Statistique des assurances - Projet

Isabelle Ajtay (41010932) Smail Chabane (38012939) Yuxuan Zhang (38019811)

26 février 2023

Table des matières

1 Contexte et objectifs

Dans le cadre du présent projet, nous somme une compagnie d'assurance non-vie, qui dispose d'un jeu de données historiques sur des ménages ayant souscrit ses polices d'assurance.

Les objectifs ? Déterminer la prime pure pour un ménage intéressé par l'assurance proposée par notre compagnie. On souhaite construire un modèle qui explique les demandes d'indemnisation (Sinistres 1, 2 et 3) en utilisant les données qu'on possède. On veut aussi des modèles pour les prix des polices d'assurance précédemment vendues, le nombre de sinistres, ainsi que la durée de vie d'un contrat d'assurance.

Pour ce faire, nous employerons des méthodes et outils vus en cours de Statistiques des Assurances, et d'autres cours, sous le logiciel *R*.

La mtd train du package caret fait de la cv + boot, et permet d'ajuster des centaines de modèles prédictifs différents, spécifiés facilement avec l'argument method. VerboseIter donne un log du progress, pe masura ce le modèle est ajusté.

On va choisir lequel des deux on met dans le modèle: RUC ou full income. Mais pas les 2 car très fortement correlees. Il y a aussi les quantiles de cet income.

Anat non signif. A suppr.

Police i corresp a sin. i. Il a une assurance de base et rajoute des additionnelles. 0-> le type n'a pas pris cette assurance là.

Sini1 plus facile a modéliser, moins de zeros. 2 et 3 en ont bcp.

1283 outlier pê. On doit les enlever.

2. Description des données

Import des données

La procédure pour lire ligne par ligne ces données est longue. Donc nous les avons exportées dans un fichier .txt pour aller plus vite.

```
cat("\f")                                    # clears the console, by sending Ctrl + L
```

```r
data = read.table("data.txt", sep = " ", header=T , encoding = "UTF-8")
#data = read.table("assurance_complete_corrige.R") #, sep = "", header=T)
```

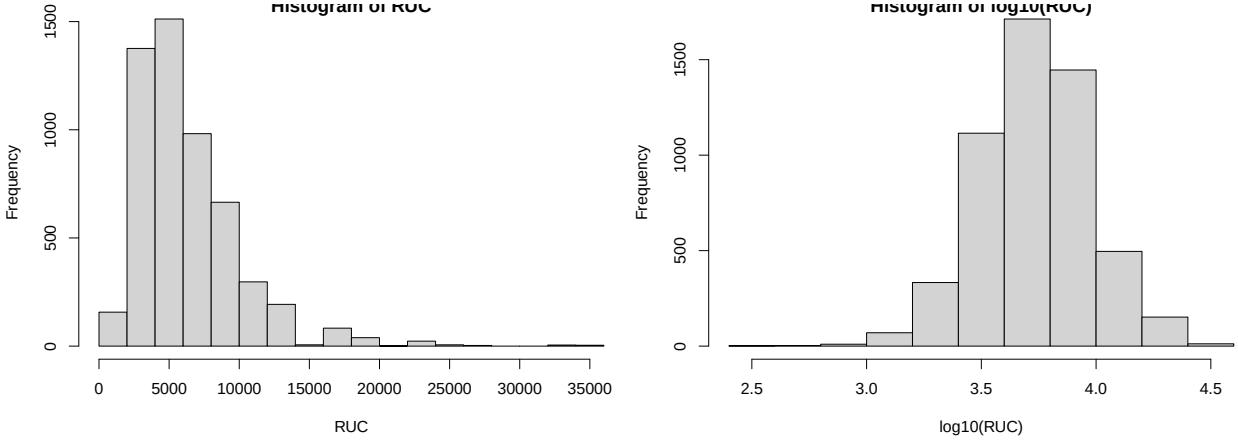
2.1 Analyses univariées

On a utilisé str pour afficher les informations simples concernant les variables, et summary pour afficher les données statistiques pour chaque variable.

Voici les variables: [1] pcs 2 RUC 3 cs 4 reves 5 crevpp 6 region 7 habi 8 Ahabi 9 Atyph 10 agecat 11 Acompm 12 nbpers 13 enfants 14 Anat 15 Bauto 16 “Nbadulte” 17 Sinistre1” 18 Sinistre2 19 Sinistre3” 20 Police1 21 “Police2” 22 “Police3” 23 “durPolice1” 24 Durée” 25 NSin” 26 censure” 27 Sinistre0

On observe que les variables *pcs, cs, region, crevpp, agecat* et *habi* sont qualitatives, malgré leur format caractères ou numérique. On rémedie.

```
## [1] 1 2 3 4 5 7 8 9
```

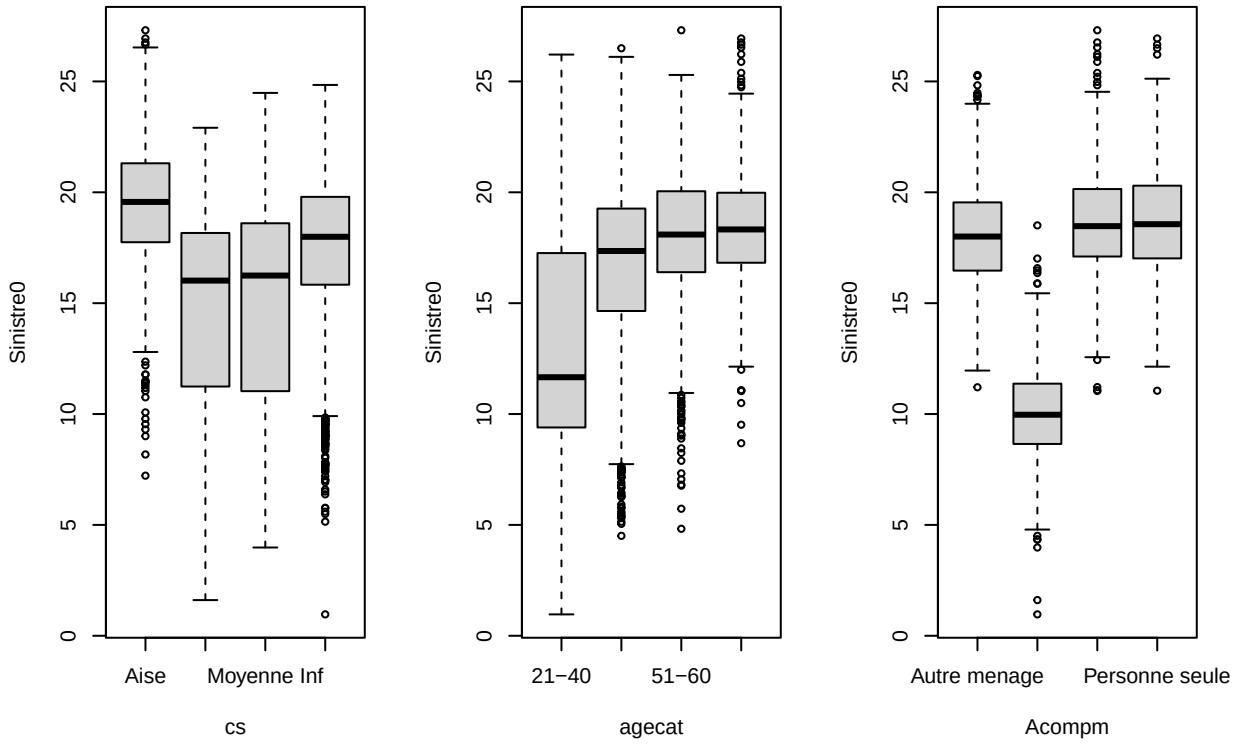



```
## Le chargement a nécessité le package : Hmisc
## Le chargement a nécessité le package : lattice
## Le chargement a nécessité le package : Formula
##
## Attachement du package : 'Hmisc'
## L'objet suivant est masqué depuis 'package:psych':
##
##     describe
## Les objets suivants sont masqués depuis 'package:questionr':
##
##     describe, wtd.mean, wtd.table, wtd.var
## Les objets suivants sont masqués depuis 'package:dplyr':
##
##     src, summarize
## Les objets suivants sont masqués depuis 'package:base':
##
##     format.pval, units
## funModeling v.1.9.4 :)
## Examples and tutorials at livebook.datascienceheroes.com
##  / Now in Spanish: librovivodecienciadedatos.ai
```

```
##
## Attachement du package : 'funModeling'

## L'objet suivant est masqué depuis 'package:questionr':
##
##     freq

##      variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1         pcs       0    0.00    0    0     0     0  factor      8
## 2         RUC       0    0.00    0    0     0     0 numeric    249
## 3          cs       0    0.00    0    0     0     0  factor      4
## 4       reves       0    0.00    0    0     0     0 integer     24
## 5      crevpp       0    0.00    0    0     0     0  factor      4
## 6      region       0    0.00    0    0     0     0  factor      8
## 7        habi    1352   25.26    0    0     0     0  factor      9
## 8       Ahabi       0    0.00    0    0     0     0  factor      5
## 9       Atyph       0    0.00    0    0     0     0  factor      3
## 10     agecat       0    0.00    0    0     0     0  factor      4
## 11     Acompm       0    0.00    0    0     0     0  factor      4
## 12     nbpers       0    0.00    0    0     0     0 integer     10
## 13    enfants       0    0.00    0    0     0     0  factor      2
## 14       Anat       0    0.00    0    0     0     0  factor      3
## 15      Bauto     443    8.28    0    0     0     0 numeric      2
## 16   Nbadulte       0    0.00    0    0     0     0 integer      8
## 17  Sinistre1    4085   76.33    0    0     0     0 numeric    298
## 18  Sinistre2    4797   89.63    0    0     0     0 numeric    112
## 19  Sinistre3    1780   33.26    0    0     0     0 numeric    818
## 20    Police1     773   14.44    0    0     0     0 numeric   1894
## 21    Police2     102    1.91    0    0     0     0 numeric   3740
## 22    Police3     412    7.70    0    0     0     0 numeric   1863
## 23 durPolice1     773   14.44    0    0     0     0 numeric   1296
## 24      Durée      23    0.43    0    0     0     0 integer    454
## 25       NSin    1392   26.01    0    0     0     0 integer     20
## 26    censure    2045   38.21    0    0     0     0 integer      2
## 27  Sinistre0       0    0.00    0    0     0     0 numeric   5352

##      variable         mean      std_dev variation_coef        p_01        p_05
## 1         RUC 6.277521e+03 3.709063e+03      0.5908484 1451.613000 2291.667000
## 2       reves 1.487995e+04 7.460939e+04      5.0140883 3500.000000 4500.000000
## 3      nbpers 3.038303e+00 1.409790e+00      0.4640055    1.000000    1.000000
## 4       Bauto 9.172272e-01 2.755642e-01      0.3004318    0.000000    0.000000
## 5    Nbadulte 2.388453e+00 1.049460e+00      0.4393889    1.000000    1.000000
## 6   Sinistre1 1.242663e+00 9.060978e+00      7.2915835    0.000000    0.000000
## 7   Sinistre2 1.615049e-01 1.150240e+00      7.1220173    0.000000    0.000000
## 8   Sinistre3 1.837128e+00 2.733674e+00      1.4880148    0.000000    0.000000
## 9     Police1 3.750700e+00 5.020503e+00      1.3385510    0.000000    0.000000
## 10    Police2 1.301746e+01 1.326108e+01      1.0187154    0.000000    0.642750
## 11    Police3 2.110487e+00 2.422230e+00      1.1477112    0.000000    0.000000
## 12 durPolice1 5.190665e+08 3.797352e+10     73.1573447    0.000000    0.000000
## 13      Durée 2.491054e+02 6.299362e+02      2.5287941    1.000000    4.000000
## 14       NSin 4.249253e+00 3.811811e+00      0.8970544    0.000000    0.000000
## 15    censure 6.178999e-01 4.859462e-01      0.7864482    0.000000    0.000000
## 16  Sinistre0 1.617321e+01 4.295643e+00      0.2656023    6.266625    8.219549
##            p_25         p_50         p_75        p_95        p_99   skewness
## 1  3.823529e+03 5.500000e+03  7812.500000 13235.29000 19117.65000  2.1021792
```

```
## 2  8.500000e+03 1.125000e+04 16250.000000 27500.00000 40000.00000 42.4537837
## 3  2.000000e+00 3.000000e+00     4.000000     5.00000     6.00000  0.4071934
## 4  1.000000e+00 1.000000e+00     1.000000     1.00000     1.00000 -3.0284495
## 5  2.000000e+00 2.000000e+00     3.000000     4.00000     5.00000  1.0696991
## 6  0.000000e+00 0.000000e+00     0.000000     3.90000    25.89800 19.4471373
## 7  0.000000e+00 0.000000e+00     0.000000     0.65000     3.09800 16.4665539
## 8  0.000000e+00 7.050000e-01     2.560000     7.29900    12.33215  2.8096043
## 9  5.375000e-01 1.950000e+00     5.017000    13.45450    23.08685  2.8813781
## 10 3.828750e+00 9.060000e+00    17.950000    39.27640    60.74648  2.1722796
## 11 5.200000e-01 1.420000e+00     2.832750     6.63580    11.69000  2.9798673
## 12 7.254295e-02 4.387771e-01     2.327894    10.82940    40.35390 73.1368593
## 13 2.300000e+01 4.250000e+01   233.000000  1073.00000  2944.00000  7.0239595
## 14 0.000000e+00 4.000000e+00     6.000000    12.00000    16.00000  1.1551963
## 15 0.000000e+00 1.000000e+00     1.000000     1.00000     1.00000 -0.4852836
## 16 1.327694e+01 1.725581e+01    19.271757    21.77459    23.64269 -0.6010396
##       kurtosis         iqr                              range_98
## 1    10.998320 3988.971000                  [1451.613, 19117.65]
## 2  1848.653446 7750.000000                         [3500, 40000]
## 3     2.739259    2.000000                                [1, 6]
## 4    10.171507    0.000000                                [0, 1]
## 5     4.563682    1.000000                                [1, 5]
## 6   553.060764    0.000000                           [0, 25.898]
## 7   342.795883    0.000000                 [0, 3.09799999999996]
## 8    17.255933    2.560000                         [0, 12.33215]
## 9    15.749212    4.479500                         [0, 23.08685]
## 10   10.342418   14.121250                 [0, 60.7464799999999]
## 11   19.228753    2.312750                            [0, 11.69]
## 12 5350.000187    2.255351                 [0, 40.3538973725331]
## 13   71.130104  210.000000                              [1, 2944]
## 14    4.912461    6.000000                                [0, 16]
## 15    1.235500    1.000000                                 [0, 1]
## 16    2.532559    5.994818 [6.26662546758083, 23.6426937685246]
##                                range_80
## 1                  [2741.936, 11029.41]
## 2                         [6500, 22500]
## 3                                [1, 5]
## 4                                [1, 1]
## 5                                [1, 4]
## 6                            [0, 1.25]
## 7                            [0, 0.23]
## 8                            [0, 5.29]
## 9                           [0, 9.6745]
## 10                       [1.3805, 30.05]
## 11              [0.1, 4.85400000000001]
## 12                [0, 7.05746495927491]
## 13                             [7, 639]
## 14                              [0, 10]
## 15                               [0, 1]
## 16 [9.39437980247863, 20.9060359215896]
```

```
##         0%        25%        50%        75%       100%
##   277.7778  3823.5290  5500.0000  7812.5000 35294.1200
```

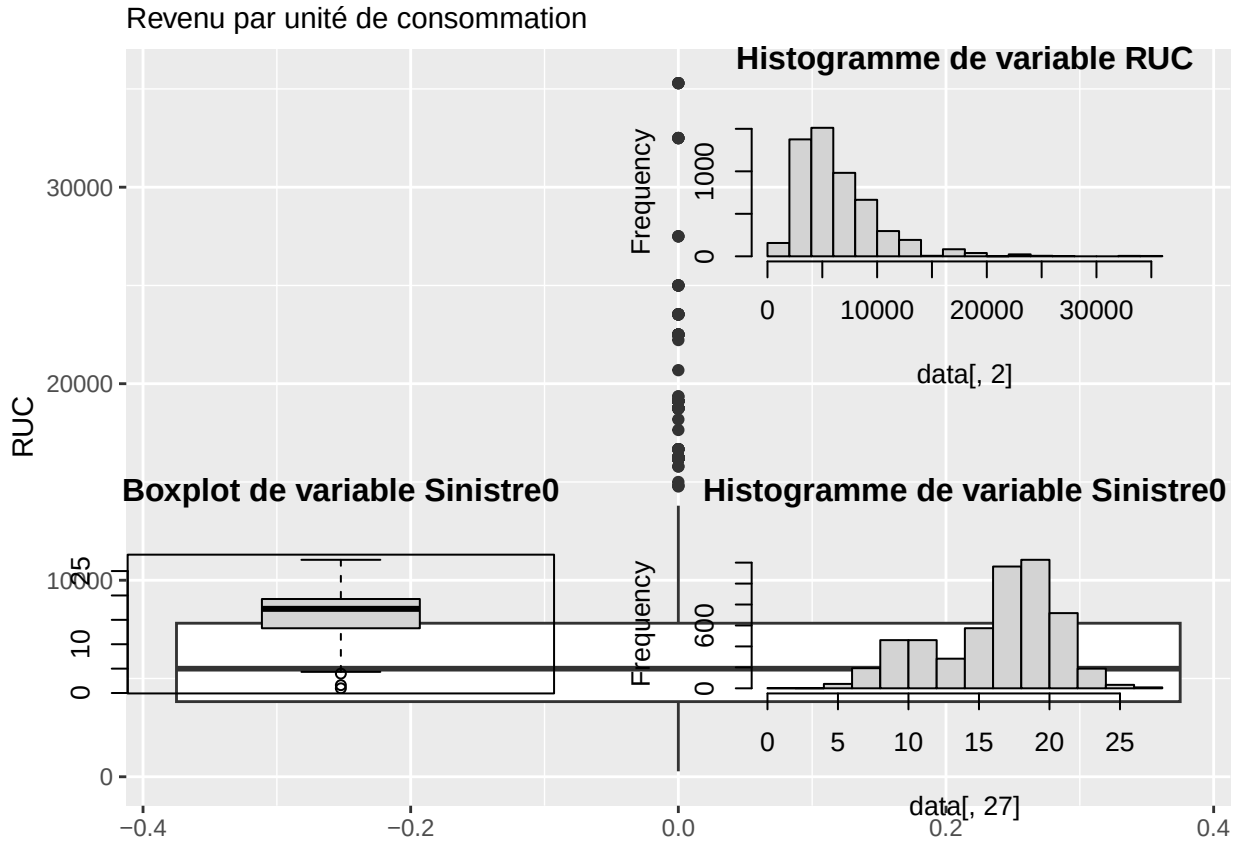


On a representé les boxplots des variables RUC et Sinistre0. Pour écraser les grandes valeurs, on utilise la fonction log.

```
par(mfrow=c(2,2))
boxplot(log(data[,2]),main="Boxplot de variable RUC")
ggplot(data, aes(y=RUC, fill=Durée)) + geom_boxplot(orientation = "x") + labs(subtitle = "Revenu par uni
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

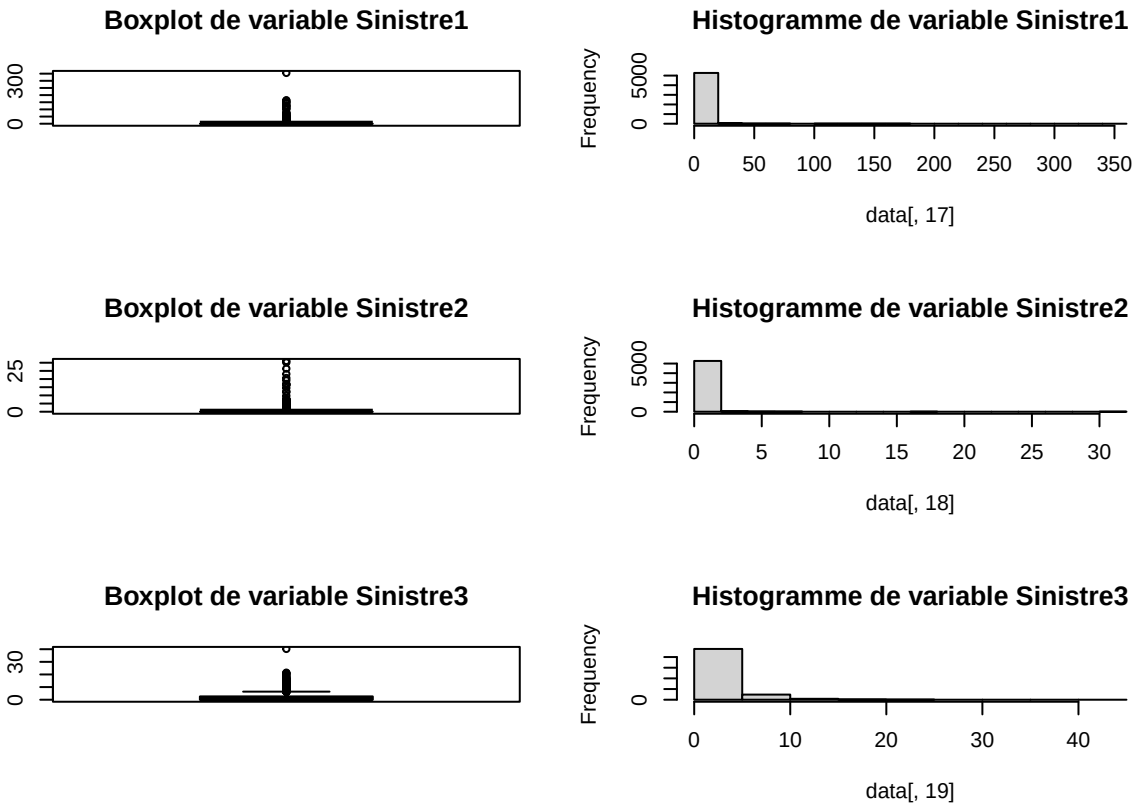
```
hist(data[,2],main="Histogramme de variable RUC")
#  erreur ici chez Eva : stat_density requires an x or y aesthetic ggplot(data = data.frame(data[,2]))
boxplot(data[,27],main="Boxplot de variable Sinistre0")
hist(data[,27],main="Histogramme de variable Sinistre0")
```




```
density(data[,27])
```

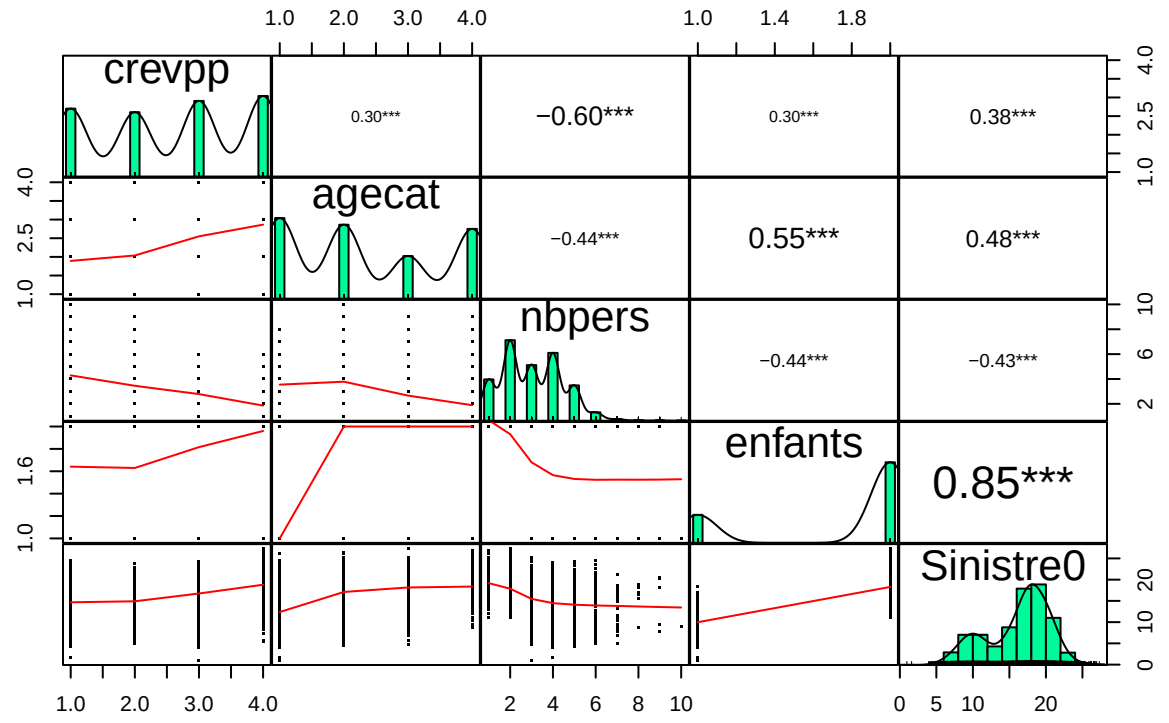
```
##
## Call:
##  density.default(x = data[, 27])
##
## Data: data[, 27] (5352 obs.);    Bandwidth 'bw' = 0.6943
##
##        x                y
##  Min.   :-1.118   Min.   :1.270e-06
##  1st Qu.: 6.509   1st Qu.:8.995e-04
##  Median :14.136   Median :2.309e-02
##  Mean   :14.136   Mean   :3.275e-02
##  3rd Qu.:21.763   3rd Qu.:4.731e-02
##  Max.   :29.390   Max.   :1.226e-01
```

```
par(mfrow=c(3,2))
boxplot(data[,17],main="Boxplot de variable Sinistre1")
hist(data[,17],main="Histogramme de variable Sinistre1")
boxplot(data[,18],main="Boxplot de variable Sinistre2")
hist(data[,18],main="Histogramme de variable Sinistre2")
boxplot(data[,19],main="Boxplot de variable Sinistre3")
hist(data[,19],main="Histogramme de variable Sinistre3")
```



2.2 Analyses bivariées

A travers un graphique des variables numériques deux à deux, nous regardons comment évoluent les variables ensemble, et s'il y a des "tendances" reconnaissables. Par exemple, sur le graphique ci-dessous, qui contient les analyses bivariées complètes, on voit une tendance linéaire croissante (et corrélation positive significative) entre *pib* et *recc.* (fonction trouvée à ref. 7: analyse bi + corrélations).



On peut constater que les variabilités des trois types de *Sinistres* sont toutes grandes.

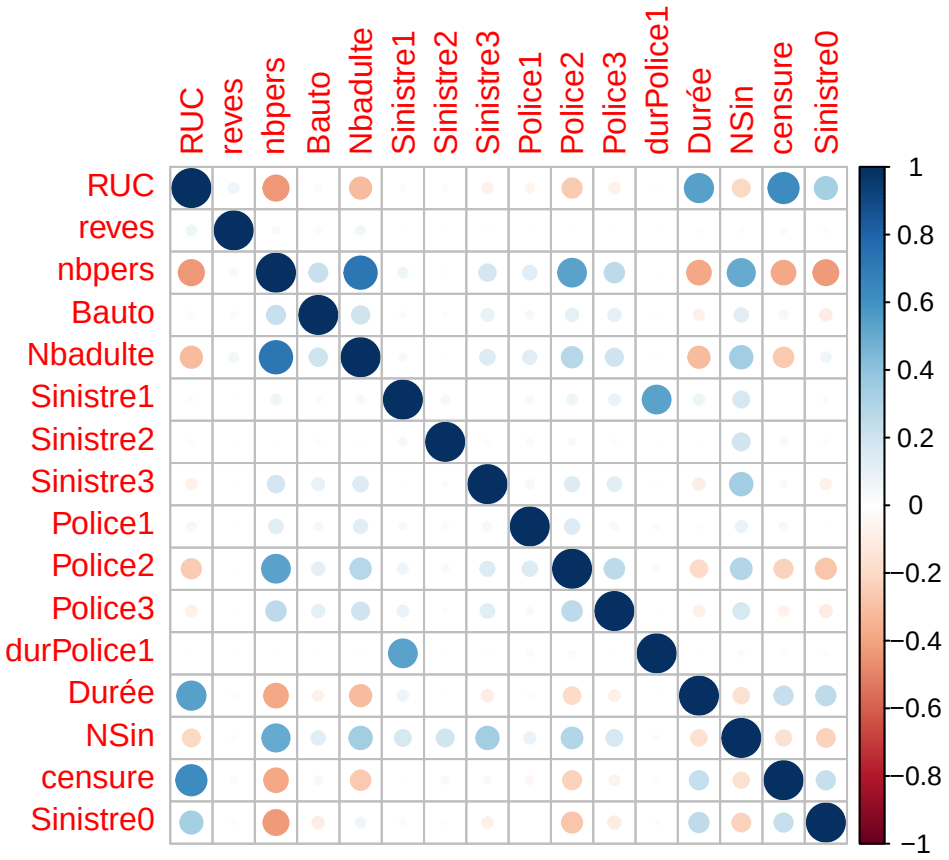
```
aggregate(data[,c(17,18,19,27)],list(data[,1]),mean)
```

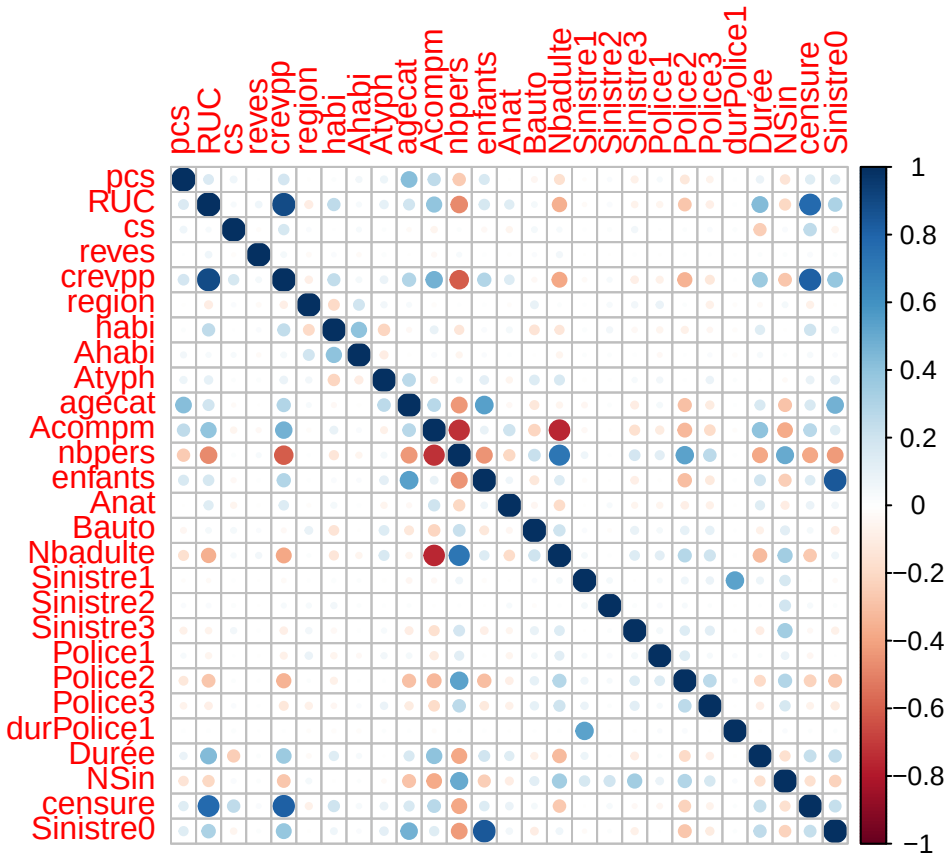
```
##                                Group.1 Sinistre1   Sinistre2 Sinistre3
## 1                     Agr. exploitants 0.4860776 0.009051724  2.983017
## 2        Artisans, comm., chefs d'ent. 0.3559116 0.097458564  1.836381
## 3     Autres pers. sans activite prof. 1.4143169 0.107540984  1.761721
## 4 Cadres et prof. intellectuelles sup. 1.6937083 0.184058333  2.249010
## 5                              Employes 1.2392184 0.159409429  1.771824
## 6                              Ouvriers 1.6113979 0.128146194  1.841745
## 7           Professions intermediaires 1.7689008 0.207553551  2.214859
## 8                              Retraites 0.5480867 0.185876093  1.395999
##   Sinistre0
## 1  15.01849
## 2  16.12462
## 3  17.41922
## 4  16.83139
## 5  15.57469
## 6  14.30565
## 7  15.66231
## 8  18.36901
```

```
#Representation des correlations
variables_quantitatives = data %>% select_if(is.numeric) %>% cor()
kable(variables_quantitatives, digits=3)
```

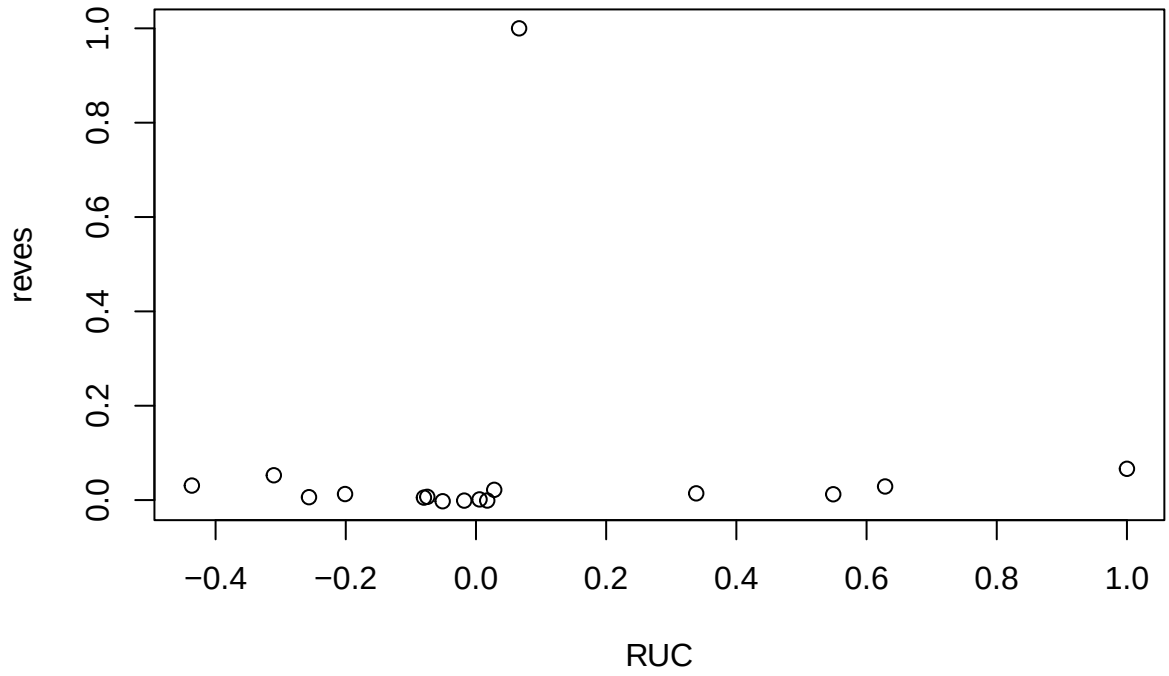
	RUC	reves	nbpers	Bauto	Nbadulte	Sinistre1	Sinistre2	Sinistre3	Police1	Police2	Police3	durPolice1	Durée	NSin	censure	Sinistre0
RUC	1.000	0.066	-0.436	0.028	-0.311	-0.018	0.017	-0.075	-0.051	-0.256	-0.080	0.005	0.549	-0.201	0.628	0.338
reves	0.066	1.000	0.031	0.022	0.053	-0.001	-0.001	0.007	-0.002	0.006	0.005	0.001	0.012	0.013	0.029	0.014
nbpers	-0.436	0.031	1.000	0.228	0.720	0.064	0.006	0.188	0.129	0.531	0.262	0.009	-0.387	0.508	-0.385	-0.429
Bauto	0.028	0.022	0.228	1.000	0.208	0.023	-0.013	0.099	0.040	0.110	0.107	0.004	-0.072	0.127	0.044	-0.091
Nbadulte	-0.311	0.053	0.720	0.208	1.000	0.042	0.008	0.148	0.124	0.288	0.202	0.008	-0.312	0.346	-0.262	0.066
Sinistre1	-0.018	-0.001	0.064	0.023	0.042	1.000	0.042	-0.001	0.030	0.070	0.082	0.534	0.073	0.175	-0.001	-0.021
Sinistre2	0.017	-0.001	0.006	-0.013	0.008	0.042	1.000	0.015	0.029	0.031	0.010	-0.002	0.004	0.200	0.037	0.012
Sinistre3	-0.075	0.007	0.188	0.099	0.148	-0.001	0.015	1.000	0.048	0.140	0.130	-0.006	-0.090	0.346	-0.032	-0.077
Police1	-0.051	-0.002	0.129	0.040	0.124	0.030	0.029	0.048	1.000	0.146	0.035	0.020	0.013	0.087	-0.044	0.000
Police2	-0.256	0.006	0.531	0.110	0.288	0.070	0.031	0.140	0.146	1.000	0.265	0.028	-0.193	0.297	-0.228	-0.280
Police3	-0.080	0.005	0.262	0.107	0.202	0.082	0.010	0.130	0.035	0.265	1.000	-0.001	-0.082	0.172	-0.069	-0.103
durPolice1	0.005	0.001	0.009	0.004	0.008	0.534	-0.002	-0.006	0.020	0.028	-0.001	1.000	-0.005	0.021	0.011	0.011
Durée	0.549	0.012	-0.387	-0.072	-0.312	0.073	0.004	-0.090	0.013	-0.193	-0.082	0.005	1.000	-0.167	0.231	0.256
NSin	-0.201	0.013	0.508	0.127	0.346	0.175	0.200	0.346	0.087	0.297	0.172	0.021	-0.167	1.000	-0.162	-0.221
censure	0.628	0.029	-0.385	0.044	-0.262	-0.001	0.037	-0.032	-0.044	-0.228	-0.069	0.011	0.231	-0.162	1.000	0.231
Sinistre0	0.338	0.014	-0.429	-0.091	0.066	-0.021	0.012	-0.077	0.000	-0.280	-0.103	0.011	0.256	-0.221	0.231	1.000

```
corrplot(variables_quantitatives)
```


RUC
reves
nbpers
Bauto
Nbadulte
Sinistre1
Sinistre2
Sinistre3
Police1
Police2
Police3
durPolice1
Durée
NSin
censure
Sinistre0
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



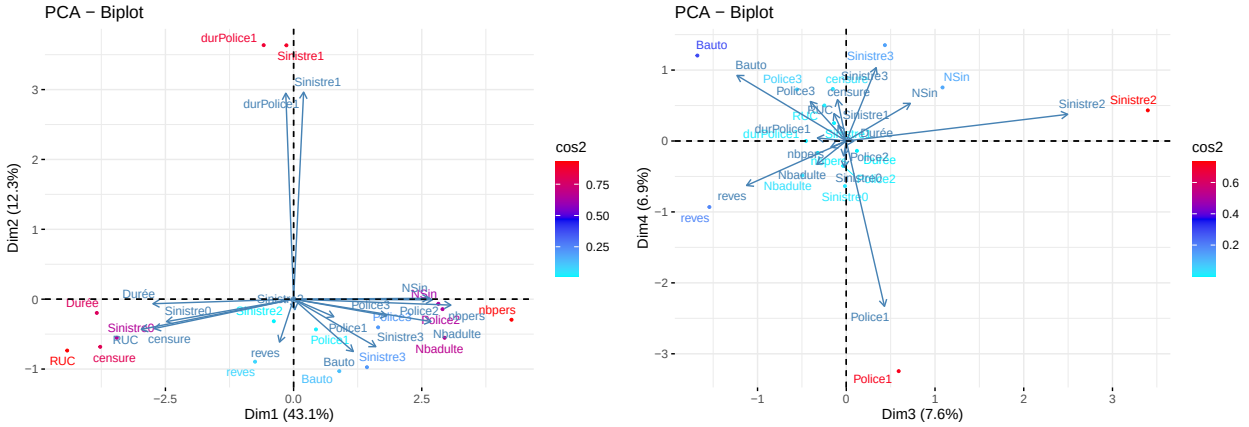

```
# Boxplot
# boxplot(data$column_name, main = "Boxplot of column_name", xlab = "Column name", ylab = "Values")

# Histogram
# hist(data$column_name, main = "Histogram of column_name", xlab = "Values", ylab = "Frequency", col =

# Estimateurs de la densitÃ©
# density_plot <- density(data$column_name, main = "Density Plot of column_name", xlab = "Values", ylab
# lines(density_plot, col = "red")

# Statistiques basiques
# summary(data$column_name)

plot(variables_quantitatives)
```




```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 16 individuals, described by 16 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

```
##         eigenvalue percentage of variance cumulative percentage of variance
## comp 1        6.90                  43.11                             43.11
## comp 2        1.97                  12.28                             55.39
## comp 3        1.21                   7.59                             62.98
## comp 4        1.10                   6.88                             69.87
## comp 5        1.08                   6.73                             76.59
```

```
## comp 6       0.86          5.35                    81.94
## comp 7       0.81          5.03                    86.98
## comp 8       0.69          4.30                    91.28
## comp 9       0.42          2.65                    93.93
## comp 10      0.33          2.08                    96.02
## comp 11      0.29          1.80                    97.82
## comp 12      0.17          1.04                    98.86
## comp 13      0.12          0.78                    99.64
## comp 14      0.04          0.26                    99.90
## comp 15      0.02          0.10                   100.00
```



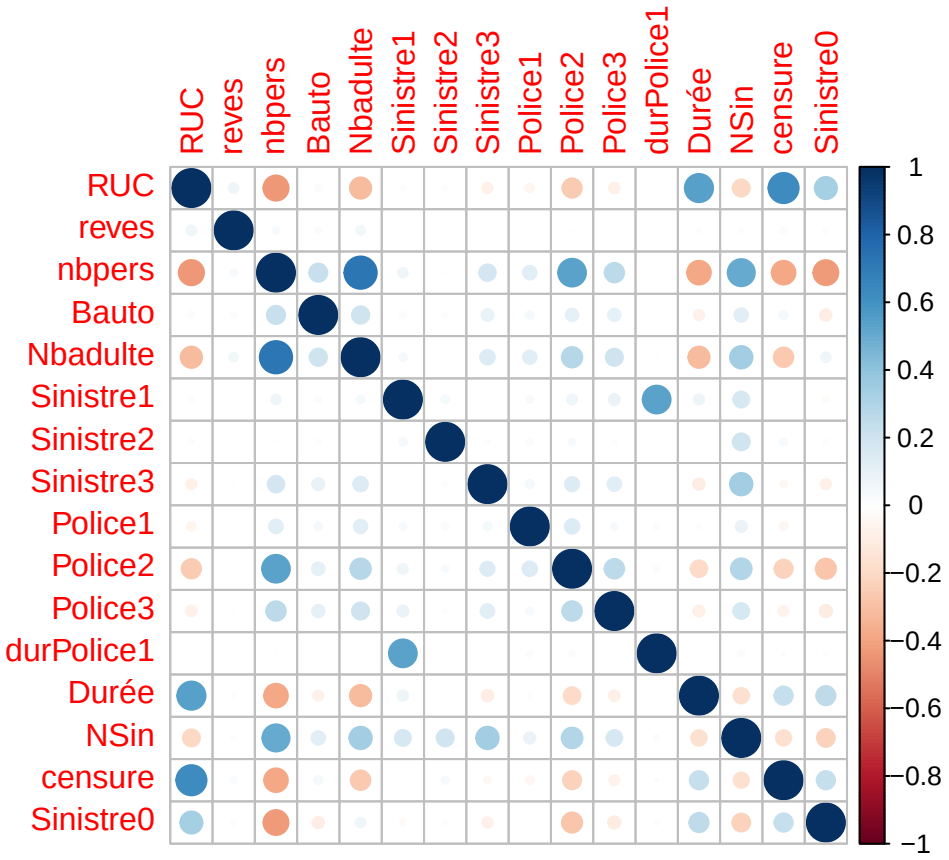
2.3 Dépendances entre les variables

Si les variables explicatives sont fortement corrélées entre elles, cela peut rendre l'interprétation du coefficient de corrélation plus difficile. Nous allons vérifier la corrélation entre les variables explicatives en utilisant le coefficient V de Cramèr. On peut utiliser la fonction assocstats pour calculer les coefficients V de Cramer pour toutes les paires de variables et stocker les résultats dans une matrice, puis créer une heatmap à partir de cette matrice en utilisant la fonction heatmap :

3. Modélisation des sinistres et des primes pures

3.1 Problème d'endogénéïté dans les variables

```
# Selection des variables quantitatives
quant_vars <- sapply(data, is.numeric)

# Matrice de correlation
cor_matrix <- cor(data[, quant_vars])
corrplot(cor_matrix)
```




```
# Tests de causalité de Granger pour toutes les paires de variables
# la fonction grangertest() permet de tester si une variable X est un prédicteur significatif d'une aut
quant_vars <- as.matrix(quant_vars)
d = data[, quant_vars]
for(i in 1:(ncol(d) - 1)){
  for(j in (i + 1):ncol(d)){

    result <- grangertest(d[,i], d[,j], order = 2)

    print(paste("Granger causality test entre ", colnames(d)[i],
                " et ", colnames(d)[j], ":", result[2,4]))
  }
}
```

```
## [1] "Granger causality test entre  RUC  et  reves : 0.0923837756616444"
## [1] "Granger causality test entre  RUC  et  nbpers : 0.00124220485708754"
## [1] "Granger causality test entre  RUC  et  Bauto : 0.000242581095264626"
## [1] "Granger causality test entre  RUC  et  Nbadulte : 0.0172836315703496"
## [1] "Granger causality test entre  RUC  et  Sinistre1 : 0.709483878283177"
## [1] "Granger causality test entre  RUC  et  Sinistre2 : 0.0951611194510821"
## [1] "Granger causality test entre  RUC  et  Sinistre3 : 0.0981000677158548"
## [1] "Granger causality test entre  RUC  et  Police1 : 0.737522698532588"
## [1] "Granger causality test entre  RUC  et  Police2 : 0.0123843096814322"
## [1] "Granger causality test entre  RUC  et  Police3 : 0.912874664962623"
## [1] "Granger causality test entre  RUC  et  durPolice1 : 0.234660821645245"
## [1] "Granger causality test entre  RUC  et  Durée : 0.00278908899804597"
## [1] "Granger causality test entre  RUC  et  NSin : 0.283574612045495"
```

```
## [1] "Granger causality test entre  RUC  et  censure : 5.81376380854141e-09"
## [1] "Granger causality test entre  RUC  et  Sinistre0 : 0.00736870516547412"
## [1] "Granger causality test entre  reves  et  nbpers : 0.18968408454619"
## [1] "Granger causality test entre  reves  et  Bauto : 0.00138772460027939"
## [1] "Granger causality test entre  reves  et  Nbadulte : 0.36176438407316"
## [1] "Granger causality test entre  reves  et  Sinistre1 : 0.970631057503471"
## [1] "Granger causality test entre  reves  et  Sinistre2 : 0.934632216327838"
## [1] "Granger causality test entre  reves  et  Sinistre3 : 0.87081127385712"
## [1] "Granger causality test entre  reves  et  Police1 : 0.465186907224791"
## [1] "Granger causality test entre  reves  et  Police2 : 0.366262300097807"
## [1] "Granger causality test entre  reves  et  Police3 : 0.38259062929957"
## [1] "Granger causality test entre  reves  et  durPolice1 : 0.99606824048349"
## [1] "Granger causality test entre  reves  et  Durée : 0.89938550170445"
## [1] "Granger causality test entre  reves  et  NSin : 0.956546525314893"
## [1] "Granger causality test entre  reves  et  censure : 0.45554355536769"
## [1] "Granger causality test entre  reves  et  Sinistre0 : 0.365219373654749"
## [1] "Granger causality test entre  nbpers  et  Bauto : 0.212820782315102"
## [1] "Granger causality test entre  nbpers  et  Nbadulte : 0.193083295586577"
## [1] "Granger causality test entre  nbpers  et  Sinistre1 : 0.558110531140403"
## [1] "Granger causality test entre  nbpers  et  Sinistre2 : 0.49305523182955"
## [1] "Granger causality test entre  nbpers  et  Sinistre3 : 0.944191036071153"
## [1] "Granger causality test entre  nbpers  et  Police1 : 0.0381762395658839"
## [1] "Granger causality test entre  nbpers  et  Police2 : 0.38758443522889"
## [1] "Granger causality test entre  nbpers  et  Police3 : 0.79937810796575"
## [1] "Granger causality test entre  nbpers  et  durPolice1 : 0.272894547109501"
## [1] "Granger causality test entre  nbpers  et  Durée : 0.0905118575346298"
## [1] "Granger causality test entre  nbpers  et  NSin : 0.0480562871228446"
## [1] "Granger causality test entre  nbpers  et  censure : 0.00455159012190975"
## [1] "Granger causality test entre  nbpers  et  Sinistre0 : 0.362139149025635"
## [1] "Granger causality test entre  Bauto  et  Nbadulte : 0.210225721893678"
## [1] "Granger causality test entre  Bauto  et  Sinistre1 : 0.999921147172683"
## [1] "Granger causality test entre  Bauto  et  Sinistre2 : 0.73890292789447"
## [1] "Granger causality test entre  Bauto  et  Sinistre3 : 0.993727606162102"
## [1] "Granger causality test entre  Bauto  et  Police1 : 0.246862506387724"
## [1] "Granger causality test entre  Bauto  et  Police2 : 0.210858520434915"
## [1] "Granger causality test entre  Bauto  et  Police3 : 0.247303227505292"
## [1] "Granger causality test entre  Bauto  et  durPolice1 : 0.916976635532682"
## [1] "Granger causality test entre  Bauto  et  Durée : 0.189726004205516"
## [1] "Granger causality test entre  Bauto  et  NSin : 0.507864976570639"
## [1] "Granger causality test entre  Bauto  et  censure : 0.00198971669715613"
## [1] "Granger causality test entre  Bauto  et  Sinistre0 : 0.124812334872418"
## [1] "Granger causality test entre  Nbadulte  et  Sinistre1 : 0.358541171759482"
## [1] "Granger causality test entre  Nbadulte  et  Sinistre2 : 0.683643172870649"
## [1] "Granger causality test entre  Nbadulte  et  Sinistre3 : 0.954146981569278"
## [1] "Granger causality test entre  Nbadulte  et  Police1 : 0.130836636428486"
## [1] "Granger causality test entre  Nbadulte  et  Police2 : 0.659774129946827"
## [1] "Granger causality test entre  Nbadulte  et  Police3 : 0.50444501848523"
## [1] "Granger causality test entre  Nbadulte  et  durPolice1 : 0.275432907148821"
## [1] "Granger causality test entre  Nbadulte  et  Durée : 0.0210922967848045"
## [1] "Granger causality test entre  Nbadulte  et  NSin : 0.381605011258915"
## [1] "Granger causality test entre  Nbadulte  et  censure : 0.00262936227638448"
## [1] "Granger causality test entre  Nbadulte  et  Sinistre0 : 0.348576217380067"
## [1] "Granger causality test entre  Sinistre1  et  Sinistre2 : 0.827711518387258"
## [1] "Granger causality test entre  Sinistre1  et  Sinistre3 : 0.570989882925704"
```

```
## [1] "Granger causality test entre  Sinistre1  et  Police1 : 0.736836059511227"
## [1] "Granger causality test entre  Sinistre1  et  Police2 : 0.288643457072958"
## [1] "Granger causality test entre  Sinistre1  et  Police3 : 0.6554707617976"
## [1] "Granger causality test entre  Sinistre1  et  durPolice1 : 0.976563675926668"
## [1] "Granger causality test entre  Sinistre1  et  Durée : 0.473758676644474"
## [1] "Granger causality test entre  Sinistre1  et  NSin : 0.36196682034159"
## [1] "Granger causality test entre  Sinistre1  et  censure : 0.38432432846914"
## [1] "Granger causality test entre  Sinistre1  et  Sinistre0 : 0.134747038950314"
## [1] "Granger causality test entre  Sinistre2  et  Sinistre3 : 0.300524713276513"
## [1] "Granger causality test entre  Sinistre2  et  Police1 : 0.10246562700862"
## [1] "Granger causality test entre  Sinistre2  et  Police2 : 0.158801981749613"
## [1] "Granger causality test entre  Sinistre2  et  Police3 : 0.370757787884079"
## [1] "Granger causality test entre  Sinistre2  et  durPolice1 : 0.98055544905637"
## [1] "Granger causality test entre  Sinistre2  et  Durée : 0.520300685937179"
## [1] "Granger causality test entre  Sinistre2  et  NSin : 0.498259153148845"
## [1] "Granger causality test entre  Sinistre2  et  censure : 0.652951329078083"
## [1] "Granger causality test entre  Sinistre2  et  Sinistre0 : 0.355856344901385"
## [1] "Granger causality test entre  Sinistre3  et  Police1 : 0.201632295337263"
## [1] "Granger causality test entre  Sinistre3  et  Police2 : 0.0553060810276148"
## [1] "Granger causality test entre  Sinistre3  et  Police3 : 0.688841817622712"
## [1] "Granger causality test entre  Sinistre3  et  durPolice1 : 0.820036164981182"
## [1] "Granger causality test entre  Sinistre3  et  Durée : 0.0979144699157492"
## [1] "Granger causality test entre  Sinistre3  et  NSin : 0.0590920136482579"
## [1] "Granger causality test entre  Sinistre3  et  censure : 0.420641926037821"
## [1] "Granger causality test entre  Sinistre3  et  Sinistre0 : 0.541651999430448"
## [1] "Granger causality test entre  Police1  et  Police2 : 0.667133611759315"
## [1] "Granger causality test entre  Police1  et  Police3 : 0.00354879403230157"
## [1] "Granger causality test entre  Police1  et  durPolice1 : 0.785179474093198"
## [1] "Granger causality test entre  Police1  et  Durée : 0.132813120000447"
## [1] "Granger causality test entre  Police1  et  NSin : 0.0203246545252461"
## [1] "Granger causality test entre  Police1  et  censure : 0.204917633529259"
## [1] "Granger causality test entre  Police1  et  Sinistre0 : 0.846171264937825"
## [1] "Granger causality test entre  Police2  et  Police3 : 0.310224237087878"
## [1] "Granger causality test entre  Police2  et  durPolice1 : 0.055655850228255"
## [1] "Granger causality test entre  Police2  et  Durée : 0.277199668621766"
## [1] "Granger causality test entre  Police2  et  NSin : 0.0239566356380064"
## [1] "Granger causality test entre  Police2  et  censure : 0.0692500258718777"
## [1] "Granger causality test entre  Police2  et  Sinistre0 : 0.269125934485199"
## [1] "Granger causality test entre  Police3  et  durPolice1 : 0.653405172596898"
## [1] "Granger causality test entre  Police3  et  Durée : 0.0113542062495392"
## [1] "Granger causality test entre  Police3  et  NSin : 0.000464286639831717"
## [1] "Granger causality test entre  Police3  et  censure : 3.1569081679773e-05"
## [1] "Granger causality test entre  Police3  et  Sinistre0 : 0.0436598567098487"
## [1] "Granger causality test entre  durPolice1  et  Durée : 0.961630268620227"
## [1] "Granger causality test entre  durPolice1  et  NSin : 0.939390629133645"
## [1] "Granger causality test entre  durPolice1  et  censure : 0.567426085872365"
## [1] "Granger causality test entre  durPolice1  et  Sinistre0 : 0.0396821681991153"
## [1] "Granger causality test entre  Durée  et  NSin : 0.572377359207435"
## [1] "Granger causality test entre  Durée  et  censure : 0.000658970825685853"
## [1] "Granger causality test entre  Durée  et  Sinistre0 : 0.0983144988858764"
## [1] "Granger causality test entre  NSin  et  censure : 0.774359173306188"
## [1] "Granger causality test entre  NSin  et  Sinistre0 : 0.647017398450364"
## [1] "Granger causality test entre  censure  et  Sinistre0 : 0.151342377417034"
```

Si on fixe $\alpha = 0.05$, alors il y a une causalité entre Sinistre0 et les variables suivantes : RUC/durPolice1. La méthode des MCO donne l'estimateur le plus efficient s'il n'y a pas d'endogéneïté.

S'il y a de l'endogéneïté, OLS (MCO) va donner des résultats inconsistants. L'estimateur des variables instrumentales va être consistant, mais inéfficient.

```
# Régression linéaire multiple
model1 <- lm(Sinistre0 ~ ., data = data)

# Afficher le résumé du modèle
summary(model1)
```

```
##
## Call:
## lm(formula = Sinistre0 ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5038 -1.4182 -0.0375  1.4986  8.0183
##
## Coefficients: (5 not defined because of singularities)
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             1.722e+01  5.198e-01  33.135   <2e-16
## pcsArtisans, comm., chefs d'ent.        1.050e-01  2.608e-01   0.403   0.6873
## pcsAutres pers. sans activite prof.     3.288e-02  2.660e-01   0.124   0.9016
## pcsCadres et prof. intellectuelles sup. 7.633e-02  2.369e-01   0.322   0.7473
## pcsEmployes                            -6.494e-02  2.232e-01  -0.291   0.7711
## pcsOuvriers                            -1.258e-01  2.145e-01  -0.587   0.5575
## pcsProfessions intermediaires          -1.499e-01  2.227e-01  -0.673   0.5009
## pcsRetraites                           -3.271e-02  2.485e-01  -0.132   0.8953
## RUC                                     2.154e-04  2.187e-05   9.849   <2e-16
## csModeste                               6.520e-02  2.566e-01   0.254   0.7995
## csMoyenne Inf                           1.390e-01  2.043e-01   0.680   0.4963
## csMoyenne Sup                           1.672e-01  1.570e-01   1.065   0.2868
## reves                                  -3.535e-07  3.969e-07  -0.891   0.3731
## crevpp2eme quartile                    -5.961e-02  1.399e-01  -0.426   0.6701
## crevpp3eme quartile                    -1.857e-01  1.962e-01  -0.946   0.3440
## crevpp4eme quartile                    -2.174e-01  2.445e-01  -0.889   0.3741
## region2                                -3.011e-01  2.113e-01  -1.425   0.1542
## region3                                -2.570e-01  2.304e-01  -1.116   0.2646
## region4                                -2.103e-01  2.198e-01  -0.957   0.3388
## region5                                -1.379e-01  2.136e-01  -0.645   0.5187
## region7                                 7.066e-02  2.208e-01   0.320   0.7490
## region8                                -2.338e-01  2.167e-01  -1.079   0.2807
## region9                                -1.347e-01  2.222e-01  -0.606   0.5443
## habi1                                   4.859e-02  1.266e-01   0.384   0.7011
## habi2                                   9.944e-02  1.435e-01   0.693   0.4885
## habi3                                  -1.135e-01  1.439e-01  -0.789   0.4304
## habi4                                  -1.486e-01  1.289e-01  -1.153   0.2488
## habi5                                  -9.448e-02  1.280e-01  -0.738   0.4603
## habi6                                  -9.277e-03  1.207e-01  -0.077   0.9387
## habi7                                   5.967e-02  9.343e-02   0.639   0.5231
## habi8                                  -1.180e-01  2.221e-01  -0.531   0.5954
## AhabiParis + Agglomeration                     NA         NA      NA       NA
## AhabiUn. urb. de 10 000 a 99 999 hab.          NA         NA      NA       NA
```

```
## AhabiUn. urb. de 100 000 hab. et +              NA         NA      NA       NA
## AhabiUn. urb. de 2 000 a 9 999 hab.             NA         NA      NA       NA
## AtyphNon declare                          1.006e-01  2.579e-01   0.390   0.6964
## AtyphProprietaire                        -1.998e-02  6.863e-02  -0.291   0.7710
## agecat41-50                               2.780e-03  9.272e-02   0.030   0.9761
## agecat51-60                              -7.727e-02  1.126e-01  -0.686   0.4925
## agecat61-96                              -1.032e-01  1.668e-01  -0.619   0.5362
## AcompmCouple avec enfant(s)              -8.070e+00  1.256e-01 -64.245   <2e-16
## AcompmCouple sans enfant                  1.130e-01  1.127e-01   1.002   0.3162
## AcompmPersonne seule                      1.101e-01  1.628e-01   0.676   0.4988
## nbpers                                    1.817e-02  6.336e-02   0.287   0.7743
## enfantsPas d'enfants                             NA         NA      NA       NA
## AnatMenage francais                      -2.970e-01  2.053e-01  -1.447   0.1481
## AnatNon declare                          -4.437e-01  2.357e-01  -1.882   0.0599
## Bauto                                     1.930e-02  1.198e-01   0.161   0.8720
## Nbadulte                                 -2.155e-02  6.865e-02  -0.314   0.7536
## Sinistre1                                -5.460e-03  3.989e-03  -1.369   0.1711
## Sinistre2                                 2.712e-02  2.643e-02   1.026   0.3049
## Sinistre3                                 5.726e-03  1.178e-02   0.486   0.6269
## Police1                                   8.886e-03  6.099e-03   1.457   0.1452
## Police2                                   2.874e-03  2.730e-03   1.053   0.2924
## Police3                                   6.160e-03  1.321e-02   0.466   0.6411
## durPolice1                                1.137e-12  9.263e-13   1.227   0.2197
## Durée                                     2.832e-05  6.300e-05   0.450   0.6531
## NSin                                      1.200e-02  9.962e-03   1.205   0.2284
## censure                                   9.184e-02  1.205e-01   0.762   0.4459
##
## (Intercept)                              ***
## pcsArtisans, comm., chefs d'ent.
## pcsAutres pers. sans activite prof.
## pcsCadres et prof. intellectuelles sup.
## pcsEmployes
## pcsOuvriers
## pcsProfessions intermediaires
## pcsRetraites
## RUC                                      ***
## csModeste
## csMoyenne Inf
## csMoyenne Sup
## reves
## crevpp2eme quartile
## crevpp3eme quartile
## crevpp4eme quartile
## region2
## region3
## region4
## region5
## region7
## region8
## region9
## habi1
## habi2
## habi3
## habi4
```

```
## habi5
## habi6
## habi7
## habi8
## AhabiParis + Agglomeration
## AhabiUn. urb. de 10 000 a 99 999 hab.
## AhabiUn. urb. de 100 000 hab. et +
## AhabiUn. urb. de 2 000 a 9 999 hab.
## AtyphNon declare
## AtyphProprietaire
## agecat41-50
## agecat51-60
## agecat61-96
## AcompmCouple avec enfant(s)             ***
## AcompmCouple sans enfant
## AcompmPersonne seule
## nbpers
## enfantsPas d'enfants
## AnatMenage francais
## AnatNon declare                         .
## Bauto
## Nbadulte
## Sinistre1
## Sinistre2
## Sinistre3
## Police1
## Police2
## Police3
## durPolice1
## Durée
## NSin
## censure
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.148 on 5298 degrees of freedom
## Multiple R-squared:  0.7524, Adjusted R-squared:  0.7499
## F-statistic: 303.8 on 53 and 5298 DF,  p-value: < 2.2e-16
```

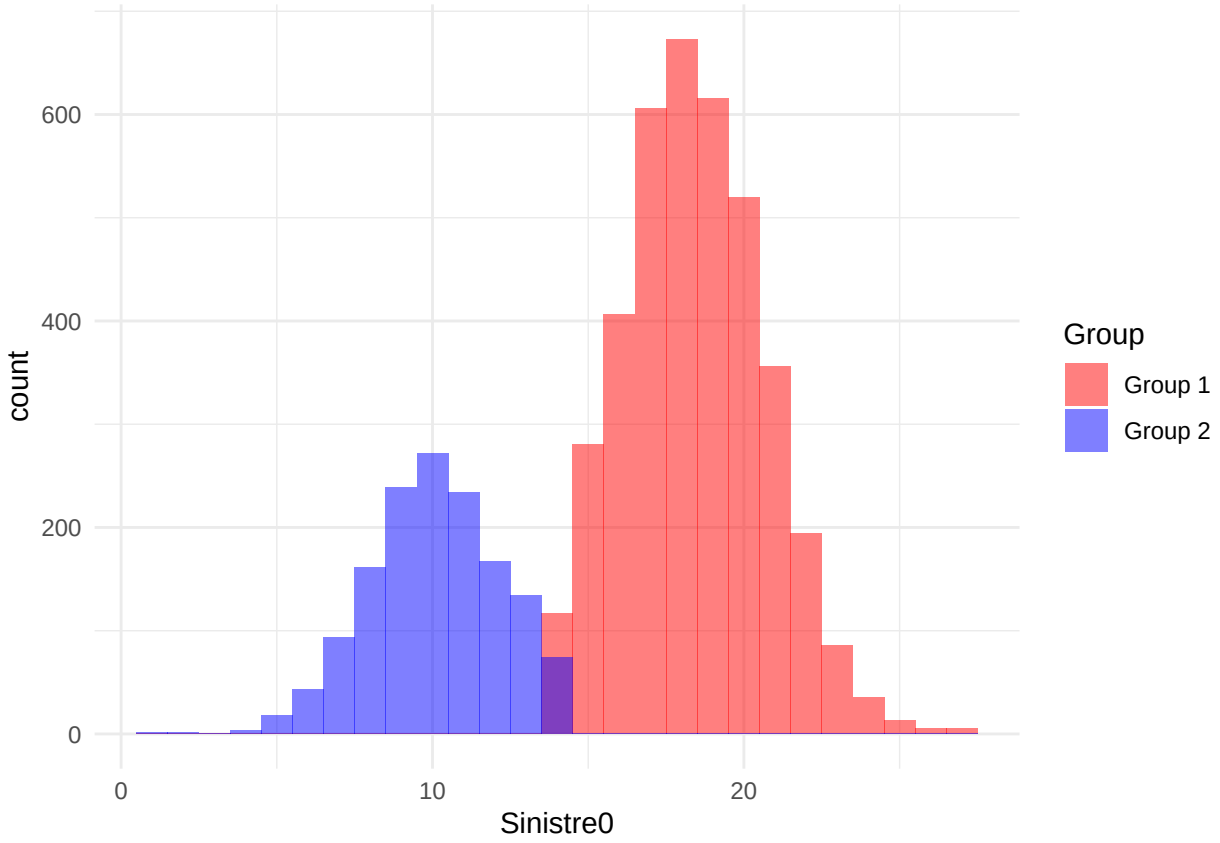
```
summary(selectionAIC)
```

```
##
## Call:
## lm(formula = Sinistre0 ~ RUC + Acompm + Police1 + NSin, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4423 -1.4309 -0.0272  1.4871  7.8258
##
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  1.682e+01  8.389e-02  200.442   <2e-16 ***
## RUC                          2.081e-04  8.701e-06   23.923   <2e-16 ***
## AcompmCouple avec enfant(s) -8.009e+00  7.598e-02 -105.410   <2e-16 ***
## AcompmCouple sans enfant     3.721e-02  8.351e-02    0.446   0.6560
```

```
## AcompmPersonne seule        -1.015e-02  1.025e-01   -0.099   0.9211
## Police1                      1.050e-02  5.889e-03    1.782   0.0748 .
## NSin                         1.602e-02  8.531e-03    1.878   0.0604 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.146 on 5345 degrees of freedom
## Multiple R-squared:  0.7507, Adjusted R-squared:  0.7504
## F-statistic:  2682 on 6 and 5345 DF,  p-value: < 2.2e-16
```

3.2 Modélisation de *Sinistre0*

On observe sur l'histogramme de la variable *Sinistre0* qu'il y a deux sous-populations distinctes, qu'on sépare.



Analyse multivariée:

```
## Le chargement a nécessité le package : rgl
```

```
## Le chargement a nécessité le package : mgcv
```

Comme la variable *Sinistre0* n'a pas de zéros (toutes les valeurs observées sont positives), on va la modéliser avec un linéaire:

```
##
## Call:
## lm(formula = Sinistre0 ~ groupe + RUC + crevpp + Acompm + nbpers +
##     Anat + Durée, data = data2)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.4291 -1.4599  0.0031  1.4972  7.8096
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  5.755e+00  1.084e+00    5.309 1.14e-07 ***
## groupe                             NA         NA       NA       NA
## RUC                          1.512e+00  1.359e-01   11.125  < 2e-16 ***
## crevpp2eme quartile         -3.442e-01  1.133e-01   -3.038 0.002392 **
## crevpp3eme quartile         -5.250e-01  1.481e-01   -3.544 0.000397 ***
## crevpp4eme quartile         -4.560e-01  2.128e-01   -2.143 0.032129 *
## AcompmCouple avec enfant(s) -8.091e+00  7.868e-02 -102.836  < 2e-16 ***
## AcompmCouple sans enfant     5.430e-02  1.039e-01    0.523 0.601313
## AcompmPersonne seule        -4.929e-03  1.491e-01   -0.033 0.973623
## nbpers                       4.535e-02  3.990e-02    1.137 0.255794
## AnatMenage francais         -3.533e-01  2.042e-01   -1.730 0.083634 .
## AnatNon declare             -4.496e-01  2.348e-01   -1.915 0.055557 .
## Durée                        1.942e-04  5.834e-05    3.329 0.000878 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.158 on 5340 degrees of freedom
## Multiple R-squared:  0.7481, Adjusted R-squared:  0.7476
## F-statistic:  1441 on 11 and 5340 DF,  p-value: < 2.2e-16
```

Prof: , on peut essayer le modéliser par un modèle linéaire f. de: compo du ménage, catégorie d'âge, type d'habitation, nationalité, voiture ou pas, catégorie socio-prof, la région, le revenu (un d'eux!!). On ne va pas mettre la variable censure ou pas, qui nous dit si l'individu est dans la base ou pas. Il faut pas la mettre. Durée? Ca peut faire du sens, mais ça risque de compliquer un peu le modèle. Mais durée de Police est celle de la Police1, or je ne sais à quoi c lié le Sinistre0. Les enfants, ça peut être rendondant avec d'autres variables (type de ménage pê). Ca donne un premier modèle.

Plein de choses ne sont pas signif. Couple avec enfant c *très signif*; la compo du mènage, et le revenu aussi, très signif. On va faire le tri, regarder les AIC.

Et si on fait des analyses numériques, notamment des graphes, il y a des phénomènes un peu bizzares: des gros packets. Cad on a des individus dont les fitted values sont très petites; et pour d'autres, très grosses. Donc c un mélange. Il y a vraiment DEUX POPULATIONS la dédans - une certaine **hétérosced**.

Pour modéliser l'hétérosced, il y a qqch de très simple dans un 1er temps: prendre les résidus du modèle et les mettre au carré. Puis regresser sur les variables mises dans le modèles. Car si on regresse et on voit qu'il y a des variables qui sont significatives, càd que les residus dépendent des variables observées. Donc un moyen très simple, lin_modele_1.1, si on plot les residus, on va les mettre au carré + nommer () et on va les regresser (LM) sur les variables que j'ai vu qu'étaient significatives: RUC, Acompm (compo du menage). On voit que RUC est très signif - donc il y a de l'hétérosced. Donc faudra ut. les moindres carrés linéaires généralisés.

GLM: on peut essayer de modéliser. On rajoute une nouvelle var, delta: le fait que le sinistre1 soit >0. Cad j'ai une sinistre, vs. j'ai pas de sinistre. Donc j'ai une nouvelle var, que je vais modéliser par un probit: modèle lin gen, var. delta expliquée par : cs, anat, type... Fam Binomiale, avec modèle soit Probit ou logit. Cloglog (double exponentielle). Rcmd donne le modèle: glm, famille de lien binomial, avec une famille logit. Ca sort tous les estimateurs, et faudra choisir quelles sont pertinentes pour savoir si on aura un sinistre de type 1 ou pas.

On voit que la catégorie d'âge est importante en particulier pour les personnes agées; la région aussi, mais aussi la compo du ménage; et le type être proprio ou pas est légéremment signif. Faudra qu'on choisisse nous les variables.

On va voir, très souvent, que les modèles linéaire, gamma, autre modèles, ne sont pas très différents, a la fin. En terme des coefficients ou des residus. Mais s'il y a beaucoup de zéros, ça va être plus compliqué. Sur Sinistre0 on pourra essayer deja de faire des choses.2.3 Modélisation de Sinistre 1 ou 2 ou 3 (au moins un)

notamment pour Sinistre1 à 3 on choisira entre modèle gamma combiné à probit/logit, tobit, tobit généralisé ou double hurdle pour des variables bien choisies

3.4 Modèle pour le prix de Police 1 ou 2 ou 3 (au moins un)

3.5 Modèle retenu au final

Le choix du modèle retenu au final et les critères choisis devront être justifiés.

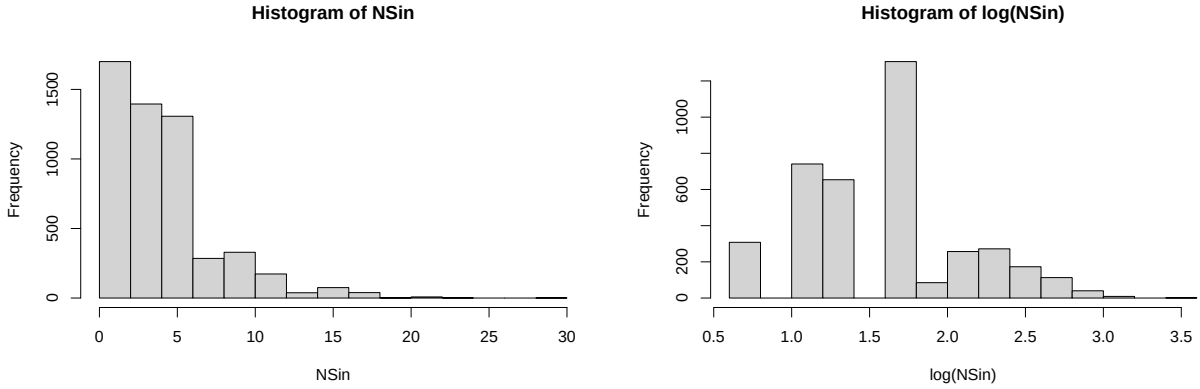
IV regressions

The four kinds of variables in IV

- Y = outcome variables
- X = endogenous, causal variable(s)
- Z = instrument(s): doivent être exogenes, càd leur influence sur Y se fait seulement via leur influence sur X, la var endogene
- W = any exogenous variables not including instruments

4. Modélisation pour les prix : le nombre de sinistres et la tarification des nouveaux arrivants

4.1 Modèle pour le nombre de sinistres, NSin



4.2 Méthode de tarification pour les nouveaux arrivants

On a deux types de modèles pour la tarification :

- *tarification a priori* : pour une nouvelle police d'assurance souscrite, nous ne savons pas quelles garanties ont été souscrites, et connaissons uniquement les caractéristiques du ménage qui a souscrit le contrat. Concrétement, nous n'utiliserons pas les variables *Police*.
- *tarification a posteriori* : nous savons ici quelles garanties ont été souscrites, et le prix payé pour celles-ci. On souhaite savoir le coût estimé pour l'assureur de ce ménage. Ce modèle est différent car il s'avère qu'une plus grande couverture en assurance est associée à des coûts plus importants pour l'assureur. Ces modèles sont plus compliquées car on aura un souci d'endogénéité entre les variables.

5. Estimation des durées

5.1 Estimateur de Kaplan-Meier

5.2 Modèle de Cox