Projet modèle de régression

Présentation des données

Le fichier **Assurance.txt** contient des données recueillies par une assurance concernant des ménages assurés habitant Paris et son agglomération : des données personnelles du ménage, les montants des dépenses du ménage pour se couvrir contre certains risques et les montants des dommages de sinistres pour trois types de risques.

Question A

Préciser le nombre de ménages assurés étudiés. Définir le type de chacune des variables recueillies et résumer numériquement ces variables sur l'ensemble des ménages.
En particulier,

- quelle est la distribution d'effectifs observée de la CSP ?
- quels sont les montants minimum, maximum, moyen, médian observés, la variance et l'écart-type observés des dommages des sinistres de type 1 ?

On dispose de n = 235 ménages observés pour lesquels ont été relevées 16 variables : 8 variables quantitatives continues REVENU, RUC, POL1, POL2, POL3, DOM1, DOM2, DOM3, 3 variables quantitatives discrètes NBPERS, NBAD, NBSIN et 5 variables qualitatives CSP, CR, STOCC, COMP et AUTO.
Les distributions d'effectifs des variables qualitatives montrent des répartitions assez uniformes pour la CSP, variable qualitative à 3 modalités : *Cadres et prof. intellectuelles sup.*, *Employes*, *Professions intermediaires* (cf Tableau A.1) la catégorie de revenu, variable qualitative à 3 modalités : *Aise*, *Moyenne Inf*, *Moyenne Sup* (cf Tableau A.2 légèrement moins de catégorie *Moyenne Inf*) le statut d'occupation de l'habitation, variable qualitative à 2 modalités : *Locataire*, *Proprietaire* (cf Tableau A.3) la composition du ménage, variable qualitative à 3 modalités : *Couple avec enfant(s)*, *Couple sans enfant*, *Personne seule* (cf Tableau A.4 un peu moins de ménages *Couple sans enfant*) ; mais pour la possession d'un véhicule, variable qualitative à 2 modalités : *Au - 1 vehicule*, *Pas de vehicule*, presque quatre fois plus de ménages possèdent *Au - 1 vehicule* (cf Tableau A.5).

Le revenu observé varie de 5500 à 60000, est en moyenne de 1.7609×10^4 ; le revenu médian observé est de 16250 (inférieur au revenu moyen observé), le 1er quartile observé est de 1.125×10^4 et le 3ème de 2.25×10^4 (cf Tableau A.6).
Le nombre observé de personnes par ménage varie de 1 à 7, est en moyenne de 2.532 ; le nombre médian observé est de 2, le 1er quartile observé de 1 et le 3ème de 4 (cf Tableau A.6). Le nombre observé d'adultes par ménage est de 1 ou 2 (cf Tableau A.6).
Le nombre observé de sinistres antérieurs varie de 0 à 15, est en moyenne de 3.821 ; le nombre médian observé est de 4, le 1er quartile observé de 2 et le 3ème de 5 (cf Tableau A.6).
Le montant observé des dommages des sinistres de type 1 (cf Tableau A.6) varie de 8.862 à 18.943, est en moyenne de 13.914 avec une variance observée de 3.333 et un écart-type observé de 1.826 ; le montant médian des dommages des sinistres de type 1 observé est de 13.768 (proche de la moyenne observée).

Les commandes R *suivantes permettent de stocker les données du fichier* Assurance.txt *dans l'objet* données*, d'afficher le nombre d'observations, le nombre, les noms et types des variables.*
Les résultats des deux dernières commandes ne sont pas affichés.

```r
# lecture du fichier
données <- read.table("Assurance.txt") #, header=TRUE)
nrow(données) # nombre d'observations

[1]  235
```

```r
str(données)
head(données)
```

Les résumés numériques des variables quantitatives (minimum, maximum, quartiles et moyenne observés) sont donnés par la commande R ci-dessous :

```r
summary(données) # résumés des variables

      CSP                 REVENU            RUC              CR
 Length:235          Min.   : 5500    Min.   : 3519   Length:235
 Class :character    1st Qu.:11250    1st Qu.: 6944   Class :character
 Mode  :character     Median :16250    Median : 8523   Mode  :character
                     Mean   :17609    Mean   :10331
                     3rd  Qu.:22500   3rd  Qu.:11250
                     Max.   :60000    Max.   :35294
     STOCC               COMP              NBPERS          NBAD
 Length:235          Length:235          Min.   :1.000   Min.   :1.000
 Class :character    Class :character    1st Qu.:1.000   1st Qu.:1.000
 Mode  :character    Mode  :character    Median :2.000   Median :2.000
                                         Mean   :2.532   Mean   :1.685
                                         3rd  Qu.:4.000  3rd  Qu.:2.000
                                         Max.   :7.000   Max.   :2.000
     AUTO                POL1              POL2             POL3
 Length:235          Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000
 Class :character    1st Qu.: 0.270   1st Qu.: 2.718   1st Qu.: 0.4025
 Mode  :character    Median : 1.512   Median : 6.875   Median : 0.9750
                     Mean   : 2.845   Mean   :10.508   Mean   : 1.5873
                     3rd Qu.: 3.783   3rd Qu.:13.540   3rd Qu.: 1.9650
                     Max.   :18.720   Max.   :75.635   Max.   :21.3050
      DOM1             DOM2             DOM3             NBSIN
 Min.   : 8.862   Min.   : 4.192   Min.   : 0.000   Min.   : 0.000
 1st Qu.:12.670   1st Qu.: 9.067   1st Qu.: 0.000   1st Qu.: 2.000
 Median :13.768   Median :10.440   Median : 0.520   Median : 4.000
 Mean   :13.914   Mean   :10.421   Mean   : 1.204   Mean   : 3.821
 3rd Qu.:15.109   3rd Qu.:12.621   3rd Qu.: 1.610   3rd Qu.: 5.000
 Max.   :18.943   Max.   :15.476   Max.   :12.220   Max.   :15.000

attach(données)
```

Les distributions des effectifs observées des variables qualitatives sont données par les commandes R suivantes :

```r
# distributions d'effectifs observées des variables qualitatives
table(CSP); table(CR); table(STOCC); table(COMP); table(AUTO)
# résumés numériques des variables quantitatives
summary(data.frame(REVENU,RUC,NBPERS,NBAD,  DOM1,NBSIN))
```

Tableau A.1

CSP	Freq
Cadres et prof. intellectuelles sup.	77
Employes	85
Professions intermediaires	73

Tableau A.2

CR	Freq
Aise	86
Moyenne Inf	53
Moyenne Sup	96

Tableau A.3

STOCC	Freq
Locataire	126
Proprietaire	109

Tableau A.4

COMP	Freq
Couple avec enfant(s)	106
Couple sans enfant	55
Personne seule	74

Tableau A.5

AUTO	Freq
Au - 1 vehicule	186
Pas de vehicule	49

Tableau A.6

REVENU	RUC	NBPERS	NBAD	DOM1	NBSIN
Min. : 5500	Min. : 3519	Min. :1.000	Min. :1.000	Min. : 8.862	Min. : 0.000
1st Qu. :11250	1st Qu. : 6944	1st Qu. :1.000	1st Qu. :1.000	1st Qu. :12.670	1st Qu. : 2.000
Median :16250	Median : 8523	Median :2.000	Median :2.000	Median :13.768	Median : 4.000
Mean :17609	Mean :10331	Mean :2.532	Mean :1.685	Mean :13.914	Mean : 3.821
3rd Qu. :22500	3rd Qu. :11250	3rd Qu. :4.000	3rd Qu. :2.000	3rd Qu. :15.109	3rd Qu. : 5.000
Max. :60000	Max. :35294	Max. :7.000	Max. :2.000	Max. :18.943	Max. :15.000

Les variances observées (biaisées) et les écart-types observés des variables quantitatives sont donnés par les commandes R *suivantes :*

```
# variance et écart-type observés de DOM1
var(DOM1)*(nrow(données)-1)/nrow(données); sd(DOM1)*sqrt((nrow(données)-1)/nrow(données))

[1] 3.33277
[1] 1.825588
```

Question B

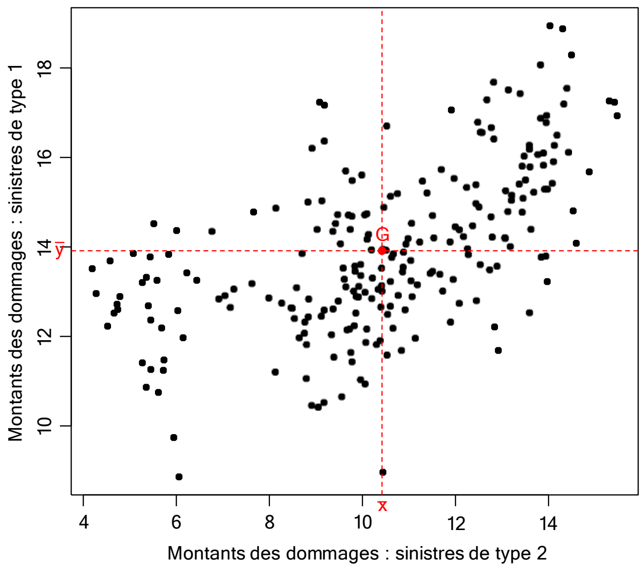
On s'intéresse à la relation entre le montant des dommages des sinistres de type 1 et celui des dommages des sinistres de type 2.

1. Représenter le nuage de points du montant des dommages des sinistres de **type 1** en fonction de celui des dommages des sinistres de **type 2**, ainsi que son centre de gravité.
   Calculer puis commenter le cœfficient de corrélation linéaire observé entre les deux variables.

   Le nuage de points (cf Graphique B.1) est réalisé avec les commandes R ci-dessous :

```
# Nuage de points de DOM1 en fonction de DOM2
plot(DOM2,DOM1, xlab="Montants des dommages : sinistres de type 2",
     ylab="Montants des dommages  :  sinistres de type  1", pch=20)
points(mean(DOM2),mean(DOM1), col='red', pch=16) # centre de gravité du nuage
abline(v=mean(DOM2), h=mean(DOM1), col='red', lty=2)
text(mean(DOM2), mean(DOM1), "G", col='red', pos=3)
text(mean(DOM2), min(DOM1)-0.7, expression(bar(x)), col='red', xpd=TRUE)
text(min(DOM2)-0.7, mean(DOM1), expression(bar(y)), col='red', xpd=TRUE)
```

Graphique B.1



Le nuage de points et son centre de gravité, point G de coordonnées $G = (\overline{\text{DOM2}}, \overline{\text{DOM1}})$ Ä $(10.42,\ 13.91)$ sont représentés sur le Graphique B.1.
Le nuage de points (cf Graphique B.1) a une forme relativement linéaire pour les valeurs de DOM2 supérieures à 8 mais il n'est pas très homogène ; la linéarité est moins observée pour les valeurs de DOM2 plus faibles.
Le cœfficient de corrélation observé entre les montants des dommages des sinistres de type 1 et 2 vaut $r(\text{DOM1}, \text{DOM2})$ Ä 0.5809 : il est positif, d'intensité modérée. De prime abord, on observe une corrélation positive modérée entre les montants des dommages des deux types de sinistres : ils varient dans le même sens.
Les résultats numériques ont été obtenus grâce aux commandes R ci-dessous :

```
mean(DOM1); mean(DOM2) # moyenne de DOM1 et DOM2

[1]  13.914
[1]  10.42121

cor(DOM1,DOM2)  #  cœfficient  de  corrélation  entre  DOM1  et  DOM2

[1]  0.5808849
```

2. On considère le modèle de régression linéaire simple du montant des dommages des sinistres de **type 1** sur celui des dommages des sinistres de **type 2** et la constante (modèle (1)).

   (a) Écrire le modèle correspondant, donner sa forme matricielle et préciser ses hypothèses.

En notant $y_1, \dots, y_n$ les montants observés des dommages des sinistres de type 1, $Y_1, \dots, Y_n$ les variables aléatoires correspondantes et $x_1, \dots, x_n$ les montants observés des dommages des sinistres de type 2 pour les $n = 235$ ménages assurés, le modèle de régression linéaire simple du montant des dommages des sinistres de type 1 sur celui des dommages des sinistres de type 2 et la constante (modèle (1)) s'écrit :

(1) $\quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad$ pour $i = 1, \dots, n$

ou, de manière matricielle : (1) $\quad Y = X\beta + \varepsilon$

- $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ étant les vecteurs aléatoires $(n, 1)$
- $X = (1_n \;\; x) = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ la matrice (n, p) du modèle, de rang $p = 2$ puisqu'il existe au moins deux valeurs différentes de x vecteur du montant des dommages des sinistres de type 2, variable DOM2 : en effet, la variance observée de DOM2 $s_x^2 \approx 7.32 \neq 0$
- $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ le vecteur $(p, 1)$ des cœfficients inconnus du modèle, à estimer.

Les hypothèses probabilistes d'un modèle linéaire gaussien sont que, conditionnellement à la variable explicative DOM2, ou DOM2 étant considérée comme déterministe, les erreurs ε_i sont indépendantes deux à deux, centrées, de même variance et de loi normale : $\varepsilon \sim \mathcal{N}_n\left(0_n; \sigma^2 I_n\right)$ ou ε_i i.i.d. de loi $\mathcal{N}\left(0; \sigma^2\right)$ pour tout $i = 1, \dots, n$.

La variance des erreurs σ^2 inconnue doit être estimée.

Le modèle (1) *est implémenté dans* R *et stocké dans l'objet* mod1 *avec la commande suivante :*

```r
# modèle (1) : régression simple de DOM1 sur DOM2
mod1 <- lm(DOM1 ~ DOM2)
```

* L'hypothèse d'indépendance des erreurs découle de celle des observations des ménages assurés, et les conditions de linéarité et d'homoscédasticité des erreurs sont suggérées par la linéarité et l'homogénéité de la dispersion du nuage de points : ces deux conditions étant relativement peu établies, elles devront être vérifiées a posteriori à partir des résidus.

L'hypothèse de normalité des erreurs est vérifiée a priori si la variable à expliquer DOM1 peut être considérée comme gaussienne ; graphiquement l'histogramme observé de la variable DOM1 (cf Graphique B.2a) est légèrement décalé par rapport à celui d'une loi gaussienne de mêmes moyenne et écart-type que DOM1, la droite de Henry (Q-Q plot) de la variable DOM1 (cf Graphique B.2b) montre des points assez alignés, mais le test de normalité de Shapiro-Wilk pour cette variable rejette l'hypothèse nulle de normalité de la variable DOM1 au risque $\alpha = 20\%$ puisque la p-valeur du test $0.1271 > \alpha = 0.2$ (il est préférable de faire un test plus puissant en augmentant le seuil d'erreur de première espèce α par exemple à 20%).

Le nombre d'observations $n = 235$ étant grand, il est néanmoins acceptable de considérer des modèles linéaires sur la variable DOM1.

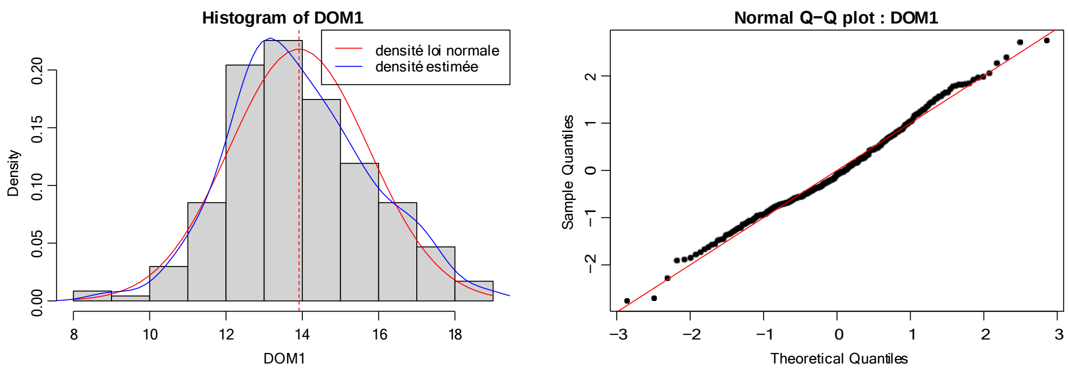
L'hypothèse de normalité des erreurs devra être vérifiée a posteriori à partir des résidus.

```r
hist(DOM1, freq=F) # histogramme de DOM1
curve(dnorm(x,mean(DOM1),sd(DOM1)), add=T, col='red') # densité loi normale
abline(v=mean(DOM1), lty=2, col='red')
lines(density(DOM1), col='blue') # densité estimée de DOM1
legend('topright', legend=c('densité loi normale', 'densité estimée'),
       col=c('red','blue'), lty=1)
```

```r
qqnorm(scale(DOM1), main='Normal Q-Q plot : DOM1", pch=20)
abline(0,1, col='red') # droite d'équation y = x
shapiro.test(DOM1) # test de normalité de DOM1
```

```
Shapiro-Wilk normality test

data:    DOM1
W = 0.99048, p-value = 0.1271
```

Graphique B.2



(a) **(b)**

(b) Donner les estimations des moindres carrés des cœfficients et l'estimation de la variance des erreurs du modèle (1).

L'estimateur des moindres carrés (EMC) des cœfficients : $\hat{\beta} = ({}^tXX)^{-1}\,{}^tXY$ est un estimateur sans biais de β ;

les estimations de β_0 et β_1 : $\hat{\beta}_0 \approx 9.8212$ et $\hat{\beta}_1 \approx 0.3927$

L'estimateur sans biais de la variance des erreurs : $S^2 = \dfrac{\|\hat{\varepsilon}\|^2}{n-p} = \dfrac{SSE}{n-p}$

où $SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$ est la somme des carrés résiduelle,

et son estimation $s^2 \approx 2.2272$

La pente de la droite de régression de la variable DOM1 sur la variable DOM2 est estimée à 0.3927 et le terme constant à 9.8212 ; la variance des erreurs du modèle (1) est estimée à 2.2272.

Les résultats numériques ont été obtenus grâce aux commandes R ci-dessous :

```r
mod1$coef # cœfficients estimés du modèle (1)

(Intercept)        DOM2
  9.8211643   0.3927412

sigma(mod1)^2 # variance estimée des erreurs du modèle (1)

[1] 2.227157
```

(c) Donner l'équation de la droite de régression observée, puis la représenter graphiquement sur le nuage de points.

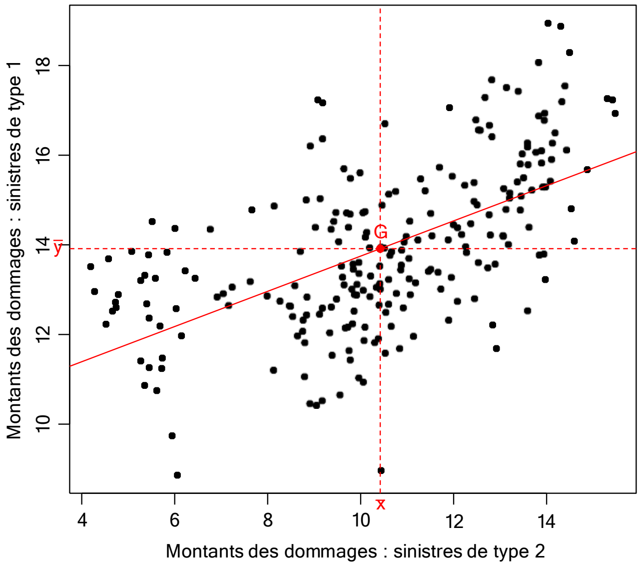
La droite de régression observée a pour équation : $y = \hat{\beta}_0 + \hat{\beta}_1 x \approx 9.8212 + 0.3927\,x$

La droite de régression observée passe par le point de coordonnées $(x = 0, y = 9.8212)$ et par le centre de gravité du nuage, puisque $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ (cf Graphique B.3) ;
lorsque le montant des dommages des sinistres de type 2 augmente de 1 unité, le modèle (1) estime que celui des dommages des sinistres de type 1 augmente de 0.3927 unité.

La droite de régression observée est ajoutée au nuage de points (cf Graphique B.3) avec la commande R ci-dessous :

```
abline(mod1, col='red') # droite de régression du modèle (1)
```

Graphique B.3



(d) La pente de la droite de régression est-elle significativement positive ? Préciser les hypothèses testées et la statistique de test utilisée, sa loi sous l'hypothèse nulle, relever sa valeur observée et donner la p-valeur ; indiquer la décision prise et le risque d'erreur encouru.

On teste l'hypothèse nulle $H_0 : \beta_1 = 0$ contre l'alternative unilatérale droite $H_1 : \beta_1 > 0$ au niveau de risque α ; la statistique de test de Student $T_1 = \frac{\hat\beta_1}{S\sqrt{\Gamma_{1,1}}} = \frac{\hat\beta_1}{\sqrt{\widehat{\text{var}}(\hat\beta_1)}}$ suit sous H_0 une loi de Student $\mathcal{T}(n-p) = \mathcal{T}(233)$ où $\Gamma_{1,1}$ est le second élément diagonal (correspondant à β_1) de la matrice $({}^tXX)^{-1}$ de sorte que l'estimateur de la variance de $\hat\beta_1$ $\widehat{\text{var}}(\hat\beta_1) = S^2\,\Gamma_{1,1}$

La valeur observée de T_1 vaut $\frac{0.3927}{0.0361} = 10.893$

et la p-valeur bilatérale $2\,P(T_1 > |10.893|) = 2\,(1 - \Phi_{\mathcal{T}}(|10.893|)) \approx 1.33 \times 10^{-22}$ (cf Tableau B.1) où $\Phi_{\mathcal{T}}$ est la fonction de répartition de la loi $\mathcal{T}(233)$;
puisque la valeur observée de la pente est positive, cohérente avec l'alternative du test, la p-valeur unilatérale vaut : $P_{H_0}(T_1 > 10.893) = 1 - \Phi_{\mathcal{T}}(10.893) \approx 1.33 \times 10^{-22}/2 = 6.65 \times 10^{-23}$
La p-valeur étant inférieure au niveau de risque $\alpha = 5\%$ on rejette l'hypothèse nulle en faveur de l'alternative au risque maximum $\alpha = 5\%$.

Au risque maximum $\alpha = 5\%$ la pente de la droite du montant des dommages des sinistres de type 1 en fonction de celui des sinistres de type 2 est significativement positive.
On estime avec une confiance de 95% que la pente de la droite se situe entre 0.3217 et 0.4638 environ (cf Tableau B.2).

Les résultats numériques ont été obtenus grâce aux commandes R ci-dessous :

```
summary(mod1)$coef # tests de nullité des cœfficients
confint(mod1) # IC à 95% des cœfficients
```

Tableau B.1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.8211643	0.3881347	25.30349	8.667010e-69
DOM2	0.3927412	0.0360541	10.89309	1.330026e-22

Tableau B.2

	2.5 %	97.5 %
(Intercept)	9.0564622	10.585866
DOM2	0.3217074	0.463775

(e) Quelle est la qualité de l'ajustement réalisé ?

La qualité globale de l'ajustement réalisé est mesurée par le cœfficient de détermination

$$R^2 = \frac{SSR}{SST}$$

où SSR est la somme des carrés expliquée par le modèle (1) et SST la somme des carrés totale, numérateur de la variance empirique de la variable à expliquer DOM1

$$SSR = \|\hat{Y} - \overline{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \quad \text{et} \quad SST = \|Y - \overline{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Pour le modèle de régression linéaire simple, $R^2 = r(x, y)^2 = r(\text{DOM2}, \text{DOM1})^2$

Le cœfficient de détermination observé vaut 0.3374 : environ 33.74% de la variabilité du montant des dommages des sinistres de type 1 est expliquée par sa régression sur celui des sinistres de type 2. Il est significatif au risque 5% puisque la pente est significativement non nulle au risque maximum 5% (p-valeur bilatérale de $1.33 \times 10^{-22} < \alpha = 5\%$, cf Tableau B.1).

< Une part significative (environ 33.74%) de la variabilité du montant des dommages des sinistres de type 1 est expliquée par sa régression sur celui des sinistres de type 2.

Les résultats numériques ont été obtenus grâce aux commandes R ci-dessous :

```
summary(mod1)$r.squared   ;   cor(DOM1,DOM2)^2

[1]  0.3374273
[1]  0.3374273
```

(f) Étudier la validité du modèle (1).

Les graphiques obtenus (cf Graphique B.4) ne permettent pas d'invalider la linéarité (on n'observe pas de tendance mais une légère courbure sur le Graphique B.4a) ni l'homoscédasticité des erreurs (légère hétérogénéité pour les valeurs prévues faibles sur le Graphique B.4a et le Graphique B.4c) ; la normalité des erreurs est vérifiée visuellement par le bon alignement des points le long de la droite de Henry (cf Graphique B.4b) et la compatibilité de l'histogramme des résidus standardisés avec une loi $\mathcal{N}(0; 1)$ (cf Graphique B.5).

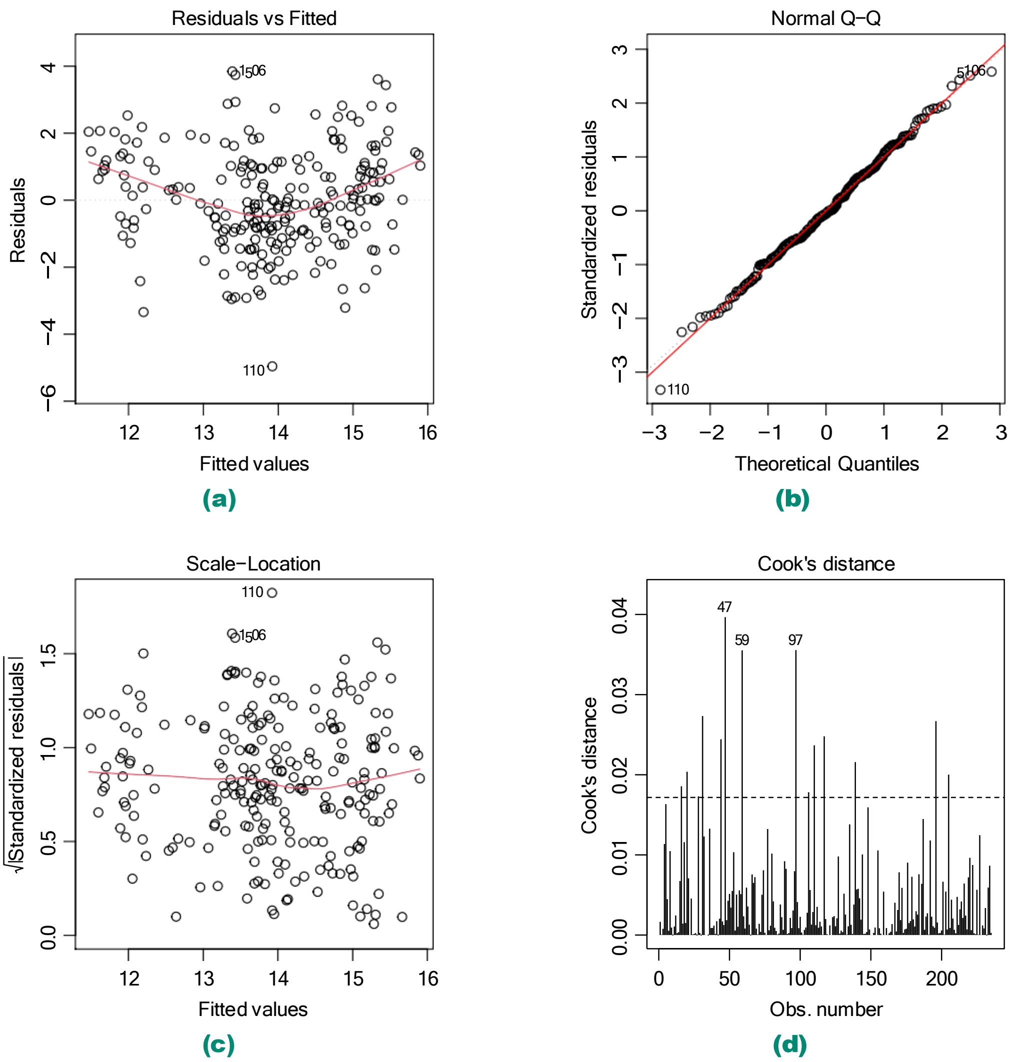
On observe 7 résidus extrêmes (résidus standardisés supérieurs à 2 en valeur absolue) soit 2.98% et 14 points influents soit 5.96%, correspondants aux distances de Cook élevées (supérieures à $4/233 \approx 0.0172$, cf Graphique B.4d), soit moins de 5% de valeurs mal prévues et un peu plus de 5% de valeurs influentes.

Aucun de ces éléments ne permet de remettre en cause la validité du modèle (1).

Les graphiques ont été obtenus grâce aux commandes R qui suivent.

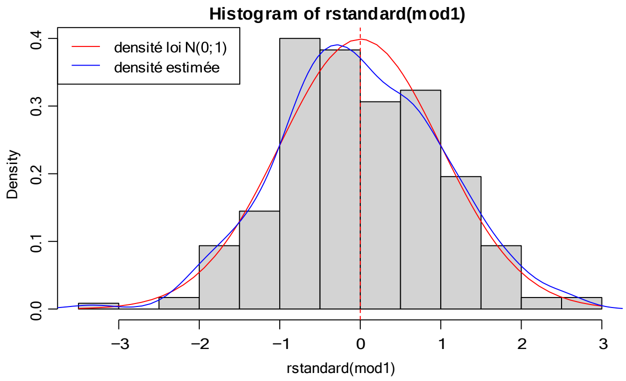
```
plot(mod1,1:2)
abline(0,1, col='red') # droite d'équation y=x
plot(mod1,3:4)
abline(h=4/mod1$df, lty=2) # limite point influent
which(abs(rstandard(mod1))>2) ; length(which(abs(rstandard(mod1))>2))

   5   16   47   59   97  106  110
   5   16   47   59   97  106  110
[1]  7

which(cooks.distance(mod1)>4/mod1$df);length(which(cooks.distance(mod1)>4/mod1$df))

  16   20   28   31   44   47   59   97  106  110  117  139  196  205
  16   20   28   31   44   47   59   97  106  110  117  139  196  205
[1]  14
```

Graphique B.4



```r
hist(rstandard(mod1), freq=F) # histogramme résidus standardisés
curve(dnorm(x,0,1), add=T, col='red') # densité loi N(0;1)
abline(v=0, lty=2, col='red')
lines(density(rstandard(mod1)), col='blue') # densité estimée résidus standardisés
legend('topleft', legend=c('densité loi N(0;1)', 'densité estimée'),
       col=c('red','blue'), lty=1)
```

Graphique B.5



La part de variabilité du montant des dommages de type 1 expliquée par le modèle (1) étant faible (33.74%), il est légitime de chercher à augmenter la qualité de l'ajustement en introduisant de nouvelle(s) variable(s) dans le modèle.

Question C

On se demande s'il est pertinent de pendre en compte la CSP du ménage dans le modèle (1).

1. Représenter le nuage de points du montant des dommages des sinistres de **type 1** en fonction de celui des dommages des sinistres de **type 2**, en utilisant une couleur spécifique pour chaque niveau de la variable CSP. Préciser la signification de chaque couleur dans une légende au graphique.

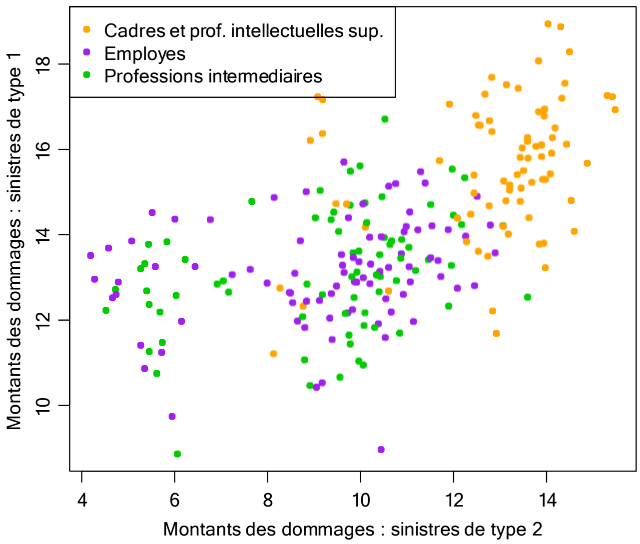
```
CSP <- as.factor(CSP) # définir le facteur CSP
levels(CSP) # niveaux du facteur CSP

[1] "Cadres et prof. intellectuelles sup."
[2] "Employes"
[3] "Professions intermediaires"

couleurCSP  <-  c('orange','purple','green3')
coulCSP <- ifelse(CSP==levels(CSP)[1], couleurCSP[1],
                  ifelse(CSP==levels(CSP)[2], couleurCSP[2], couleurCSP[3]))
```

La variable CSP est une variable qualitative à 3 modalités, transformée en facteur (qualitatif) à 3 niveaux : *Cadres et prof. intellectuelles sup.*, *Employes*, *Professions intermediaires*.

Le nuage de points (cf Graphique C.1) est réalisé avec les commandes R ci-dessous :

```
# Nuage de points de DOM1 en fonction de DOM2 selon CSP
plot(DOM2,DOM1, xlab="Montants des dommages : sinistres de type 2",
    ylab="Montants des dommages : sinistres de type 1", pch=20, col=coulCSP)
legend('topleft', legend=c(levels(CSP)[1],levels(CSP)[2],levels(CSP)[3]),
       col=couleurCSP, pch=20)
```

Graphique C.1



2. Proposer plusieurs modèles distincts permettant de tenir compte de la CSP du ménage dans le modèle de régression du montant des dommages des sinistres de **type 1** sur celui des dommages des sinistres de **type 2**. Expliciter ces modèles, les écrire sous forme matricielle, préciser la(les) contrainte(s) d'identifiabilité éventuelle(s) en expliquant son(leur) utilité.

Trois modèles distincts peuvent être envisagés pour tenir compte de la CSP du ménage dans l'explication du montant des dommages des sinistres de type 1 par celui des dommages des sinistres de type 2.

(a) Le modèle additif ou sans interaction (modèle (2.a)) consiste à ajuster trois droites de même pente à chacun des trois nuages de points selon la CSP.

Il s'écrit $Y_i = \mu_0 + a\,x_i + \mu_1\,\mathrm{I}_{i1} + \mu_2\,\mathrm{I}_{i2} + \mu_3\,\mathrm{I}_{i3} + \varepsilon_i$ pour tout $i = 1, \ldots, n = 235$

et de manière matricielle : $Y = X\beta + \varepsilon$

- $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ étant les vecteurs aléatoires $(n, 1)$

- $X = (1_n \;\; x \;\; \mathrm{I}_1 \;\; \mathrm{I}_2 \;\; \mathrm{I}_3)$ la matrice (n, k) du modèle, où x désigne le vecteur des montants des dommages des sinistres de type 2 et I_1, I_2 et I_3 les indicatrices respectives de chaque niveau de CSP :

  $\mathrm{I}_{i1} = 1$ si le ménage i a pour CSP Cadres et prof. intellectuelles sup. et $\mathrm{I}_{i1} = 0$ sinon,

  $\mathrm{I}_{i2} = 1$ si le ménage i a pour CSP Employes et $\mathrm{I}_{i2} = 0$ sinon,

  $\mathrm{I}_{i3} = 1$ si le ménage i a pour CSP Professions intermediaires et $\mathrm{I}_{i3} = 0$ sinon.

  Ces trois indicatrices et le vecteur 1_n (vecteur dont toutes les composantes sont égales à 1) étant liés linéairement puisque $\mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3 = 1_n$ la matrice X du modèle n'est pas injective, elle est de rang $p = 4 < k = 5$;

  le modèle n'étant pas identifiable, il nécessite une contrainte d'identifiabilité, celle imposée par défaut par R étant : $\mu_1 = 0$

Le modèle identifiable s'écrit donc :

(2.a) $Y_i = \mu_0 + a\,x_i + \mu_2\,\mathrm{I}_{i2} + \mu_3\,\mathrm{I}_{i3} + \varepsilon_i$ pour tout $i = 1, \ldots, n = 235$

dont la matrice $X = (1_n \;\; x \;\; \mathrm{I}_2 \;\; \mathrm{I}_3)$ est de plein rang $p = 4$

ou (2.a) $\quad Y = X\beta + \varepsilon$

où le vecteur $(p = 4, 1)$ des cœfficients inconnus du modèle $\beta = \begin{pmatrix} \mu_0 \\ a \\ \mu_2 \\ \mu_3 \end{pmatrix}$ doit être estimé.

On peut donc décomposer le modèle (2.a) en trois parties :

pour un ménage i de CSP *Cadres et prof. intellectuelles sup.* : $Y_i = \mu_0 + a\,x_i + \varepsilon_i$

pour un ménage i de CSP *Employes* : $Y_i = \mu_0 + a\,x_i + \mu_2 + \varepsilon_i = \mu_0 + \mu_2 + a\,x_i + \varepsilon_i$

pour un ménage i de CSP *Professions intermediaires* : $Y_i = \mu_0 + a\,x_i + \mu_3 + \varepsilon_i = \mu_0 + \mu_3 + a\,x_i + \varepsilon_i$

on ajuste trois droites de régression de même pente a mais de termes constants différents.

Pour $\mu_2 = \mu_3 = 0$ les trois droites coïncident : on retrouve le modèle (1) ; le modèle (1) est donc un sous-modèle du modèle (2.a).

(b) Le modèle multiplicatif ou avec interaction (modèle (2.b)) consiste à ajuster trois droites distinctes à chacun des trois nuages de points selon la CSP.

Il s'écrit $Y_i = \delta_0 + a_0\,x_i + \delta_1\,\mathrm{I}_{i1} + \delta_2\,\mathrm{I}_{i2} + \delta_3\,\mathrm{I}_{i3} + a_1\,x_i\mathrm{I}_{i1} + a_2\,x_i\mathrm{I}_{i2} + a_3\,x_i\mathrm{I}_{i3} + \varepsilon_i$ pour tout $i = 1, \ldots, n = 235$

et de manière matricielle : $Y = X\beta + \varepsilon$

- $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ étant les vecteurs aléatoires $(n, 1)$

- $X = (1_n \;\; x \;\; \mathrm{I}_1 \;\; \mathrm{I}_2 \;\; \mathrm{I}_3 \;\; \mathrm{I}_1 x \;\; \mathrm{I}_2 x \;\; \mathrm{I}_3 x)$ la matrice (n, k) du modèle, où les vecteurs $\mathrm{I}_1 x$, $\mathrm{I}_2 x$ et $\mathrm{I}_3 x$ sont définis par :

  $\mathrm{I}_{i1} x_i = x_i$ si le ménage i a pour CSP Cadres et prof. intellectuelles sup. et $\mathrm{I}_{i1} x_i = 0$ sinon,

  $\mathrm{I}_{i2} x_i = x_i$ si le ménage i a pour CSP Employes et $\mathrm{I}_{i2} x_i = 0$ sinon,

  $\mathrm{I}_{i3} x_i = x_i$ si le ménage i a pour CSP Professions intermediaires et $\mathrm{I}_{i3} x_i = 0$ sinon.

  Ces trois vecteurs et le vecteur x étant liés linéairement puisque $\mathrm{I}_1 x + \mathrm{I}_2 x + \mathrm{I}_3 x = x$ la matrice X du modèle n'est pas injective, elle est de rang $p = 6 < k = 8$; le modèle n'étant pas identifiable, il nécessite deux contraintes d'identifiabilité, celles imposées par défaut par R étant : $\delta_1 = a_1 = 0$

Le modèle identifiable s'écrit donc :

(2.b) $\quad Y_i = \delta_0 + a_0\, x_i + \delta_2\, \mathrm{I}_{i2} + \delta_3\, \mathrm{I}_{i3} + a_2\, x_i \mathrm{I}_{i2} + a_3\, x_i \mathrm{I}_{i3} + \varepsilon_i$ pour tout $i = 1, \dots, n = 235$

dont la matrice $X = (\mathbb{1}_n \;\; x \;\; \mathrm{I}_2 \;\; \mathrm{I}_3 \;\; \mathrm{I}_2\, x \;\; \mathrm{I}_3\, x)$ est de plein rang $p = 6$

ou (2.b) $\quad Y = X\beta + \varepsilon$

où le vecteur $(p = 6, 1)$ des cœfficients inconnus du modèle $\beta = \begin{pmatrix} \delta_0 \\ a_0 \\ \delta_2 \\ \delta_3 \\ a_2 \\ a_3 \end{pmatrix}$ est à estimer.

On peut donc décomposer le modèle (2.b) en trois parties :

$$\begin{cases} \text{pour un ménage } i \text{ de CSP } \textit{Cadres et prof. intellectuelles sup.} : & Y_i = \delta_0 + a_0\, x_i + \varepsilon_i \\ \text{pour un ménage } i \text{ de CSP } \textit{Employes} : & Y_i = \delta_0 + a_0\, x_i + \delta_2 + a_2\, x_i + \varepsilon_i \\ & = \delta_0 + \delta_2 + (a_0 + a_2)\, x_i + \varepsilon_i \\ \text{pour un ménage } i \text{ de CSP } \textit{Professions intermediaires} : & Y_i = \delta_0 + a_0\, x_i + \delta_3 + a_3\, x_i + \varepsilon_i \\ & = \delta_0 + \delta_3 + (a_0 + a_3)\, x_i + \varepsilon_i \end{cases}$$

on ajuste trois droites de régression de pentes et de termes constants différents.

Pour $a_2 = a_3 = 0$ les trois droites ont la même pente : on retrouve le modèle (2.a) ; le modèle (2.a) est donc un sous-modèle du modèle (2.b) ; pour $\delta_2 = \delta_3 = a_2 = a_3 = 0$ les trois droites coïncident : on retrouve le modèle (1) ; le modèle (1) est donc un sous-modèle du modèle (2.b).

(c) Le modèle (modèle (2.c)) consiste à ajuster trois droites de même cœfficient constant à chacun des trois nuages de points selon la CSP.

Il s'écrit $\quad Y_i = \gamma + b_0\, x_i + b_1\, x_i \mathrm{I}_{i1} + b_2\, x_i \mathrm{I}_{i2} + b_3\, x_i \mathrm{I}_{i3} + \varepsilon_i \quad$ pour tout $i = 1, \dots, n = 235$

et de manière matricielle : $Y = X\beta + \varepsilon$

- $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ étant les vecteurs aléatoires $(n, 1)$
- $X = (\mathbb{1}_n \;\; x \;\; \mathrm{I}_1\, x \;\; \mathrm{I}_2\, x \;\; \mathrm{I}_3\, x)$ la matrice (n, k) du modèle, non injective puisqu'elle est de rang $p = 4 < k = 5$; le modèle n'étant pas identifiable, il nécessite une contrainte d'identifiabilité, celle imposée par défaut par R étant : $b_1 = 0$

Le modèle identifiable s'écrit donc :

(2.c) $\quad Y_i = \gamma + b_0\, x_i + b_2\, x_i \mathrm{I}_{i2} + b_3\, x_i \mathrm{I}_{i3} + \varepsilon_i \quad$ pour tout $i = 1, \dots, n = 235$

dont la matrice $X = (\mathbb{1}_n \;\; x \;\; \mathrm{I}_2\, x \;\; \mathrm{I}_3\, x)$ est de plein rang $p = 4$

ou (2.c) $\quad Y = X\beta + \varepsilon$

où le vecteur $(p = 4, 1)$ des cœfficients inconnus du modèle $\beta = \begin{pmatrix} \gamma \\ b_0 \\ b_2 \\ b_3 \end{pmatrix}$ doit être estimé.

On peut donc décomposer le modèle (2.c) en trois parties :

$$\begin{cases} \text{pour un ménage } i \text{ de CSP } \textit{Cadres et prof. intellectuelles sup.} : & Y_i = \gamma + b_0\, x_i + \varepsilon_i \\ \text{pour un ménage } i \text{ de CSP } \textit{Employes} : & Y_i = \gamma + b_0\, x_i + b_2\, x_i + \varepsilon_i \\ & = \gamma + (b_0 + b_2)\, x_i + \varepsilon_i \\ \text{pour un ménage } i \text{ de CSP } \textit{Professions intermediaires} : & Y_i = \gamma + b_0\, x_i + b_3\, x_i + \varepsilon_i \\ & = \gamma + (b_0 + b_3)\, x_i + \varepsilon_i \end{cases}$$

on ajuste trois droites de régression de mêmes termes constants.

Pour $b_2 = b_3 = 0$ les trois droites coïncident : on retrouve le modèle (1) ; le modèle (1) est donc un sous-modèle du modèle (2.c) ; le modèle (2.c) correspond au modèle (2.b) avec les valeurs particulières $\delta_2 = \delta_3 = 0$ (les trois droites ont le même cœfficient constant) ; le modèle (2.c) est donc un sous-modèle du modèle (2.b).

Pour chacun de ces trois modèles, les hypothèses probabilistes sont celles d'un modèle linéaire gaussien :
$\varepsilon \sim \mathcal{N}_n\left(\mathbf{0}_n; \sigma^2 I_n\right) \quad$ ou $\quad \varepsilon_i$ i.i.d. de loi $\mathcal{N}\left(0; \sigma^2\right)$ pour tout $i = 1, \dots, n$.

La variance des erreurs σ^2 inconnue doit être estimée pour chaque modèle.

3. Ajuster les modèles proposés.

(a) Estimation du modèle (2.a)

Les estimations des cœfficients du modèle (2.a) : $\hat{\mu}_0 \approx 12.91$, $\hat{a} \approx 0.2054$, $\hat{\mu}_2 \approx -1.696$ et $\hat{\mu}_3 \approx -1.681$ (cf Tableau C.1) de sorte que les valeurs prévues pour le montant des dommages pour les sinistres de type 1 par le modèle (2.a) sont :

- pour un ménage i de CSP *Cadres et prof. intellectuelles sup.* : $\hat{Y}_i = \hat{\mu}_0 + \hat{a}\, x_i \approx 12.91 + 0.2054\, x_i$
- pour un ménage i de CSP *Employes* : $\hat{Y}_i = \hat{\mu}_0 + \hat{\mu}_2 + \hat{a}\, x_i \approx 12.91 - 1.696 + 0.2054\, x_i \approx 11.21 + 0.2054\, x_i$
- pour un ménage i de CSP *Professions intermediaires* : $\hat{Y}_i = \hat{\mu}_0 + \hat{\mu}_3 + \hat{a}\, x_i \approx 12.91 - 1.681 + 0.2054\, x_i \approx 11.23 + 0.2054\, x_i$

Les droites ajustées observées sont représentées sur le nuage de points (cf Graphique C.2a).

L'estimation sans biais de la variance des erreurs du modèle (2.a) $S^2 \approx 1.867$

(b) Estimation du modèle (2.b)

Les estimations des cœfficients du modèle (2.b) : $\hat{\delta}_0 \approx 11.15$, $\hat{a}_0 \approx 0.3422$, $\hat{\delta}_2 \approx 0.7812$, $\hat{\delta}_3 \approx 0.001683$, $\hat{a}_2 \approx -0.2142$ et $\hat{a}_3 \approx -0.1281$ (cf Tableau C.2) de sorte que les valeurs prévues pour le montant des dommages pour les sinistres de type 1 par le modèle (2.b) sont :

- pour un ménage i de CSP *Cadres et prof. intellectuelles sup.* :
$$\hat{Y}_i = \hat{\delta}_0 + \hat{a}_0 x_i \approx 11.15 + 0.3422\, x_i$$
- pour un ménage i de CSP *Employes* :
$$\hat{Y}_i = \hat{\delta}_0 + \hat{\delta}_2 + (\hat{a}_0 + \hat{a}_2) x_i \approx 11.15 + 0.7812 + (0.3422 - 0.2142)\, x_i \approx 11.93 + 0.128\, x_i$$
- pour un ménage i de CSP *Professions intermediaires* :
$$\hat{Y}_i = \hat{\delta}_0 + \hat{\delta}_3 + (\hat{a}_0 + \hat{a}_3) x_i \approx 11.15 + 0.001683 + (0.3422 - 0.1281)\, x_i \approx 11.15 + 0.2141\, x_i$$

Les droites ajustées observées sont représentées sur le nuage de points (cf Graphique C.2b).

L'estimation sans biais de la variance des erreurs du modèle (2.b) $S^2 \approx 1.854$

(c) Estimation du modèle (2.c)

Les estimations des cœfficients du modèle (2.c) : $\hat{\gamma} \approx 11.52$, $\hat{b}_0 \approx 0.3138$, $\hat{b}_2 \approx -0.144$ et $\hat{b}_3 \approx -0.1377$ (cf Tableau C.3) de sorte que les valeurs prévues pour le montant des dommages pour les sinistres de type 1 par le modèle (2.c) sont :

- pour un ménage i de CSP *Cadres et prof. intellectuelles sup.* : $\hat{Y}_i = \hat{\gamma} + \hat{b}_0 x_i \approx 11.52 + 0.3138\, x_i$
- pour un ménage i de CSP *Employes* : $\hat{Y}_i = \hat{\gamma} + (\hat{b}_0 + \hat{b}_2) x_i \approx 11.52 + (0.3138 - 0.144)\, x_i \approx 11.52 + 0.1698\, x_i$
- pour un ménage i de CSP *Professions intermediaires* : $\hat{Y}_i = \hat{\gamma} + (\hat{b}_0 + \hat{b}_3) x_i \approx 11.52 + (0.3138 - 0.1377)\, x_i \approx 11.52 + 0.1762\, x_i$

Les droites ajustées observées sont représentées sur le nuage de points (cf Graphique C.2c).

L'estimation sans biais de la variance des erreurs du modèle (2.c) $S^2 \approx 1.845$

Les ajustements des modèles ont été réalisés grâce aux commandes R qui suivent.

```
mod2a <- lm(DOM1 ~ DOM2+CSP) # modèle (2.a) additif
mod2b <- lm(DOM1 ~ DOM2*CSP) # modèle (2.b) multiplicatif
mod2c <- lm(DOM1 ~ DOM2*CSP-CSP) # modèle (2.c)
```

```
mod2a$coef
```

Tableau C.1

(Intercept)	DOM2	CSPEmployes	CSPProfessions intermediaires
12.90924	0.2053894	-1.696005	-1.681046

```
sigma(mod2a)^2

[1]  1.86734
```

```
mod2b$coef
```

Tableau C.2

(Intercept)	DOM2	CSPEmployes	CSPProfessions intermediaires	DOM2 :CSPEmployes	DOM2 :CSPProfessions intermediaires
11.14602	0.3421873	0.7811901	0.0016825	-0.214237	-0.1280642

```
sigma(mod2b)^2

[1]  1.853991
```

```
mod2c$coef
```

Tableau C.3

(Intercept)	DOM2	DOM2 :CSPEmployes	DOM2 :CSPProfessions intermediaires
11.5178	0.3138358	-0.1439945	-0.137667

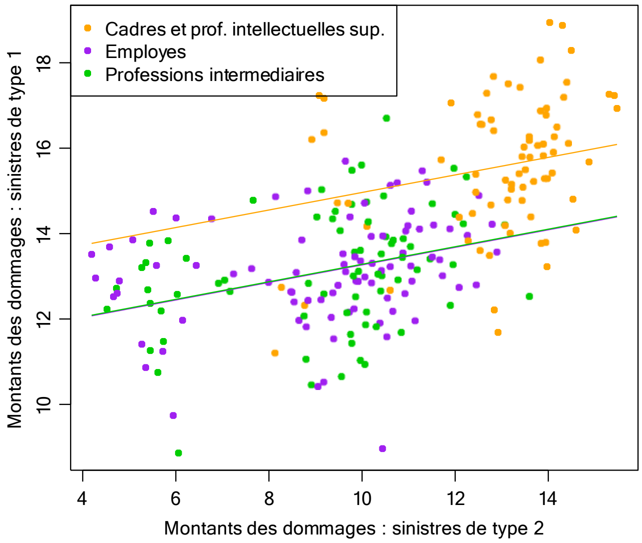
```
sigma(mod2c)^2

[1]  1.844591
```

Les nuages de points (cf Graphique C.2a, Graphique C.2b *et* Graphique C.2c*) sont réalisés avec les commandes* R *ci-dessous :*

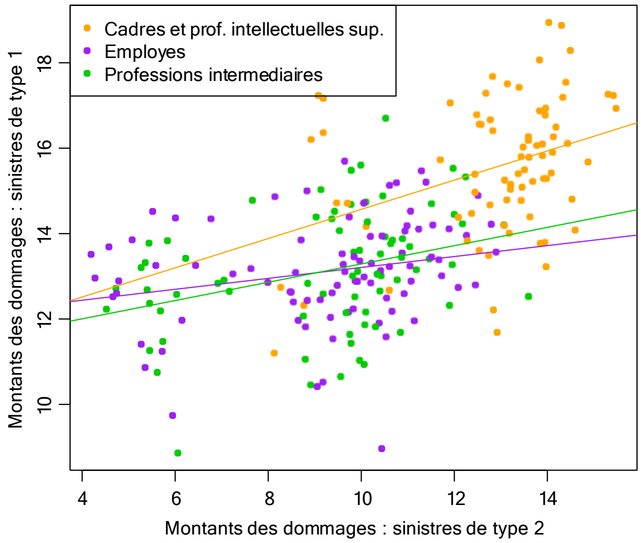
```
# Nuage de points de DOM1 en fonction de DOM2 selon CSP
plot(DOM2,DOM1, xlab="Montants des dommages : sinistres de type 2",
     ylab="Montants des dommages : sinistres de type 1", pch=20, col=coulCSP)
legend('topleft', legend=c(levels(CSP)[1],levels(CSP)[2],levels(CSP)[3]),
       col=couleurCSP, pch=20)
# droites estimées modèle (2a)
curve(mod2a$coef[1]+mod2a$coef[2]*x, col=couleurCSP[1], add=T) # CSP niveau 1
curve(sum(mod2a$coef[c(1,3)])+mod2a$coef[2]*x, col=couleurCSP[2], add=T) # CSP niveau 2
curve(sum(mod2a$coef[c(1,4)])+mod2a$coef[2]*x, col=couleurCSP[3], add=T) # CSP niveau 3
# Nuage de points de DOM1 en fonction de DOM2 selon CSP
plot(DOM2,DOM1, xlab="Montants des dommages : sinistres de type 2",
     ylab="Montants des dommages : sinistres de type 1", pch=20, col=coulCSP)
legend('topleft', legend=c(levels(CSP)[1],levels(CSP)[2],levels(CSP)[3]),
       col=couleurCSP, pch=20)
# droites estimées modèle (2b)
for(i in 1:nlevels(CSP)) {
```

```
        abline(lm(DOM1[CSP==levels(CSP)[i]] ~
                 DOM2[CSP==levels(CSP)[i]])$coef, col=couleurCSP[i], lty=1)
}
# Nuage de points de DOM1 en fonction de DOM2 selon CSP
plot(DOM2,DOM1, xlab="Montants des dommages : sinistres de type 2",
     ylab="Montants des dommages : sinistres de type 1", pch=20, col=coulCSP)
legend('topleft', legend=c(levels(CSP)[1],levels(CSP)[2],levels(CSP)[3]),
        col=couleurCSP, pch=20)
# droites estimées modèle (2c)
curve(mod2c$coef[1]+mod2c$coef[2]*x, col=couleurCSP[1], add=T) # CSP niveau 1
curve(mod2c$coef[1]+sum(mod2c$coef[c(2,3)])*x, col=couleurCSP[2], add=T) # CSP niveau  2
curve(mod2c$coef[1]+sum(mod2c$coef[c(2,4)])*x, col=couleurCSP[3], add=T) # CSP niveau 3
```

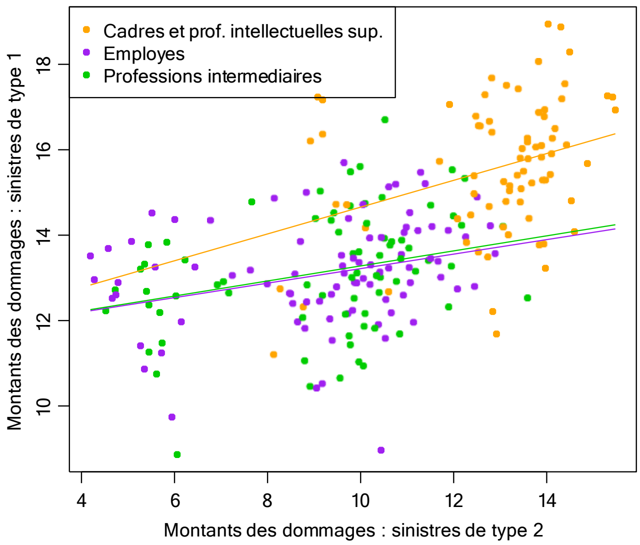
Graphique C.2



(a)



(b)



(c)

4. Comparer les modèles proposés en précisant pour chaque test, les hypothèses testées, la statistique de test utilisée et en justifiant sa loi sous l'hypothèse nulle.

- Test du modèle (2.a) contre le modèle (2.b)

  On confronte l'hypothèse nulle H_0 : modèle (2.a) ou $\alpha_2 = \alpha_3 = 0$
  à l'alternative H_1 : modèle (2.b) ou $(\alpha_2, \alpha_3) \neq (0, 0)$ au niveau de risque $\alpha = 5\%$
  ce qui revient à tester l'égalité des pentes des droites ou la nullité des effets d'interaction dans le modèle (2.b).

  Puisque le modèle (2.a) est un sous-modèle du modèle (2.b) car les vecteurs colonnes de la matrice du modèle (2.a) engendrent un sous-espace vectoriel V_0 de dimension $p_0 = 4$ inclus dans le sous-espace vectoriel V_1 de dimension $p_1 = 6$ engendré par les vecteurs colonnes de la matrice du modèle (2.b) on déduit, en appliquant le théorème de Cochran, que
  sous H_0, la statistique de test de Fisher

  $$F = \frac{\|P_{V_1}Y - P_{V_0}Y\|^2/(p_1 - p_0)}{\|Y - P_{V_1}Y\|^2/(n - p_1)} = \frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/(n - p_1)}$$

  suit la loi de Fisher $\mathcal{F}(p_1 - p_0; n - p_1) = \mathcal{F}(2; 229)$

  SSE_0 et SSE_1 étant les sommes des carrés résiduelles respectives des modèles sous H_0 modèle (2.a) et sous H_1 modèle (2.b).
  La valeur observée de F est égale à $\dfrac{(431.36 - 424.56)/(6 - 4)}{424.56/(235 - 6)} = \dfrac{6.79/2}{424.56/229} \approx 1.832$

  et la p-valeur correspondante $P_{H_0}(F > 1.832) = 1 - \Phi_F(1.832) \approx 0.1625$ (où Φ_F est la fonction de répartition de la loi $\mathcal{F}(2; 229)$) étant supérieure au seuil de risque maximum $\alpha = 5\%$, on ne rejette pas l'hypothèse nulle de nullité des effets d'interaction.

  < Les effets d'interaction ne sont pas significativement non nuls au seuil $\alpha = 5\%$, c'est-à-dire que le modèle (2.b) n'apporte pas d'information supplémentaire significative au modèle (2.a) dans l'explication de la variabilité des montants des dommages pour les sinistres de type 1. Au seuil $\alpha = 5\%$, les trois pentes du modèle (2.b) ne diffèrent pas significativement.

  Les résultats numériques ont été obtenus grâce aux commandes R suivantes :

```
anova(mod2b)

Analysis of Variance Table

Response:  DOM1
             Df Sum Sq Mean Sq   F value    Pr(>F)
DOM2          1 264.27 264.273  142.5430 < 2.2e-16 ***
CSP           2  87.57  43.786   23.6172  4.73e-10 ***
DOM2:CSP      2   6.79   3.396    1.8316    0.1625
Residuals   229 424.56   1.854
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(mod2a, mod2b)

Analysis of Variance Table

Model 1: DOM1 ~ DOM2 + CSP
Model 2: DOM1 ~ DOM2 * CSP
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    231 431.36
2    229 424.56  2    6.7917 1.8316 0.1625
```

- Test du modèle (2.c) contre le modèle (2.b)

  On confronte l'hypothèse nulle H_0 : modèle (2.c) ou $\delta_2 = \delta_3 = 0$
  à l'alternative H_1 : modèle (2.b) ou $(\delta_2, \delta_3) \neq (0, 0)$ au niveau de risque $\alpha = 5\%$
  ce qui revient à tester l'égalité des termes constants dans le modèle (2.b).
  Puisque le modèle (2.c) est un sous-modèle du modèle (2.b)(les vecteurs colonnes de la matrice du modèle (2.c) engendrent un sous-espace vectoriel V_0 de dimension $p_0 = 4$ inclus dans le sous-espace vectoriel V_1 de dimension $p_1 = 6$ engendré par les vecteurs colonnes de la matrice du modèle (2.b))
  sous H_0, la statistique de test de Fisher F suit la loi de Fisher $\mathcal{F}(2; 229)$.
  La valeur observée de F est égale à $\dfrac{(426.1 - 424.56)/(6-4)}{424.56/(235-6)} = \dfrac{1.54/2}{424.56/229} \simeq 0.414$
  et la p-valeur correspondante $P_{H_0}(F > 0.414) \simeq 0.6612$ étant supérieure au seuil de risque maximum $\alpha = 5\%$, on ne rejette pas l'hypothèse nulle d'égalité des termes constants.

  Par rapport au modèle (2.c), le modèle (2.b) n'apporte pas d'information supplémentaire significative à l'explication de la variabilité des montants des dommages des sinistres de type 1. Au seuil $\alpha = 5\%$, les termes constants du modèle (2.b) ne diffèrent pas significativement.

  Les résultats numériques ont été obtenus grâce à la commande R ci-dessous :

```
anova(mod2c, mod2b)

Analysis of Variance Table

Model 1: DOM1 ~ DOM2 * CSP - CSP
Model 2: DOM1 ~ DOM2 * CSP
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    231 426.10
2    229 424.56  2    1.5366 0.4144 0.6612
```

- Comparaison des modèles (2.a) et (2.c)

  Les modèles (2.a) et (2.c) ne sont pas emboîtés : ils ont la même dimension $p = 4$; on ne peut pas les comparer par un test de Fisher.
  Un critère de comparaison possible est le cœfficient de détermination R^2 (les dimensions des modèles étant identiques, les critères AIC et BIC aboutiront à la même conclusion que le R^2) : $R^2_{(2.c)} \simeq 0.4559$ est très légèrement supérieur à $R^2_{(2.a)} \simeq 0.4492$.

  Les résultats numériques ont été obtenus grâce aux commandes R suivantes :

```
summary(mod2a)$r.squared;  summary(mod2c)$r.squared

[1] 0.4492402
[1] 0.4559499
```

Le modèle retenu est donc le modèle (2.c) qui ajuste trois droites de même terme constant et explique 45.59% de la variabilité des montants des dommages des sinistres de type 1.

5. Pour le modèle retenu (modèle (2)) :
   - donner une interprétation des cœfficients du modèle ;
   - représenter graphiquement sur le nuage de points, les valeurs prévues par le modèle ; commenter.

Le modèle (2) retenu est le modèle (2.c) qui ajuste les trois droites d'équation

pour un ménage de CSP *Cadres et prof. intellectuelles sup.* : $y = \hat{\gamma} + \hat{b}_0 x_i \simeq 11.52 + 0.3138\, x_i$

pour un ménage de CSP *Employes* : $y = \hat{\gamma} + (\hat{b}_0 + \hat{b}_2)\, x_i \simeq 11.52 + 0.1698\, x_i$

pour un ménage de CSP *Professions intermediaires* : $y = \hat{\gamma} + (\hat{b}_0 + \hat{b}_3)\, x_i \simeq 11.52 + 0.1762\, x_i$

Lorsque le montant des dommages des sinistres de type 2 augmente d'une unité,

- pour un ménage de CSP *Cadres et prof. intellectuelles sup.* le montant estimé des dommages des sinistres de type 1 augmente de 0.3138 unité,
- pour un ménage de CSP *Employes* le montant estimé des dommages des sinistres de type 1 augmente de 0.1698 unité et
- pour un ménage de CSP *Professions intermediaires* le montant estimé des dommages des sinistres de type 1 augmente de 0.1762 unité.

Les pentes estimées des droites pour les CSP *Employes* et *Professions intermediaires* sont très proches ; on pourrait regrouper ces deux catégories.

L'estimation sans biais de la variance des erreurs du modèle (2.c) S^2 Ä 1.845
Les droites ajustées observées sont représentées sur le nuage de points (cf Graphique C.2c).

* Dans le modèle (2.c) les pentes des droites pour les CSP *Employes* et *Professions intermediaires* sont significativement différentes au risque $\alpha = 5\%$, de la pente de la droite pour la CSP *Cadres et prof. intellectuelles sup.* puisque les deux tests de Student (séparés) de nullité des cœfficients b_2 (H_0 : $b_2 = 0$) et b_3 (H_0 : $b_3 = 0$) ont pour p-valeurs bilatérales respectives 8.58×10^{-10} et 1.45×10^{-8} (calculées avec la loi $\mathcal{Y}$ $(235 - 231) = \mathcal{Y}$ (231), cf Tableau C.4) toutes deux très largement inférieures au seuil $\alpha = 5\%$.

Cette paramétrisation du modèle (2.c) ne permet pas de comparer de façon simple les pentes des droites pour les CSP *Employes* et *Professions intermediaires* c'est-à-dire de tester l'égalité des pentes, soit l'hypothèse nulle $H_0 : b_2 = b_3$; les estimations des pentes sont proches numériquement puisque $\hat{b}_2$ Ä -0.144 et $\hat{b}_3$ Ä -0.1377 (cf Tableau C.4).
Pour faire le test, il faudrait modifier le niveau de référence de la CSP *Cadres et prof. intellectuelles sup.* par défaut de R en le fixant à *Employes* par exemple.

```
summary(mod2c)$coef
```

Tableau C.4

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.5178026	0.42672807	26.990965	2.234766e-73
DOM2	0.3138358	0.03465056	9.057164	5.781151e-17
DOM2 :CSPEmployes	-0.1439945	0.02250072	-6.399550	8.584796e-10
DOM2 :CSPProfessions intermediaires	-0.1376670	0.02342566	-5.876760	1.451360e-08

6. Entre les modèles (1) et (2), lequel préférer (justifier) ?

On teste l'hypothèse nulle H_0 : modèle (1) ou $b_2 = b_3 = 0$
contre l'alternative H_1 : modèle (2.c) ou $(b_2, b_3) \neq (0, 0)$ au niveau de risque $\alpha = 5\%$
ce qui revient à tester l'égalité des pentes dans le modèle (2.c).
Puisque le modèle (1) est un sous-modèle du modèle (2.c)(les vecteurs colonnes de la matrice du modèle (1) engendrent un sous-espace vectoriel V_0 de dimension $p_0 = 2$ inclus dans le sous-espace vectoriel V_1 de dimension $p_1 = 4$ engendré par les vecteurs colonnes de la matrice du modèle (2.c))
sous H_0, la statistique de test de Fisher F suit la loi de Fisher $7(2; 231)$.
La valeur observée de F est égale à $\dfrac{(518.93 - 426.1)/(4-2)}{426.1/(235-4)} = \dfrac{92.83/2}{426.1/231}$ Ä 25.162
et la p-valeur correspondante $P_{H_0}(F > 25.162)$ Ä 1.3×10^{-10} étant inférieure au seuil de risque maximum $\alpha = 5\%$, on rejette l'hypothèse nulle d'égalité des pentes.

< Par rapport au modèle (1), le modèle (2.c) apporte une information supplémentaire significative à l'explication de la variabilité des montants des dommages des sinistres de type 1.
Les trois pentes du modèle (2.c) diffèrent significativement au risque $\alpha = 5\%$.

Les résultats numériques ont été obtenus grâce à la commande R qui suit.

```
anova(mod1, mod2c)

Analysis of Variance Table

Model 1: DOM1 ~ DOM2
Model 2: DOM1 ~ DOM2 * CSP - CSP
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    233 518.93
2    231 426.10  2    92.827 25.162 1.3e-10 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

<u>Remarque</u>

Dans le cas où le modèle (2.a) aurait été retenu à la question 5,
on teste l'hypothèse nulle H_0 : modèle (1) ou $\mu_2 = \mu_3 = 0$
contre l'alternative H_1 : modèle (2.a) ou $(\mu_2, \mu_3) \neq (0, 0)$ au niveau de risque $\alpha = 5\%$
ce qui revient à tester l'égalité des termes constants dans le modèle (2.a).
La valeur observée de la statistique de test de Fisher F vaut

$$\frac{(518.93 - 431.36)/(4-2)}{431.36/(235-4)} = \frac{87.57/2}{431.36/231} \text{ Ä } 23.448$$

et la p-valeur correspondante $P_{H_0}(F > 23.448)$ Ä 5.354×10^{-10} (où $F \sim 7(2; 231)$ sous H_0)
étant inférieure au seuil de risque maximum $\alpha = 5\%$, on rejette l'hypothèse nulle d'égalité des termes constants.

< Par rapport au modèle (1), le modèle (2.a) apporte une information supplémentaire significative à l'explication de la variabilité des montants des dommages des sinistres de type 1 ; il explique 44.92% de cette variabilité.
Au risque $\alpha = 5\%$, les trois termes constants du modèle (2.a) diffèrent significativement.

Les résultats numériques ont été obtenus grâce à la commande R suivante :

```
anova(mod1, mod2a)

Analysis of Variance Table

Model 1: DOM1 ~ DOM2
Model 2: DOM1 ~ DOM2 + CSP
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    233 518.93
2    231 431.36  2    87.572 23.448 5.354e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dans le modèle (2.a) les termes constants des droites pour les CSP *Employes* et *Professions intermediaires* sont significativement différents au risque $\alpha = 5\%$ du terme constant de la droite pour la CSP *Cadres et prof. intellectuelles sup.* puisque les deux tests de Student (séparés) de nullité des cœfficients μ_2 et μ_3 ont pour p-valeurs bilatérales respectives 1.03×10^{-9} et 3.31×10^{-9} (calculées avec la loi $\mathcal{Y}(235-231) = \mathcal{Y}(231)$, cf Tableau C.5) toutes deux très inférieures au seuil $\alpha = 5\%$.
Cette paramétrisation du modèle (2.a) ne permet pas de tester simplement l'égalité des termes constants des droites pour les CSP *Employes* et *Professions intermediaires* soit l'hypothèse nulle $H_0 : \mu_2 = \mu_3$; les estimations des termes constants sont proches numériquement puisque $\hat{\mu}_2$ Ä -1.696 et $\hat{\mu}_3$ Ä -1.681 (cf Tableau C.5).

```
summary(mod2a)$coef
```

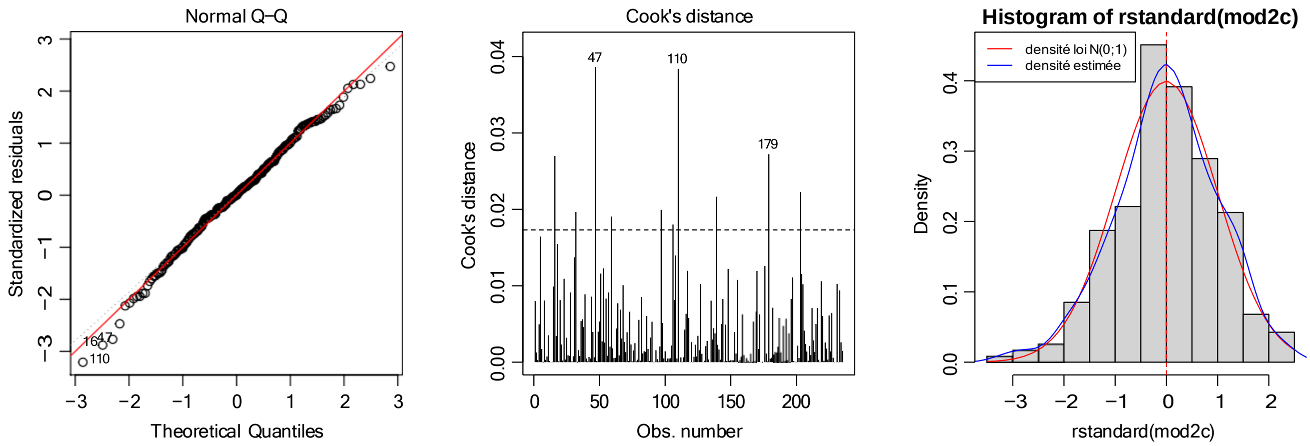
Tableau C.5

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.9092422	0.57417659	22.483052	4.366334e-60
DOM2	0.2053894	0.04287714	4.790184	2.979748e-06
CSPEmployes	-1.6960046	0.26641197	-6.366098	1.033698e-09
CSPProfessions intermediaires	-1.6810457	0.27317670	-6.153694	3.311390e-09

* Vérification des conditions de validité du modèle retenu (2.c)

```
plot(mod2c,2)
abline(0,1, col='red') # droite d'équation y=x
plot(mod2c,4)
abline(h=4/mod2c$df, lty=2) # limite point influent
# which(abs(rstandard(mod2c))>2) ; length(which(abs(rstandard(mod2c))>2))
# which(cooks.distance(mod2c)>4/mod2c$df)
# length(which(cooks.distance(mod2c)>4/mod2c$df))
hist(rstandard(mod2c), freq=F) # histogramme résidus standardisés
curve(dnorm(x,0,1), add=T, col='red') # densité loi N(0;1)
abline(v=0, lty=2, col='red')
lines(density(rstandard(mod2c)), col='blue') # densité estimée résidus standardisés
legend('topleft', legend=c('densité loi N(0;1)', 'densité estimée'),
       col=c('red','blue'), lty=1, cex=0.75)
```

Graphique C.3



(a) **(b)** **(c)**

Graphiquement la normalité des erreurs est vérifiée par le bon alignement des points le long de la droite de Henry (cf Graphique C.3a) et la compatibilité de l'histogramme des résidus standardisés avec une loi U (0; 1) (cf Graphique C.3c).
On observe 11 résidus extrêmes (supérieurs à 2 en valeur absolue) soit 4.68% et 10 points influents soit 4.26%, correspondants aux distances de Cook élevées (supérieures à 4/231 Ä 0.0173, cf Graphique C.3b), soit moins de 5% de valeurs mal prévues et de valeurs influentes.
Les conditions d'application du modèle (2.c) sont validées.

Si le modèle retenu est le modèle (2.a) on observe des graphiques similaires : ses conditions d'application sont validées.

Question D

On s'intéresse aux liaisons entre les variables.

1. Calculer les cœfficients de corrélation linéaire observés des variables quantitatives deux à deux.
   Les représenter avec la fonction `corrplot` du package `corrplot`.
   Quelles sont les deux variables les plus corrélées positivement, négativement ?

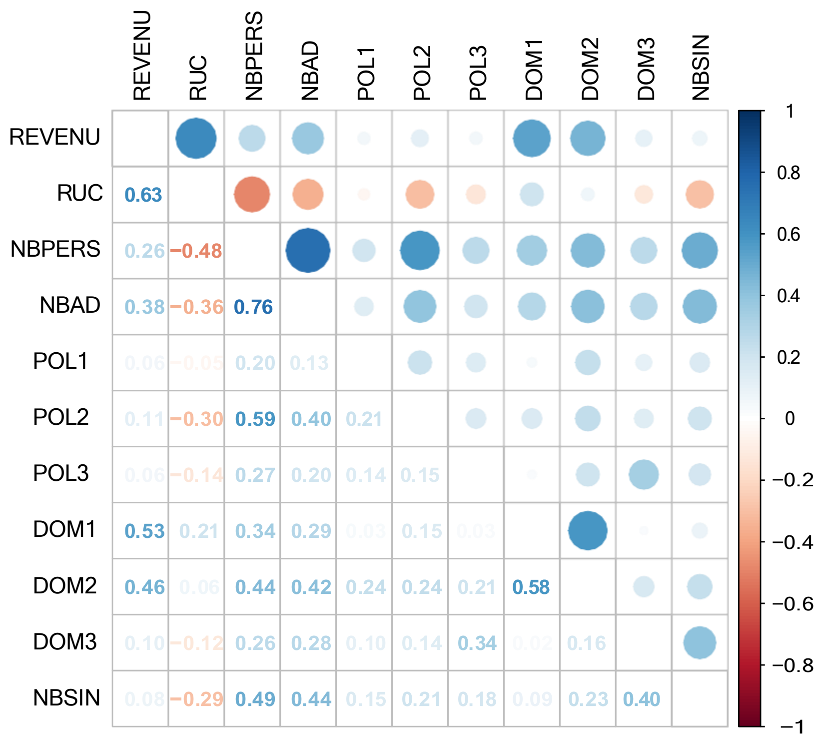
   La matrice des cœfficients de corrélations linéaires observés des variables quantitatives prises deux à deux est donnée dans le tableau suivant et représentée graphiquement (cf Graphique D.1).

```
# matrice des cœfficients de corrélation linéaire observés
C <- cor(données[,c(2,3,7,8,10:16)])
C

             REVENU         RUC      NBPERS        NBAD        POL1      POL2
REVENU 1.00000000  0.63420309  0.2636399  0.3791357  0.05666257  0.1108644
RUC    0.63420309  1.00000000 -0.4834875 -0.3559997 -0.04930355 -0.3033332
NBPERS 0.26363990 -0.48348745  1.0000000  0.7551451  0.19548946  0.5873916
NBAD   0.37913570 -0.35599974  0.7551451  1.0000000  0.13442008  0.3993619
POL1   0.05666257 -0.04930355  0.1954895  0.1344201  1.00000000  0.2148444
POL2   0.11086439 -0.30333321  0.5873916  0.3993619  0.21484441  1.0000000
POL3   0.05891816 -0.13950154  0.2677085  0.1991923  0.14162251  0.1527422
DOM1   0.53216107  0.20744859  0.3437179  0.2874623  0.03468403  0.1544001
DOM2   0.46358546  0.06344385  0.4370912  0.4163304  0.23540661  0.2402998
DOM3   0.10301787 -0.12174009  0.2644209  0.2752976  0.10069525  0.1389262
NBSIN  0.07993291 -0.29320861  0.4923674  0.4378654  0.15066076  0.2098857
              POL3       DOM1       DOM2        DOM3       NBSIN
REVENU  0.05891816 0.53216107 0.46358546  0.10301787  0.07993291
RUC    -0.13950154 0.20744859 0.06344385 -0.12174009 -0.29320861
NBPERS  0.26770849 0.34371791 0.43709124  0.26442090  0.49236739
NBAD    0.19919232 0.28746229 0.41633040  0.27529761  0.43786538
POL1    0.14162251 0.03468403 0.23540661  0.10069525  0.15066076
POL2    0.15274219 0.15440010 0.24029984  0.13892616  0.20988575
POL3    1.00000000 0.02901375 0.20862051  0.33522233  0.18213897
DOM1    0.02901375 1.00000000 0.58088489  0.02035049  0.08977223
DOM2    0.20862051 0.58088489 1.00000000  0.16075980  0.23399893
DOM3    0.33522233 0.02035049 0.16075980  1.00000000  0.40115118
NBSIN   0.18213897 0.08977223 0.23399893  0.40115118  1.00000000
```

```
library(corrplot)
corrplot.mixed(C, tl.col='black', number.cex = .9, tl.pos='lt', tl.cex=1)
```

   Les deux variables les plus corrélées positivement sont le nombre de personnes du ménage NBPERS et le nombre d'adultes du ménage NBAD puisque leur cœfficient de corrélation linéaire observé 0.7551 est maximum ; la variable NBAD ne possédant que deux valeurs, ce cœfficient de corrélation linéaire ne peut pas s'interpréter comme une mesure d'une liaison linéaire entre les deux variables.
   Les deux variables les plus corrélées négativement sont le revenu par unité de consommation du ménage RUC et le nombre de personnes du ménage NBPERS puisque leur cœfficient de corrélation linéaire observé -0.4835 est minimum.

Graphique D.1



2. Soient les commandes R suivantes :

```
COMP <- as.factor(COMP)
I3 <- ifelse(COMP==levels(COMP)[3],1,0)
cor(NBAD,I3)
```

   – Quelles variables sont définies par ces commandes ?
   – Que déduire du résultat du calcul effectué ?

```
COMP <- as.factor(COMP)
nlevels(COMP); levels(COMP)

[1] 3
[1] "Couple avec enfant(s)" "Couple sans enfant"    "Personne seule"

I3 <- ifelse(COMP==levels(COMP)[3],1,0)
table(I3,COMP)

   COMP
I3  Couple avec enfant(s) Couple sans enfant Personne seule
  0                   106                 55              0
  1                     0                  0             74

cor(NBAD,I3)

[1] -1
```

L'objet COMP est un vecteur (n = 235, 1) représentant le facteur correspondant à la composition du ménage COMP à 3 niveaux : *Couple avec enfant(s)*, *Couple sans enfant*, *Personne seule*.
L'objet I3 est un vecteur (n = 235, 1) contenant l'indicatrice du niveau 3 de la composition du ménage *Personne seule* : cette variable vaut 1 pour tout ménage composé de *Personne seule* et 0 sinon.

La variable nombre d'adultes NBAD est parfaitement anti-corrélée avec cette indicatrice, ce qui veut dire que les deux vecteurs centrés correspondants sont linéairement dépendants ; NBAD est parfaitement déterminée par la donnée de I3 puisque, si la composition du ménage est *Personne seule* alors le nombre d'adultes du ménage est de 1, et sinon le nombre d'adultes du ménage est de 2 : NBAD $= 2\ \mathbb{1}_n - \mathrm{I}_3$ si I_3 désigne l'indicatrice du niveau 3 de la variable COMP.

Question E

On considère un modèle complet identifiable de régression multiple du montant des dommages des sinistres de **type 2** sur les autres variables, sans interaction.

1. Expliquer pourquoi il est nécessaire d'éliminer la variable représentant le nombre d'adultes en tant que variable explicative dans un modèle incluant déjà la composition du ménage comme variable explicative.

   La matrice d'un modèle incluant la composition du ménage COMP comme variable explicative a parmi ses vecteurs colonnes, l'indicatrice du niveau 3 de la composition du ménage *Personne seule*.

   La matrice d'un modèle incluant simultanément les variables nombre d'adultes NBAD et composition du ménage COMP comme covariables ne sera pas injective puisque le vecteur indicatrice du niveau 3 de COMP, le vecteur NBAD et le vecteur $\mathbb{1}_n$ sont linéairement dépendants : NBAD $+ \mathrm{I}_3 = 2\ \mathbb{1}_n$;
   le modèle ne sera donc pas identifiable. En d'autres termes, lorsque la variable composition du ménage COMP fait partie des variables explicatives, la variable nombre d'adultes NBAD est redondante.

   La variable composition du ménage COMP à trois modalités *Couple avec enfant(s)*, *Couple sans enfant*, *Personne seule* contenant une information plus détaillée que la variable nombre d'adultes NBAD (1 ou 2) il semble a priori préférable de choisir de l'inclure dans un modèle explicatif.

2. Définir le modèle comportant toutes les variables appropriées (modèle (3)).

   On considère le modèle de régression multiple de la variable à expliquer, montant des dommages des sinistres de type 2 DOM2 sur toutes les covariables (sans interaction), excepté la variable nombre d'adultes NBAD ; il s'écrit, pour $i = 1, \ldots, n$

$$
(3)\qquad
\begin{aligned}
\mathrm{DOM2}_i = \quad & \beta_0 + \beta_1\,\mathrm{REVENU}_i + \beta_2\,\mathrm{RUC}_i + \beta_3\,\mathrm{NBPERS}_i + \beta_4\,\mathrm{POL1}_i + \beta_5\,\mathrm{POL2}_i + \beta_6\,\mathrm{POL3}_i \\
+ \quad & \beta_7\,\mathrm{DOM1}_i + \beta_8\,\mathrm{DOM3}_i + \beta_9\,\mathrm{NBSIN}_i + \vartheta_1\,\mathrm{I}_{i\mathrm{CSP1}} + \vartheta_2\,\mathrm{I}_{i\mathrm{CSP2}} + \vartheta_3\,\mathrm{I}_{i\mathrm{CSP3}} \\
+ \quad & \delta_1\,\mathrm{I}_{i\mathrm{CR1}} + \delta_2\,\mathrm{I}_{i\mathrm{CR2}} + \delta_3\,\mathrm{I}_{i\mathrm{CR3}} + \lambda_1\,\mathrm{I}_{i\mathrm{STOCC1}} + \lambda_2\,\mathrm{I}_{i\mathrm{STOCC2}} \\
+ \quad & \gamma_1\,\mathrm{I}_{i\mathrm{COMP1}} + \gamma_2\,\mathrm{I}_{i\mathrm{COMP2}} + \gamma_3\,\mathrm{I}_{i\mathrm{COMP3}} + \psi_1\,\mathrm{I}_{i\mathrm{AUTO1}} + \psi_2\,\mathrm{I}_{i\mathrm{AUTO2}} + \varepsilon_i
\end{aligned}
$$

   où $\mathrm{I}_{\mathrm{CSP1}}$, $\mathrm{I}_{\mathrm{CSP2}}$ et $\mathrm{I}_{\mathrm{CSP3}}$ désignent les indicatrices des niveaux 1,2,3 respectifs de la variable CSP, la notation étant similaire pour les autres indicatrices des niveaux des variables qualitatives CR, STOCC, COMP, AUTO.

   L'identifiabilité du modèle nécessite d'imposer 5 contraintes (une pour chacune des 5 variables qualitatives) ; par défaut R pose $\vartheta_1 = \delta_1 = \lambda_1 = \gamma_1 = \psi_1 = 0$

   Sous forme matricielle le modèle identifiable s'écrit : (3) $\mathrm{DOM2} = X\beta + \varepsilon$ où

   - DOM2 est le vecteur aléatoire ($n = 235$, 1) représentant les montants des dommages des sinistres de type 2
   - $X = (\mathbb{1}_n \quad \mathrm{REVENU} \quad \mathrm{RUC} \quad \mathrm{NBPERS} \quad \mathrm{POL1} \quad \mathrm{POL2} \quad \mathrm{POL3} \quad \mathrm{DOM1} \quad \mathrm{DOM3} \quad \mathrm{NBSIN}$
     $\mathrm{I}_{\mathrm{CSP2}} \quad \mathrm{I}_{\mathrm{CSP3}} \quad \mathrm{I}_{\mathrm{CR2}} \quad \mathrm{I}_{\mathrm{CR3}} \quad \mathrm{I}_{\mathrm{STOCC2}} \quad \mathrm{I}_{\mathrm{COMP2}} \quad \mathrm{I}_{\mathrm{COMP3}} \quad \mathrm{I}_{\mathrm{AUTO2}})$
     est la matrice ($n = 235$, $p = 18$) des covariables, de rang $p = 18$
   - $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_9, \vartheta_2, \vartheta_3, \delta_2, \delta_3, \lambda_2, \gamma_2, \gamma_3, \psi_2)$ est le vecteur ($p = 18$, 1) des cœfficients inconnus à estimer
   - ε est le vecteur aléatoire ($n = 235$, 1) des erreurs, supposées indépendantes deux à deux, d'espérance nulle, de même variance σ^2 et de loi gaussienne $\varepsilon \sim \mathrm{U}_n(\mathbf{0}_n\,;\,\sigma^2 I_n)$
     où la variance des erreurs σ^2 inconnue doit être estimée.

3. Tester la significativité globale du modèle (3) : préciser notamment les hypothèses du test, la statistique de test utilisée et sa loi sous l'hypothèse nulle (à justifier).

On teste l'hypothèse nulle $H_0 : (\beta_1, \beta_2, \cdots, \gamma_3, \psi_2) = (0, 0, \cdots, 0, 0)$ ou modèle nul
contre l'alternative $H_1 : (\beta_1, \beta_2, \cdots, \gamma_3, \psi_2) \neq (0, 0, \cdots, 0, 0)$ ou modèle (3) de dimension $p = 18$
au niveau de risque α.
Le modèle nul étant un sous-modèle du modèle (3) de dimension $p_0 = 1$, la statistique de test de Fisher

$$F = \frac{\|P_{V_1} Y - P_{V_0} Y\|^2/(p-p_0)}{\|Y - P_{V_1} Y\|^2/(n-p_1)} = \frac{(SSE_0 - SSE_1)/(p-p_0)}{SSE_1/(n-p)} = \frac{SSR_1/(p-1)}{SSE_1/(n-p)} = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}$$

suit sous H_0 la loi de Fisher $7(p-1; n-p) = 7(17; 217)$
SSR_1 et SSE_1 étant les sommes des carrés expliquée et résiduelle, et R^2 le cœfficient de détermination du modèle (3) sous H_1.
La valeur observée de F est égale à

$$\frac{0.937/(18-1)}{(1-0.937)/(235-18)} = \frac{1605.43/(18-1)}{107.9/(235-18)} = \frac{1605.43/17}{107.9/217} \text{ Ä } 189.926$$

et la p-valeur correspondante $P_{H_0}(F > 189.926)$ Ä 5.546×10^{-120} étant inférieure au seuil de risque maximum $\alpha = 5\%$, on rejette l'hypothèse nulle de nullité globale des cœfficients.

Les résultats numériques ont été obtenus grâce aux commandes R suivantes :

```
mod3 <- lm(DOM2 ~ . -NBAD, data=données)
summary(mod3)

Call:
lm(formula = DOM2 ~ . - NBAD, data = données)

Residuals:
      Min        1Q    Median        3Q       Max
-2.25319  -0.43381   0.00342   0.46921   2.51815

Coefficients:
                             Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                  1.202e+01  7.012e-01  17.144  < 2e-16 ***
CSPEmployes                 -3.244e+00  1.478e-01 -21.952  < 2e-16 ***
CSPProfessions intermediaires -3.231e+00 1.497e-01 -21.578 < 2e-16 ***
REVENU                       1.960e-05  2.202e-05   0.890   0.3743
RUC                          7.134e-06  2.927e-05   0.244   0.8077
CRMoyenne Inf                1.544e-01  2.279e-01   0.677   0.4990
CRMoyenne Sup                4.723e-02  1.523e-01   0.310   0.7568
STOCCProprietaire           -9.325e-02  1.015e-01  -0.919   0.3593
COMPCouple sans enfant      -4.089e-01  2.143e-01  -1.908   0.0577 .
COMPPersonne seule          -2.004e-01  3.325e-01  -0.603   0.5473
NBPERS                       2.620e-01  1.174e-01   2.231   0.0267 *
AUTOPas de vehicule         -3.919e+00  1.289e-01 -30.394  < 2e-16 ***
POL1                         1.037e-01  1.339e-02   7.745 3.62e-13 ***
POL2                         7.113e-03  4.934e-03   1.442   0.1509
POL3                         1.992e-02  2.547e-02   0.782   0.4349
DOM1                         2.714e-03  3.948e-02   0.069   0.9453
DOM3                        -5.804e-02  3.002e-02  -1.933   0.0545 .
NBSIN                        2.380e-02  1.769e-02   1.345   0.1799
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7051 on 217 degrees of freedom
Multiple R-squared:  0.937,Adjusted R-squared:  0.9321
F-statistic: 189.9 on 17 and 217 DF,  p-value: < 2.2e-16
```

```r
anova(lm(DOM2 ~ 1), mod3) # comparaison modèle nul au modèle (3)

Analysis of Variance Table

Model 1: DOM2 ~ 1
Model 2: DOM2 ~ (CSP + REVENU + RUC + CR + STOCC + COMP + NBPERS + NBAD +
    AUTO + POL1 + POL2 + POL3 + DOM1 + DOM3 + NBSIN) - NBAD
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    234 1713.3
2    217  107.9 17    1605.4 189.93 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$<$ Le modèle (3) sur toutes les covariables est globalement significatif au risque $\alpha = 5\%$: au moins une covariable ou une combinaison linéaire de covariables, explique une part significative (93.7024%) de la variabilité des montants des dommages des sinistres de type 2.

4. Quelle(s) variable(s) apporte(nt) une contribution significative à l'explication du montant des dommages des sinistres de type 2 (justifier) ? Le modèle vous semble-t-il parcimonieux ?

- Pour chaque variable quantitative, on teste la nullité de son cœfficient, soit avec le test de Student, soit avec le test de Fisher comparant le modèle sous H_0 où le cœfficient est nul au modèle (3) sous H_1.
  Par exemple pour la variable NBPERS
  on teste l'hypothèse nulle $H_0 : \beta_3 = 0$ contre l'alternative $H_1 : \beta_3 \neq 0$ au niveau de risque α,
  - soit avec la statistique de test de Student $T = \frac{\hat{\beta}_3}{\sqrt{\widehat{\mathrm{var}}(\hat{\beta}_3)}} \sim \mathcal{T}(n-p) = \mathcal{T}(217)$ sous H_0
    sa valeur observée vaut $\frac{0.262}{0.1174} = 2.231$ et la p-valeur bilatérale correspondante $1\,P_{H_0}(T > |2.231|)$ Ä 0.02669 étant inférieure au niveau de risque $\alpha = 5\%$ on rejette l'hypothèse nulle en faveur de l'alternative au risque maximum $\alpha = 5\%$;
  - soit en utilisant la statistique de test de Fisher qui compare le modèle sous H_0 où $\beta_3 = 0$ au modèle (3) sous H_1 et suit une loi de Fisher $\mathcal{F}(1; n-p) = \mathcal{F}(1; 217)$ sous H_0
    la valeur observée de F vaut 4.9785 et la p-valeur correspondante $P_{H_0}(F > 4.9785)$ Ä 0.02669 (égale à celle du test de Student précédent) étant inférieure à $\alpha = 5\%$, on valide l'alternative c'est-à-dire le modèle (3).

  Dans le modèle (3) le cœfficient de la variable NBPERS diffère significativement de 0 au risque $\alpha = 5\%$.
  De manière analogue, on déduit que dans le modèle (3) le cœfficient de la variable POL1 diffère significativement de 0 au risque $\alpha = 5\%$ et que celui de la variable DOM3 diffère significativement de 0 au risque $\alpha = 10\%$.

- Pour chaque variable qualitative (à plus de deux modalités), on veut tester la nullité simultanée de tous les effets.
  Par exemple pour la variable CSP,
  il s'agit de tester l'hypothèse nulle $H_0 : \vartheta_2 = \vartheta_3 = 0$ contre l'alternative H_1 modèle (3) au niveau α, en utilisant le test de Fisher de comparaison de deux modèles emboîtés, le modèle sous H_0 étant celui où on a supprimé la variable CSP modèle (3) ;
  la statistique de test de Fisher suit une loi de Fisher $\mathcal{F}(k-1; n-p) = \mathcal{F}(2; 217)$ sous H_0 où $k = 3$ est le nombre de modalités de la variable CSP ; sa valeur observée de F vaut 281.5286
  et la p-valeur correspondante $P_{H_0}(F > 281.5286)$ Ä 5.13×10^{-61} est inférieure à $\alpha = 5\%$.
  Dans le modèle (3) la variable CSP apporte une contribution significative au risque $\alpha = 5\%$ dans l'explication de la variabilité des montants des dommages des sinistres de type 2.

De manière similaire on déduit que la variable AUTO apporte une contribution significative au risque $\alpha = 5\%$ dans l'explication de la variabilité, et que la variable COMP apporte une contribution significative au risque $\alpha = 10\%$.

< Sur les 14 variables incluses dans le modèle (3), seules NBPERS, POL1, CSP et AUTO (et éventuellement DOM3, COMP), c'est-à-dire assez peu, apporte une contribution significative dans l'explication de la variabilité des montants des dommages des sinistres de type 2 : le modèle n'est donc pas parcimonieux, trop de variables augmentent la complexité du modèle sans apporter d'explication supplémentaire.

Les résultats numériques ont été obtenus grâce à la commande R ci-dessous :

```
drop1(mod3, test='F')

Single term deletions

Model:
DOM2 ~ (CSP + REVENU + RUC + CR + STOCC + COMP + NBPERS + NBAD +
    AUTO + POL1 + POL2 + POL3 + DOM1 + DOM3 + NBSIN) - NBAD
       Df  Sum of Sq      RSS        AIC   F value      Pr(>F)

                                                                      ' 1
```

Question F

On recherche un modèle parcimonieux permettant d'expliquer le montant des dommages des sinistres de **type 2** à partir de l'ensemble des variables (sans interaction).

1. Rechercher le "meilleur" modèle (modèle (4)) :
   - spécifier la(les) procédure(s) utilisée(s) et le(s) critère(s) de sélection choisi(s) ;
   - préciser quelle(s) variable(s) est(sont) sélectionnée(s) dans le modèle (4) retenu.

   On cherche un modèle parcimonieux en optimisant les critères d'information d'Akaike AIC ou bayésien BIC qui pénalisent la log-vraisemblance maximale par le nombre de paramètres du modèle, soit par une procédure de recherche pas à pas mixte, soit par une procédure de recherche exhaustive.

   - Procédure de recherche pas à pas mixte (réalisée avec la commande R suivante)

```
stAIC <- step(mod3, direction="both", trace=0)
summary(stAIC)

Call:
lm(formula = DOM2 ~ CSP + REVENU + COMP + NBPERS + AUTO + POL1 +
    POL2 + DOM3 + NBSIN, data = données)

Residuals:
     Min       1Q   Median       3Q      Max
-2.24501 -0.44578  0.00471  0.44132  2.59592

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.208e+01  3.980e-01  30.365  < 2e-16 ***
CSPEmployes                  -3.214e+00  1.224e-01 -26.262  < 2e-16 ***
CSPProfessions intermediaires -3.221e+00  1.194e-01 -26.970  < 2e-16 ***
REVENU                        1.919e-05  7.130e-06   2.692  0.00765 **
COMPCouple sans enfant       -3.968e-01  2.078e-01  -1.910  0.05741 .
COMPPersonne seule           -1.604e-01  2.890e-01  -0.555  0.57947
NBPERS                        2.812e-01  9.640e-02   2.917  0.00390 **
AUTOPas de vehicule          -3.923e+00  1.259e-01 -31.153  < 2e-16 ***
POL1                          1.058e-01  1.307e-02   8.094 3.72e-14 ***
POL2                          7.373e-03  4.880e-03   1.511  0.13226
DOM3                         -5.151e-02  2.847e-02  -1.809  0.07181 .
NBSIN                         2.473e-02  1.744e-02   1.418  0.15763
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6991 on 223 degrees of freedom
Multiple R-squared:  0.9364,Adjusted R-squared:  0.9333
F-statistic: 298.4 on 11 and 223 DF,  p-value: < 2.2e-16

AIC(stAIC)

[1] 512.3302

stBIC <- step(mod3, direction="both", trace=0, k=log(nrow(données)))
summary(stBIC)

Call:
lm(formula = DOM2 ~ CSP + NBPERS + AUTO + POL1 + RUC, data = données)

Residuals:
```

```
      Min        1Q    Median        3Q       Max
-2.47684  -0.43409  -0.03028   0.47307   2.72748

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.164e+01  2.268e-01  51.306  < 2e-16 ***
CSPEmployes                     -3.184e+00  1.221e-01 -26.076  < 2e-16 ***
CSPProfessions intermediaires  -3.201e+00  1.195e-01 -26.794  < 2e-16 ***
NBPERS                           4.724e-01  4.217e-02  11.202  < 2e-16 ***
AUTOPas de vehicule             -3.926e+00  1.217e-01 -32.265  < 2e-16 ***
POL1                             1.065e-01  1.314e-02   8.105 3.21e-14 ***
RUC                              2.420e-05  1.031e-05   2.348   0.0197 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7102 on 228 degrees of freedom
Multiple R-squared:  0.9329,Adjusted R-squared:  0.9311
F-statistic: 528.1 on 6 and 228 DF,  p-value: < 2.2e-16

BIC(stBIC)

[1] 542.6473
```

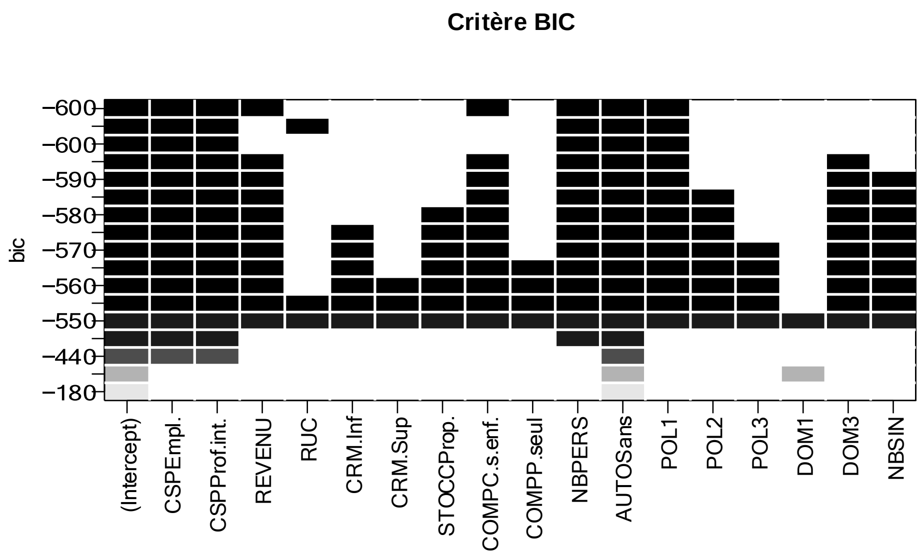
Le modèle obtenu par la recherche pas à pas mixte minimisant

- le critère AIC est le modèle incluant les 9 variables : CSP, REVENU, COMP, NBPERS, AUTO, POL1, POL2, DOM3 et NBSIN expliquant 93.64% de variabilité ; le critère AIC minimum vaut 512.33 pour ce modèle ;
- le critère BIC, inclut les 5 variables : CSP, NBPERS, AUTO, POL1 et RUC expliquant 93.29% de variabilité ; le critère BIC minimum vaut 542.65 pour ce modèle.

- Procédure de recherche exhaustive (réalisée avec la commande R suivante)

```
library(leaps)
# nb de variables max des modèles : toutes les variables (avec indicatrices)
nvm  <-  length(données)-2+3
nv <- nvm+1 # nb de variables + cste
sel <- regsubsets(DOM2 ~ . -NBAD, nvmax=nvm, data=données)
libellé <- c('(Intercept)','CSPEmpl.','CSPProf.int.','REVENU','RUC',
             'CRM.Inf','CRM.Sup','STOCCProp.','COMPC.s.enf.','COMPP.seul',
             "NBPERS","AUTOSans","POL1","POL2","POL3","DOM1","DOM3","NBSIN")
sel$xnames <- libellé
par(mgp=c(3,.6,0))
plot(sel, main="Critère BIC")#, labels=libellé)
grid(nx=nv,ny=nvm, col='white', lty=1, lwd=2)
summary(sel)$bic

 [1] -175.3234 -290.5323 -438.4163 -541.8767 -596.4112 -596.5661 -597.1569
 [8] -593.9599 -590.2575 -587.0438 -582.3980 -577.6153 -572.7638 -567.7225
[15] -562.3513 -556.9558 -551.5014

COMP2  <-  ifelse(COMP==levels(COMP)[2],1,0)  #  indicatrice  du  niveau  2  de  COMP
selBIC <- lm(DOM2 ~ CSP + REVENU + COMP2 + NBPERS + AUTO + POL1)
BIC(selBIC)

[1] 542.0565
```

Graphique F.1



Le modèle minimisant le critère BIC obtenu après une recherche exhaustive (parmi tous les modèles possibles de dimension 2 à 18) comprend les 6 variables (cf Graphique F.1 et Tableau F.1) : CSP, REVENU, NBPERS, AUTO, POL1 et le niveau 2 *Couple sans enfant* de COMP c'est-à-dire en agrégeant les niveaux 1 *Couple avec enfant(s)* et 3 *Personne seule* ; le critère BIC minimum vaut 542.06 pour ce modèle qui explique 93.46% de variabilité.
Le modèle qui minimise le critère Cp de Mallows comprend les 6 variables précédentes et la variable DOM3, et celui qui maximise le critère R2-ajusté est le même que celui qui minimise le critère AIC (cf Tableau F.1).

< Le modèle parcimonieux obtenu grâce au critère BIC, choisi parce qu'il pénalise plus fortement les modèles comportant beaucoup de variables, est le modèle de régression multiple de la variable DOM2 sur

- soit les 5 variables : CSP, NBPERS, AUTO, POL1 et RUC
- soit les 6 variables : CSP, REVENU, NBPERS, AUTO, POL1 et COMP2 variable composition du ménage à 2 niveaux : *Couple avec enfant(s) ou Personne seule* et *Couple sans enfant*.

Le tableau des valeurs observées des critères BIC, R2-ajusté et Cp de Mallows sont obtenues avec la commande R suivante :

```
with(summary(sel),data.frame(outmat, BIC=round(bic,1), R2.adj=round(adjr2,4),
     Cp=round(cp,1)))
```

Tableau F.1

	CSPEmpl.	CSPProf.int.	REVENU	RUC	CRM.Inf	CRM.Sup	STOCCProp.	COMPC.s.enf.	COMPP.seul	NBPERS	AUTOSans	POL1	POL2	POL3	DOM1	DOM3	NBSIN	BIC	R2.adj	Cp
1 (1)											*							-175.3	0.5454	1328.9
2 (1)											*				*			-290.5	0.7268	704.5
3 (1)	*	*									*							-438.4	0.8571	259.1
4 (1)	*	*								*	*							-541.9	0.9097	80.8
5 (1)	*	*								*	*	*						-596.4	0.9298	13.9
6 (1)	*	*		*						*	*	*						-596.6	0.9311	10.3
7 (1)	*	*	*					*		*	*	*						**-597.2**	**0.9326**	**6.4**
8 (1)	*	*	*					*		*	*	*				*		**-594.0**	**0.9329**	**6.3**
9 (1)	*	*	*					*		*	*	*				*	*	-590.3	0.9331	6.6
10 (1)	*	*	*					*		*	*	*	*			*	*	**-587.0**	**0.9335**	**6.5**
11 (1)	*	*	*				*	*		*	*	*	*			*	*	-582.4	0.9334	7.7
12 (1)	*	*	*		*		*	*		*	*	*	*			*	*	-577.6	0.9333	9.1
13 (1)	*	*	*		*		*	*		*	*	*	*	*		*	*	-572.8	0.9332	10.5
14 (1)	*	*	*		*		*	*	*	*	*	*	*	*		*	*	-567.7	0.9330	12.1
15 (1)	*	*	*		*	*	*	*	*	*	*	*	*	*		*	*	-562.4	0.9327	14.1
16 (1)	*	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*	-557.0	0.9324	16.0
17 (1)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	-551.5	0.9321	18.0

2. Avec le modèle (4) calculer la valeur prédite et l'intervalle de prédiction à 95% pour un nouvel assuré dont les caractéristiques sont :
Personne seule, de CSP Employés, de catégorie de revenu Aisée, de revenu (et RUC) dans la tranche 13750, locataire de son habitation, qui ne possède pas de véhicule et dont les cotisations au titre des compléments de polices de type 1, 2 et 3, les montants des dommages pour les sinistres 1 et 3 et le nombre de sinistres antérieurs couverts sont nuls.

   – Expliquer comment la valeur prédite est calculée. Interpréter les résultats obtenus.
   – Comment varie la prédiction pour un nouvel assuré ayant les mêmes caractéristiques mais possédant au moins un véhicule ?

En considérant que le modèle parcimonieux choisi s'écrit : (4) $\mathrm{DOM2} = X\beta + \varepsilon$ où X est la matrice des covariables, β le vecteur des cœfficients et ε le vecteur des erreurs du modèle (4) la valeur prédite du montant des dommages des sinistres de type 2 pour un nouvel assuré $\widehat{\mathrm{DOM2}}_0 = x_0\hat{\beta}$ où x_0 contient les valeurs des covariables du modèle (4) pour le nouvel assuré,

et l'intervalle de prédiction au niveau de confiance (sécurité) $1-\alpha$ du montant des dommages des sinistres de type 2 pour ce nouvel assuré (correspondant aux covariables x_0) est donné par :

$$IP_{1-\alpha}(\mathrm{DOM2}_0) = \left[x_0\hat{\beta} \pm \underbrace{t_{1-\alpha/2}\, S\sqrt{1+h_0}}_{\text{précision au niveau } 1-\alpha} \right]$$

où $h_0 = x_0\,({}^tXX)^{-1}\,{}^tx_0$, S étant l'estimation de l'écart-type des erreurs du modèle (4) et $t_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi de Student $\mathcal{T}(n-p)$: $\Phi_{\mathcal{T}}\left(t_{1-\alpha/2}\right) = 1-\alpha/2$.

Pour le modèle avec les 5 covariables CSP, NBPERS, AUTO, POL1 et RUC
$x_0 = (1 \quad \mathrm{I}_{0CSP2} \quad \mathrm{I}_{0CSP3} \quad \mathrm{NBPERS}_0 \quad \mathrm{I}_{0AUTO2} \quad \mathrm{POL1}_0 \quad \mathrm{RUC}_0) = (1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 13750)$
puisque le nouvel assuré étant employé et ne possédant pas de véhicule, les indicatrices ont pour valeur $\mathrm{I}_{0CSP2} = 1$ $\mathrm{I}_{0CSP3} = 0$ et $\mathrm{I}_{0AUTO2} = 1$
donc la valeur prédite pour le nouvel assuré $\widehat{\mathrm{DOM2}}_0 = x_0\hat{\beta} = (1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 13750)\,\hat{\beta}$

$$\begin{aligned}\widehat{\mathrm{DOM2}}_0 &\simeq 1 \times 11.64 + 1 \times -3.184 + 0 \times -3.201 + 1 \times 0.4724 + 1 \times -3.926 + 0 \times 0.1065 \\ &\quad +13750 \times 2.42 \times 10^{-5} \\ &\simeq 5.332\end{aligned}$$

et l'intervalle de prédiction de $\mathrm{DOM2}_0$ au niveau de sécurité 95% :
$IP_{95\%}(\mathrm{DOM2}_0) \simeq [5.332 \pm 1.421] = [3.91; 6.753]$

Pour le nouvel assuré, le modèle (4) prévoit un montant des dommages des sinistres de type 2 de 5.332 se situant entre 3.91 et 6.753 (5.332 ± 1.421) avec une confiance (ou un niveau de sécurité) de 95%.

Pour un autre assuré ayant les mêmes caractéristiques excepté le fait qu'il possède au moins un véhicule, l'indicatrice vaut $\mathrm{I}_{0AUTO2} = 0$ donc sa prédiction diffère de -3.926 (cœfficient estimé de la variable I_{AUTO2}) ; elle est de $5.332 - (-3.926)$ donc augmente de 3.926 par rapport à un assuré ne possédant pas de véhicule.

Les résultats numériques ont été obtenus grâce aux commandes R ci-dessous :

```
nouv <- data.frame(CSP="Employes", REVENU=13750, RUC=13750, CR="Aise",
                   STOCC="Locataire", COMP="Personne seule", NBPERS=1, NBAD=1,
                   AUTO="Pas de vehicule", POL1=0, POL2=0, POL3=0, DOM1=0,
                   DOM2=0, DOM3=0, NBSIN=0)
P <- predict(stBIC, nouv, interval='prediction')
P

       fit      lwr      upr
1 5.331795 3.910301 6.753288

stBIC$coef  # coefficients modèle 4
```

```
                   (Intercept)                     CSPEmployes 
                  1.163678e+01                   -3.184164e+00 
 CSPProfessions intermediaires                          NBPERS 
                 -3.201287e+00                    4.724220e-01 
            AUTOPas de vehicule                            POL1 
                 -3.926004e+00                    1.064992e-01 
                           RUC 
                  2.420095e-05 

c(1, 1, 0, 1, 1, 0, 13750) %*% stBIC$coef # prédiction pour le nouvel assuré

         [,1]
[1,]  5.331795

(P[3]-P[2])/2 # précision au niveau 95%

[1]  1.421493
```

En utilisant le modèle sur les 6 covariables CSP, REVENU, NBPERS, AUTO, POL1 et COMP2 on obtient des calculs, résultats et conclusions similaires (cf commandes R ci-dessous).

```
nouv2 <- data.frame(CSP="Employes", REVENU=13750, RUC=13750, CR="Aise",
                    STOCC="Locataire", COMP2=0, NBPERS=1, NBAD=1,
                    AUTO="Pas de vehicule", POL1=0, POL2=0, POL3=0, DOM1=0,
                    DOM2=0, DOM3=0, NBSIN=0)
( P2 <- predict(selBIC, nouv2, interval='prediction') )

       fit      lwr      upr
1 5.384051 3.977097 6.791006

(P2[3]-P2[2])/2 # précision au niveau 95%

[1]  1.406954

selBIC$coef

                   (Intercept)                     CSPEmployes 
                  1.187536e+01                   -3.187038e+00 
 CSPProfessions intermediaires                          REVENU 
                 -3.189025e+00                    1.880254e-05 
                         COMP2                          NBPERS 
                 -3.199946e-01                    3.739211e-01 
           AUTOPas de vehicule                            POL1 
                 -3.936727e+00                    1.079460e-01 
```

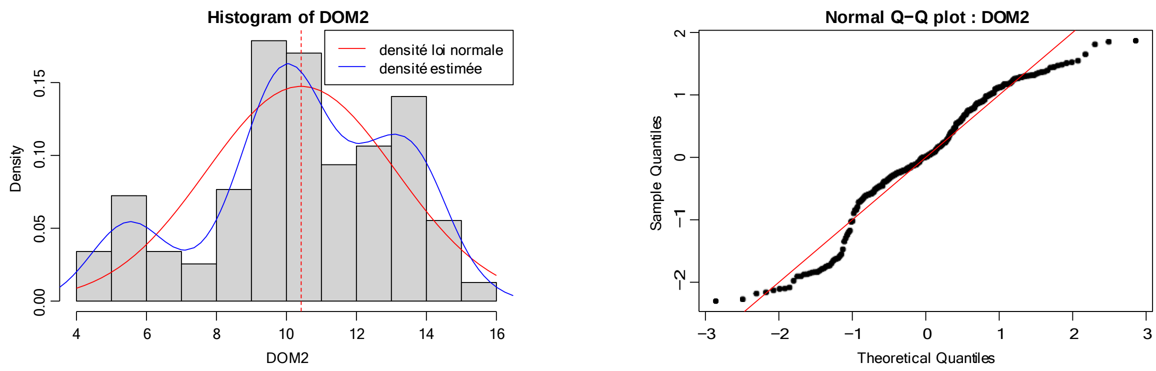
Remarque

La variable à expliquer DOM2 n'est pas gaussienne : on peut cependant la modéliser via un modèle linéaire gaussien puisque le nombre d'observations $n = 235$ est grand.

```
hist(DOM2, freq=F) # histogramme de DOM1
curve(dnorm(x,mean(DOM2),sd(DOM2)), add=T, col='red') # densité loi normale
abline(v=mean(DOM2), lty=2, col='red')
lines(density(DOM2), col='blue') # densité estimée de DOM2
legend('topright', legend=c('densité loi normale', 'densité estimée'),
```

```
        col=c('red','blue'), lty=1)
qqnorm(scale(DOM2), main='Normal Q-Q plot : DOM2', pch=20)
abline(0,1, col='red') # droite d'équation y = x
shapiro.test(DOM2) # test de normalité de DOM1

Shapiro-Wilk normality test

data:   DOM2
W  =  0.95994,  p-value  =  3.84e-06
```

Graphique F.2

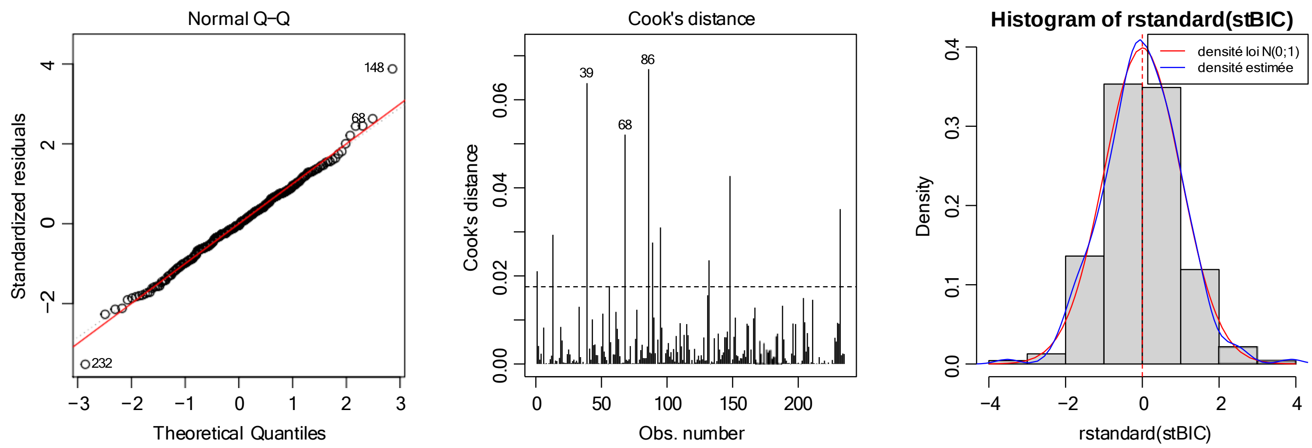


(a) **(b)**

* On vérifie néanmoins le bon comportement des erreurs du modèle (4) retenu via ses résidus.
Par exemple pour le modèle sur les 5 covariables, on valide graphiquement la normalité des erreurs (bon alignement des points sur la droite de Henry (cf Graphique F.3a) et compatibilité de l'histogramme des résidus standardisés avec la loi U (0; 1) (cf Graphique F.3c)), et on observe 10 résidus extrêmes (moins de 5%) et 10 points influents (moins de 5%).
Les conditions d'application du modèle (4) sont validées.

```
length(which(abs(rstandard(stBIC))>2));  length(which(cooks.distance(stBIC)>4/stBIC$df))

[1]  10
[1]  10
```

Graphique F.3



(a) **(b)** **(c)**