ACT6100	Analyse de données
H2019	**Solutions 4**

1. (a) Étant donné que la variable réponse est binaire (0 ou 1), l'hypothèse de normalité ne sera pas respectée. Il faut alors utiliser un modèle de régression binaire (logistique ou autre).

   (b) Le modèle complet est présenté à la Table 1 et le modèle final est présenté à la Table 2. Le code est présenté à la Figure 1.

Variable	Estimation ($\hat{\beta}$)	valeur p
Ordonnée	-0.0169	0.967
Age	0.0335	0.042
Genre_femme	2.0952	0.000
Nexp	0.0973	0.000
Prof_chomeurs	-0.7581	0.002
Prof_prive	0.7483	0.002
Prof_public	-0.4522	0.085
Rabanque	-0.0241	0.495
Revenu	0.0040	0.040
Tendett	-0.1629	0.000

Table 1 – Modèle complet

Variable	Estimation ($\hat{\beta}$)	valeur p
Ordonnée	0.2031	0.523
Age	0.0260	0.003
Genre_femme	2.0516	0.000
Nexp	0.1062	0.000
Prof_chomeurs	-0.7031	0.003
Prof_prive	0.8611	0.000
Prof_public	-0.2810	0.259
Tendett	-0.1624	0.000

Table 2 – Modèle final

   (c) Bien que la valeur p associée à **Prof_public** soit supérieure à 0.05, on ne peut la retirer du modèle. En effet, il s'agit d'une modalité de la variable **Prof** et on ne peut retirer une seule modalité : soit on retire toutes les modalités, soit on garde toutes les modalités. On pourrait éventuellement tenter de la regrouper avec une autre modalité.

   (d) Bien que la valeur p associée à l'ordonnée à l'origine soit supérieure à 0.05, on ne peut la retirer du modèle. En effet, l'élément *Intercep* du modèle comprend non seulement l'ordonnée à l'origine (le «vraie» $\hat{\beta}_0$), mais également l'effet de référence pour chacune des variables catégorielles (homme pour la variable **Genre** et libérale pour la variable **prof**). Ceci est dû au fait que l'encodage est réalisé à l'aide de 0/1 plutôt que de $-1/1$.

   (e) Le code est présenté à la Figure 2. On obtient un point de rupture environ égal à 0.63 (voir Figure 3).

   (f) Il faut d'abord calculer le score pour le nouvel individu :

$$
\begin{aligned}
\text{Score} &= 0.203 + (0.026)\text{Age} + (2.052)\text{Genre} + (0.106)\text{Nexp} \\
&\quad - (0.703)\text{Prof\_chomeur} + (0.861)\text{Prof\_prive} - (0.281)\text{Prof\_public} - (0.162)\text{Tendett} \\
&= 0.203 + (0.026)(35) + (0.106)(10) - (0.281)(1) - (0.162)(25) \\
&= -2.158.
\end{aligned}
$$

```
### Modèle complet
modele1 <- glm(as.factor(Statut) ~ Age + Revenu + Tendett + Nexp +
               Rabanque + as.factor(Prof) + as.factor(Genre),
               family = binomial(link = "logit"), data = credit)
summary(modele1)

### Sélection du modèle (critère BIC)
library(MASS)
n <- length(credit$Statut)
stepAIC(modele1, direction = "backward", k = log(n))

### Modèle final
modele2 <- glm(formula = as.factor(Statut) ~ Age + Tendett + Nexp
               + as.factor(Prof) + as.factor(Genre),
               family = binomial(link = "logit"), data = credit)
summary(modele2)
```

FIGURE 1 – Code informatique.

```
### 10-validation croisée
set.seed(100)
n <- length(credit$Statut)
indice <- matrix(sample(1:n, n, replace = FALSE), nrow = 10)

FUN <- function(tau)
{
  mod <- glm(formula = as.factor(Statut) ~ Age + Tendett + Nexp + as.factor(Prof) +
             as.factor(Genre), family = binomial(link = "logit"), data = dataTrain)
  pred <- predict(mod, type = "response", newdata = dataValidation)
  dataValidation$p <- pred > tau
  dtest <- dataValidation[dataValidation$Statut == 0,]
  sensibilite <- sum(dtest$Statut == dtest$p)/length(dtest$Statut)
  dtest1 <- dataValidation[dataValidation$Statut == 1,]
  specificite <- sum(dtest1$Statut == dtest1$p)/length(dtest1$Statut)
  c(sensibilite, specificite)
}

FUN2 <- function(x){
  train <- as.vector(indice[-x, ])
  valid <- as.vector(indice[x, ])
  dataTrain <<- credit[train, ]
  dataValidation <<- credit[valid, ]
  sapply((0:100)/100, function(y) FUN(y))
}

Taux <- sapply(1:10, function(x) FUN2(x))
TauxSe <- Taux[(0:100)*2 + 1,]
TauxSp <- Taux[(1:101)*2,]
TauxSe <- rowMeans(TauxSe)
TauxSp <- rowMeans(TauxSp)
```

FIGURE 2 – Code informatique.

On calcule ensuite la probabilité d'obtenir le crédit :

$$\begin{aligned}\Pr(\text{Statut} = 1) &= \frac{e^{-2.158}}{1 + e^{-2.158}} \\ &= 0.1036.\end{aligned}$$

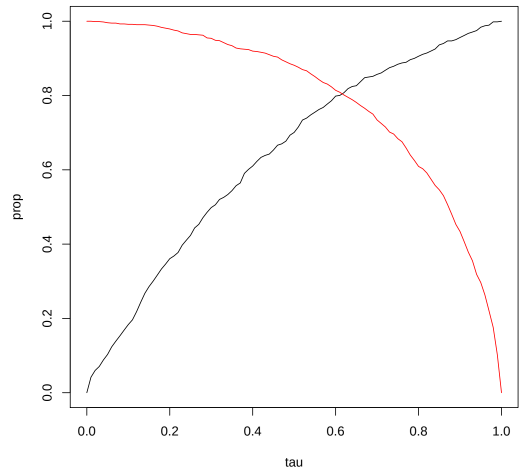

FIGURE 3 – Comportement de la sensibilité (noire) et de la spécificité (rouge) pour différentes valeurs du point de rupture.

Avec la valeur de $\tau^* = 0.63$ trouvée à la sous-question précédente, le nouvel assuré n'obtiendra pas son crédit.

2. (a) Les estimations des coefficients du modèle complet et les valeurs p sont données dans la Table 3. Le code permettant d'obtenir le modèle complet est présenté à la Figure 4.

Variable	Estimation ($\hat{\beta}$)	valeur p
Intercept	-2.536	0
veh_value	0.049	0
veh_age(2)	0.159	0
veh_age(3)	0.071	0.15
veh_age(4)	0.017	0.76
Genre(M)	-0.015	0.63
area(B)	0.099	0.03
area(C)	0.039	0.35
area(D)	-0.093	0.10
area(E)	-0.024	0.69
area(F)	0.106	0.13
agecat(2)	-0.197	0
agecat(3)	-0.221	0
agecat(4)	-0.250	0
agecat(5)	-0.439	0
agecat(6)	-0.444	0

TABLE 3 – Modèle complet

```
modele1 <- glm(as.factor(clm) ~ veh_value + as.factor(veh_age) + gender +
               area + as.factor(agecat), family = binomial(link = "logit"),
               data = carsClaims)
summary(modele1)
```

FIGURE 4 – Code informatique.

(b) Les estimations des coefficients du modèle final et les valeurs p sont données dans la Table 4. Le code permettant d'obtenir le modèle final est présenté à la Figure 4.

Variable	Estimation ($\hat{\beta}$)	valeur p
Intercept	-2.454	0
veh_value	0.052	0
agecat(2)	-0.193	0
agecat(3)	-0.222	0
agecat(4)	-0.253	0
agecat(5)	-0.444	0
agecat(6)	-0.455	0

Table 4 – Modèle final

```
library(MASS)
n <- length(carsClaims$clm)
stepAIC(modele1, direction = "backward", k = log(n))
modele2 <- glm(formula = as.factor(clm) ~ veh_value + as.factor(agecat),
               family = binomial(link = "logit"), data = carsClaims)
summary(modele2)
```

Figure 5 – Code informatique.

(c) On calcule le score de la cliente à l'aide du modèle final :

$$\begin{aligned}\text{Score} &= -2.454 + (0.052)(2.2) - (0.455)(1)\\ &= -2.7946.\end{aligned}$$

La probabilité d'avoir une réclamation est

$$\begin{aligned}\Pr(\text{clm} = 1) &= \frac{e^{-2.7946}}{1 + e^{-2.7946}}\\ &= 0.0576.\end{aligned}$$

La prime est alors donnée par

$$\begin{aligned}\Pi &= (2000)(0.0576)(1 + 0.25)\\ &= 144.\end{aligned}$$

(d) Le code est présenté à la Figure 6. On obtient un point de rupture environ égal à 0.069 (voir Figure 7).

3. — Dans l'énoncé, on a les équations

$$p(k) = \frac{e^k}{1 + e^k} \tag{1}$$

et

$$\text{logit}(k) = \ln\left(\frac{k}{1-k}\right). \tag{2}$$

Pour réaliser la démonstration, on utilise l'équation (1) et le fait que $a^{-x}a^x = 1$. On a alors

$$\begin{aligned}p(-k) &= \frac{e^{-k}}{1 + e^{-k}}\\ &= \left(\frac{e^{-k}}{1 + e^{-k}}\right)\left(\frac{e^k}{e^k}\right)\\ &= \frac{1}{e^k + 1}\\ &= 1 - \frac{e^k}{1 + e^k}.\end{aligned}$$

```
FUN <- function(tau)
{
  mod <- glm(formula = as.factor(clm) ~ veh_value + as.factor(agecat),
             family = binomial(link = "logit"), data = dataTrain)
  pred <- predict(mod, type = "response", newdata = dataValidation)
  dataValidation$p <- pred > tau
  dtest <- dataValidation[dataValidation$clm == 0,]
  sensibilite <- sum(dtest$clm == dtest$p)/length(dtest$clm)
  dtest1 <- dataValidation[dataValidation$clm == 1,]
  specificite <- sum(dtest1$clm == dtest1$p)/length(dtest1$clm)
  c(sensibilite, specificite)
}

set.seed(100)
n <- length(carsClaims$clm)
indice <- matrix(sample(1:n, n, replace = FALSE), nrow = 16)

FUN2 <- function(x){
  train <- as.vector(indice[-x, ])
  valid <- as.vector(indice[x, ])
  dataTrain <<- carsClaims[train, ]
  dataValidation <<- carsClaims[valid, ]
  sapply((50:150)/100, function(y) FUN(y))
}

Taux <- sapply(1:16, function(x) FUN2(x))
TauxSe <- Taux[(0:100)*2 + 1,]
TauxSp <- Taux[(1:101)*2,]
TauxSe <- rowMeans(TauxSe)
TauxSp <- rowMeans(TauxSp)
```

FIGURE 6 – Code informatique.

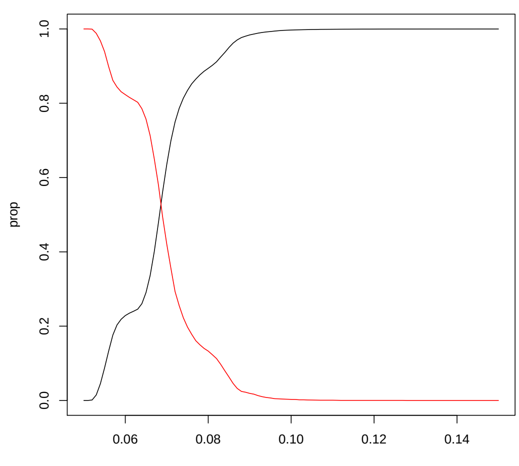

FIGURE 7 – Comportement de la sensibilité (noire) et de la spécificité (rouge) pour différentes valeurs du point de rupture.

— On rappelle que

$$\frac{d(f(x)/g(x))}{dx} = \frac{f'(x)g(x) - g'(x)f(x)}{g(x)^2},$$

où $f'(x) = df(x)/dx$ et $g'(x) = dg(x)/dx$. On a alors

$$\begin{aligned}
\frac{dp(k)}{dk} &= \frac{d\left(\frac{e^k}{1+e^k}\right)}{dk} \\
&= \frac{e^k(1+e^k) - e^k e^k}{(1+e^k)^2} \\
&= \frac{e^k}{(1+e^k)^2} \\
&= \left(\frac{e^k}{1+e^k}\right)\left(\frac{1}{1+e^k}\right) \\
&= p(k)(1-p(k)).
\end{aligned}$$

— On a

$$\begin{aligned}
\frac{dp(SCORE)}{dx_j} &= \frac{d\left(\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_q x_q}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_q x_q}}\right)}{dx_j} \\
&= \frac{e^{\beta_0+\ldots+\beta_q x_q}\beta_j(1+e^{\beta_0+\ldots+\beta_q x_q}) - e^{\beta_0+\ldots+\beta_q x_q}\beta_j e^{\beta_0+\ldots+\beta_q x_q}}{(1+e^{\beta_0+\ldots+\beta_q x_q})^2} \\
&= \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_q x_q}}{(1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_q x_q})^2}\beta_j \\
&= \left(\frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_q x_q}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_q x_q}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_q x_q}}\right)\beta_j \\
&= p(\beta_0+\beta_1 x_1+\ldots+\beta_q x_q)(1-p(\beta_0+\beta_1 x_1+\ldots+\beta_q x_q))\beta_j \\
&= p(SCORE)(1-p(SCORE))\beta_j,
\end{aligned}$$

pour $j = 1, \ldots, q$.

— Pour réaliser la démonstration, on utilise l'équation (2) et la propriété $\ln(x^a) = a\ln(x)$, où a est une constante. On obtient

$$\begin{aligned}
\text{logit}(1-k) &= \ln\left(\frac{1-k}{1-(1-k)}\right) \\
&= \ln\left(\frac{1-k}{k}\right) \\
&= \ln\left(\left(\frac{k}{1-k}\right)^{-1}\right) \\
&= -\ln\left(\frac{k}{1-k}\right) \\
&= -\text{logit}(k).
\end{aligned}$$

— On rappelle que

$$\frac{d(\ln(f(x)))}{dx} = \left(\frac{1}{f(x)}\right) f'(x),$$

avec $f'(x) = df(x)/dx$. On a

$$\begin{aligned}
\frac{d\text{logit}(k)}{dk} &= \left(\frac{1-k}{k}\right)\frac{d\left(\frac{k}{1-k}\right)}{dk} \\
&= \left(\frac{1-k}{k}\right)\left(\frac{(1)(1-k)-(-1)(k)}{(1-k)^2}\right) \\
&= \frac{1}{k(1-k)}.
\end{aligned}$$

4. (a) Afin d'obtenir la fonction de log-vraisemblance, on utilisera les propriétés $\ln(AB) = \ln(A) + \ln(B)$ et $\ln(x^a) = a\ln(x)$. On a

$$\begin{aligned}
l(\beta_0, \beta) &= \ln(L(\beta_0, \beta)) \\
&= \ln(\prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}) \\
&= \sum_{i=1}^{n} \ln(p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}) \\
&= \sum_{i=1}^{n} y_i \ln(p(x_i)) + (1 - y_i)\ln(1 - p(x_i)).
\end{aligned}$$

(b) On a, avec $S = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}$,

$$\begin{aligned}
\frac{dl(\beta_0, \beta)}{d\beta_0} &= \sum_{i=1}^{n} \frac{dy_i \ln(p(x_i))}{d\beta_0} + \frac{d(1 - y_i)\ln(1 - p(x_i))}{d\beta_0} \\
&= \sum_{i=1}^{n} (y_i) \left(\frac{1 + e^S}{e^S}\right) \left(\frac{e^S(1)(1 + e^S) - e^S(1)e^S}{(1 + e^S)^2}\right) \\
&\quad + (1 - y_i) \left(\frac{1 + e^S}{1}\right) \left(\frac{0 - e^S(1)}{(1 + e^S)^2}\right) \\
&= \sum_{i=1}^{n} y_i \left(\frac{1}{1 + e^S}\right) + (1 - y_i) \left(\frac{-e^S}{1 + e^S}\right) \\
&= \sum_{i=1}^{n} y_i(1 - p(x_i)) - (1 - y_i)p(x_i) \\
&= \sum_{i=1}^{n} y_i - p(x_i),
\end{aligned}$$

et

$$\begin{aligned}
\frac{dl(\beta_0, \beta)}{d\beta_j} &= \sum_{i=1}^{n} \frac{dy_i \ln(p(x_i))}{d\beta_j} + \frac{d(1 - y_i)\ln(1 - p(x_i))}{d\beta_j} \\
&= \sum_{i=1}^{n} (y_i) \left(\frac{1 + e^S}{e^S}\right) \left(\frac{e^S(x_{ij})(1 + e^S) - e^S(x_{ij})e^S}{(1 + e^S)^2}\right) \\
&\quad + (1 - y_i) \left(\frac{1 + e^S}{1}\right) \left(\frac{0 - e^S(x_{ij})}{(1 + e^S)^2}\right) \\
&= \sum_{i=1}^{n} y_i \left(\frac{x_{ij}}{1 + e^S}\right) + (1 - y_i) \left(\frac{-e^S x_{ij}}{1 + e^S}\right) \\
&= \sum_{i=1}^{n} x_{ij} y_i(1 - p(x_i)) - x_{ij}(1 - y_i)p(x_i) \\
&= \sum_{i=1}^{n} x_{ij} y_i - x_{ij} p(x_i).
\end{aligned}$$

(c) On a

$$
\begin{aligned}
\sum_{i=1}^{n} y_i - p(x_i) &= 0 \\
\sum_{i=1}^{n} y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}} &= 0 \\
\sum_{i=1}^{n} y_i &= \frac{ne^{\beta_0}}{1+e^{\beta_0}} \\
\sum_{i=1}^{n} y_i + e^{\beta_0} \sum_{i=1}^{n} y_i &= ne^{\beta_0} \\
\sum_{i=1}^{n} y_i &= e^{\beta_0} \left(n - \sum_{i=1}^{n} y_i \right) \\
\ln \left(\sum_{i=1}^{n} y_i \right) &= \beta_0 + \ln \left(n - \sum_{i=1}^{n} y_i \right) \\
\hat{\beta}_0 &= \ln \left(\sum_{i=1}^{n} y_i \right) - \ln \left(n - \sum_{i=1}^{n} y_i \right) \\
&= \ln \left(\frac{\sum_{i=1}^{n} y_i}{n - \sum_{i=1}^{n} y_i} \right).
\end{aligned}
$$