour Gore)
gavote$pergore <- gavote$gore/gavote$votes
plot(pergore ~ perAA, gavote)

## ------------------------------------------------
## QUESTION 3

## régression de pergore par perAA + intercept
model0 <- lm(pergore ~ perAA + 1, gavote)
plot(pergore ~ perAA, gavote, main="régression de perAA sur pergore")
coef0 <- coef(model0)
abline(coef0, col="red")
confint(model0)

## Calcul des coefficients "à la main"
X <- cbind(1,gavote$perAA)
y <- gavote$pergore
XtX <- t(X) %*% X
Xty <- crossprod(y,X)
beta <- solve(XtX) %*% t(Xty)

## ------------------------------------------------
## QUESTION 4

## dispersion de la variable equip
sorted.count.equip <- sort(table(gavote$equip),decreasing=TRUE)
par(mfrow= c(1,2))
barplot(sorted.count.equip, las=3)
pie(sorted.count.equip, col = gray(0:4/4))

## dispersion de la variable ergore en fonction de equip
boxplot(pergore ~ equip, gavote)
anova(lm(pergore ~ equip, gavote))

## ------------------------------------------------
## QUESTION 5

## régression de undercount sur perAA
model1 <- lm(undercount ~ perAA + 1, gavote)
par(mfrow=c(1,2))
plot(undercount ~ perAA + 1, gavote)
abline(coef(model1), col="red")
points(predict(model1) ~ perAA, gavote, pch="+")
plot(residuals(model1))

X <- cbind(1,gavote$perAA)
y <- gavote$undercount
```

```r
XtX <- t(X) %*% X
Xty <- crossprod(y,X)
beta <- solve(XtX) %*% t(Xty)

## calcul de la variance des estimateur
n <- nrow(X)
p <- ncol(X)
res <- y - X%*%beta
sigma <- sqrt( sum(res^2) / (n-p) )
sd.beta <- sqrt(diag(solve(XtX)) * sigma^2)

## statistique de test et p-value
z.score <- beta / sd.beta
p.val   <- 2*pt(t, df=n-p, lower.tail=FALSE)

## ------------------------------------------
## QUESTION 6

## Calcul du R2 et R2 ajusté
RSS <- sum(res^2)
SYY <- sum((y - mean(y))^2)
R2 <- 1 - RSS/SYY
adj.R2 <- 1 - (RSS/(n-p)) / (SYY/(n-1))

## Test de Fisher de la significativité du modèle
F <- (SYY - RSS) / (RSS / (n-p) )
p.val <- pf(F, df1=1, df2=n-p, lower.tail=FALSE)

## ------------------------------------------
## QUESTION 7
model2 <- lm(undercount ~ rural + 1, gavote)
summary(model2)

## ------------------------------------------
## QUESTION 8

## on centre les variables numériques (pas les variables catégorielles)

gavote$cpergore <- gavote$pergore - mean(gavote$pergore)
gavote$cperAA   <- gavote$perAA   - mean(gavote$perAA)
model30 <- lm(undercount ~ equip + cpergore + cperAA + rural, gavote)
summary(model30)
anova(model30)

## ------------------------------------------
## QUESTION 9

## On regroupe les variables equipements
## en deux classes uniquement : OSPC eet PUNCh (significatives)
## et les autres
gavote$OSPC.PUNCH <- rep(FALSE, n)
gavote$OSPC.PUNCH[gavote$equip %in% c("OS-PC", "PUNCH")] <- TRUE
model3 <- lm(undercount ~ OSPC.PUNCH + cpergore + cperAA + rural, gavote)
summary(model3)
anova(model3)


## ------------------------------------------
## QUESTION 10

## choix de modèle à l'aide du critère AIC (à maximiser, en backward)
step(model3)
model4 <- lm(undercount ~ OSPC.PUNCH + cperAA + rural, gavote)
summary(model4)
anova(model4)
anova(model30, model4)

## ------------------------------------------
## QUESTION 11

## diagnostic de la régression final
par(mfrow=c(2,3))
plot(model4, which=1:6)

set <- !(rownames(gavote) %in% c("BEN.HILL","RANDOLPH", "BACON", "MARION","TELFAIR","TAYLOR",
"TREUTLEN", "EFFINGHAM", "APPLING","EMANUEL"))
model5 <- lm(undercount ~ OSPC.PUNCH + cperAA + rural, gavote, subset=set)
```

```
summary(model5)
par(mfrow=c(2,3))
plot(model5, which=1:6)
```