Statistique des assurances - Projet

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

27 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

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.

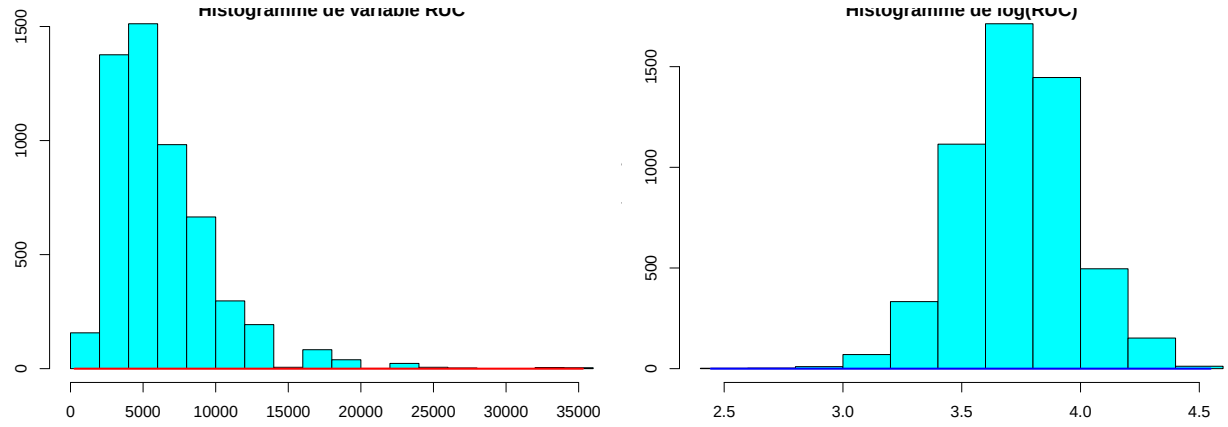
Voici les variables:

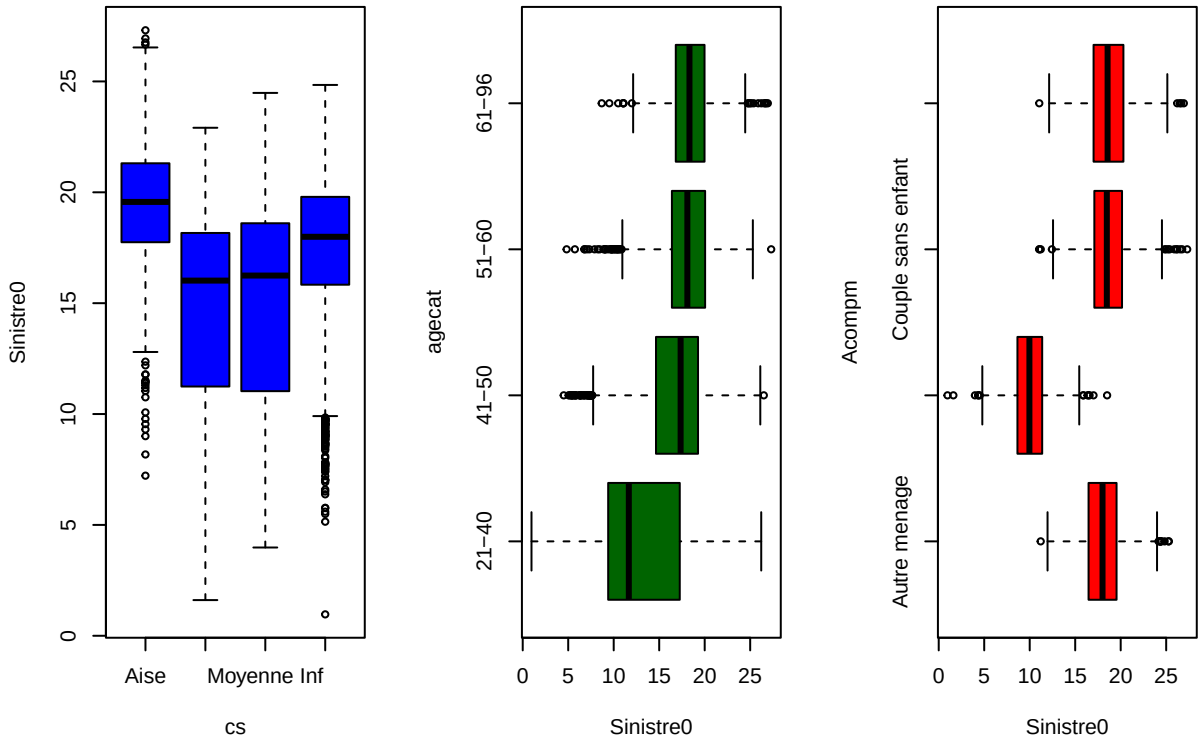
1. pcs
2. RUC
3. cs
4. reves
5. crevpp
6. region
7. habi
8. Ahabi
9. Atyph
10. agecat 11 Acompm 12 nbpers 13 enfants 14 Anat 15 Bauto 16 “Nbadulte”
11. 17 Sinistre1” 18 Sinistre2 19 Sinistre3” 20 Police1 21 “Police2” 22 “Police3”
12. 23 “durPolice1” 24 Durée” 25 NSin” 26 censure” 27 Sinistre0

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

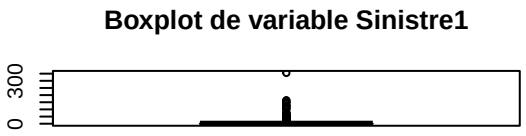
```
## [1] 1 2 3 4 5 7 8 9
```

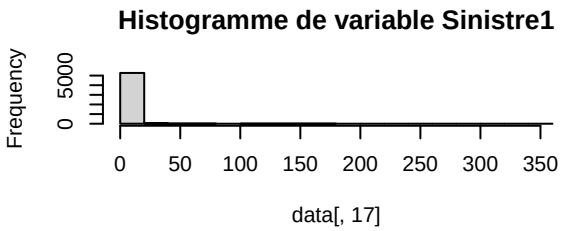
On a representé les boxplots de la variable RUC, et de son log. La transformation log rend la distribution simétrique.

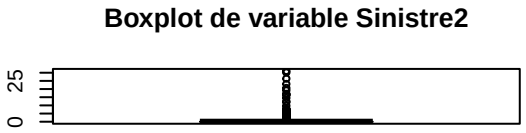


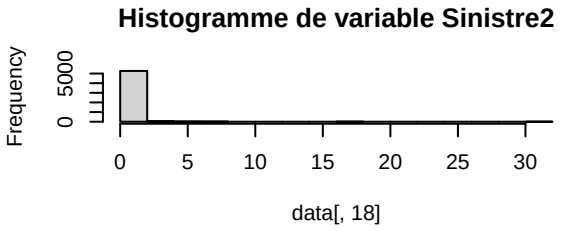


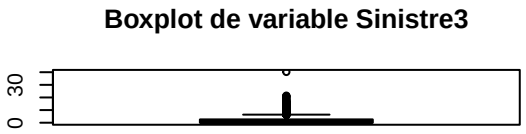

```
par(mfrow=c(3,2))
boxplot(data[,17],main="Boxplot de variable Sinistre1")
hist(data[,17],main="Histogramme de variable Sinistre1")
boxplot(data[,18],main="Boxplot de variable Sinistre2")
hist(data[,18],main="Histogramme de variable Sinistre2")
boxplot(data[,19],main="Boxplot de variable Sinistre3")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

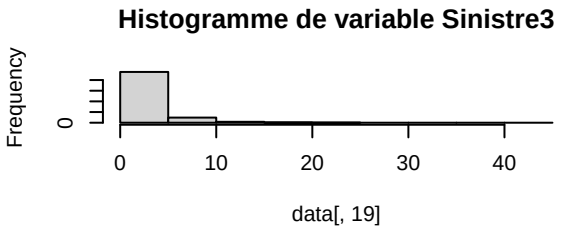












2.2 Analyses bivariées

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)
```

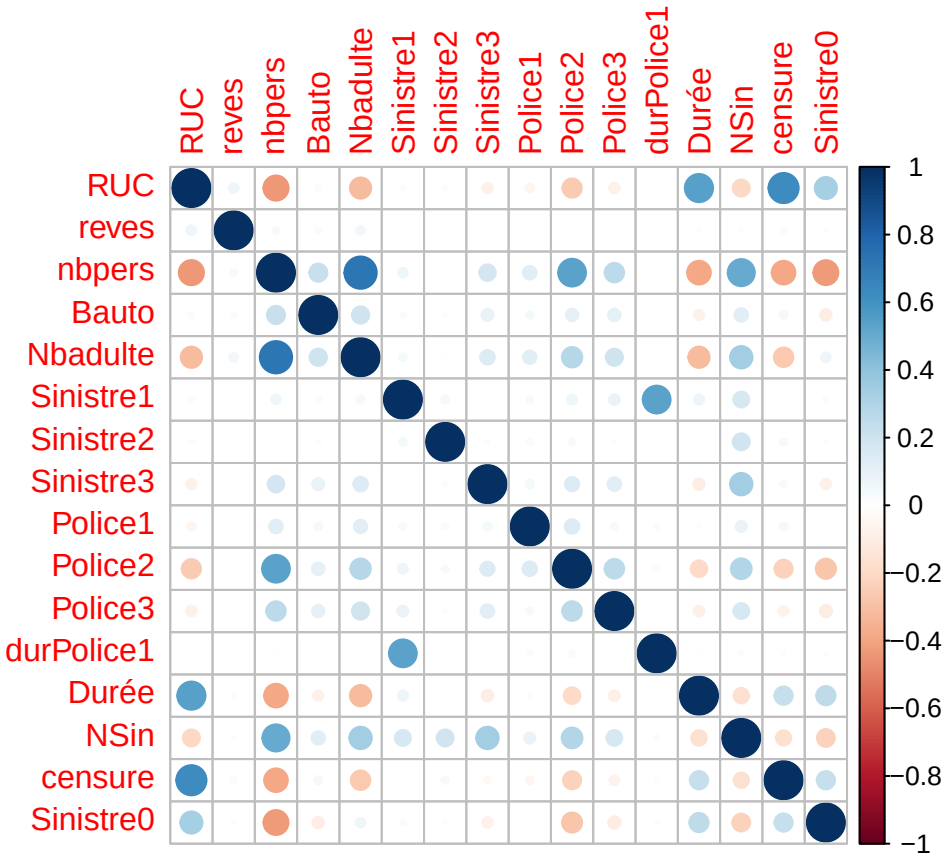
```
##                               Group.1 Sinistre1   Sinistre2 Sinistre3
## 1                     Agr. exploitants 0.4860776 0.009051724  2.983017
## 2        Artisans, comm., chefs d'ent. 0.3559116 0.097458564  1.836381
## 3     Autres pers. sans activite prof. 1.4143169 0.107540984  1.761721
## 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
```

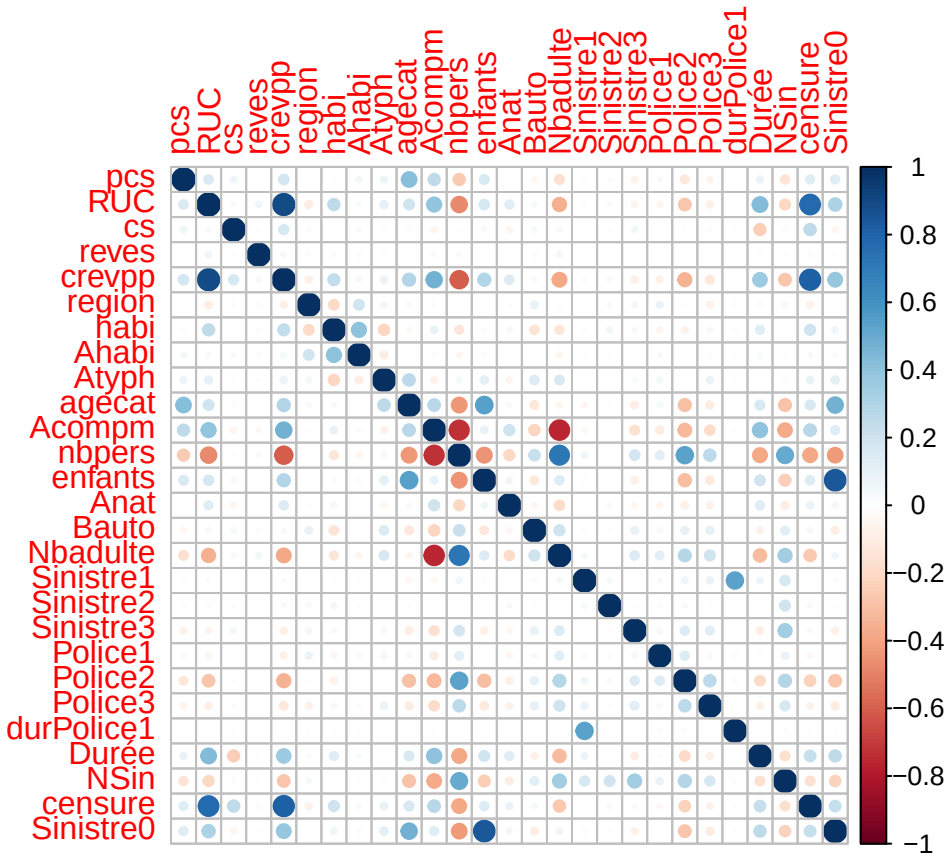
```
#Representation des correlations
variables_quantitatives = data %>% select_if(is.numeric) %>% cor()
```

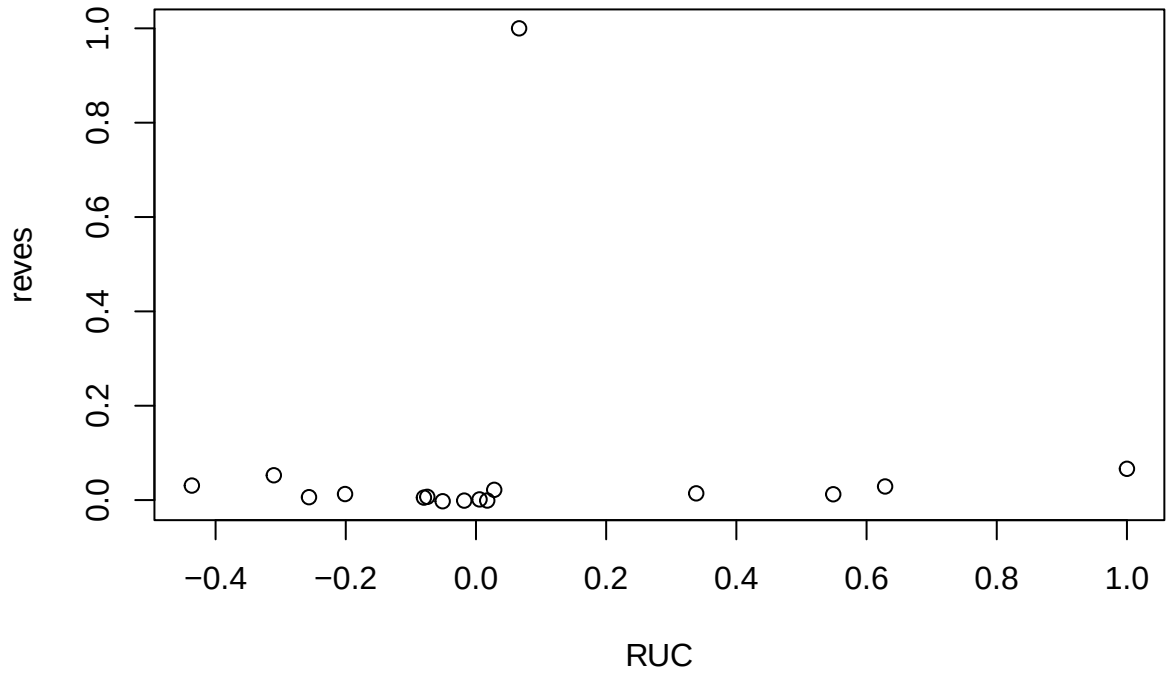
```
kable(variables_quantitatives, digits=3)
```

	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

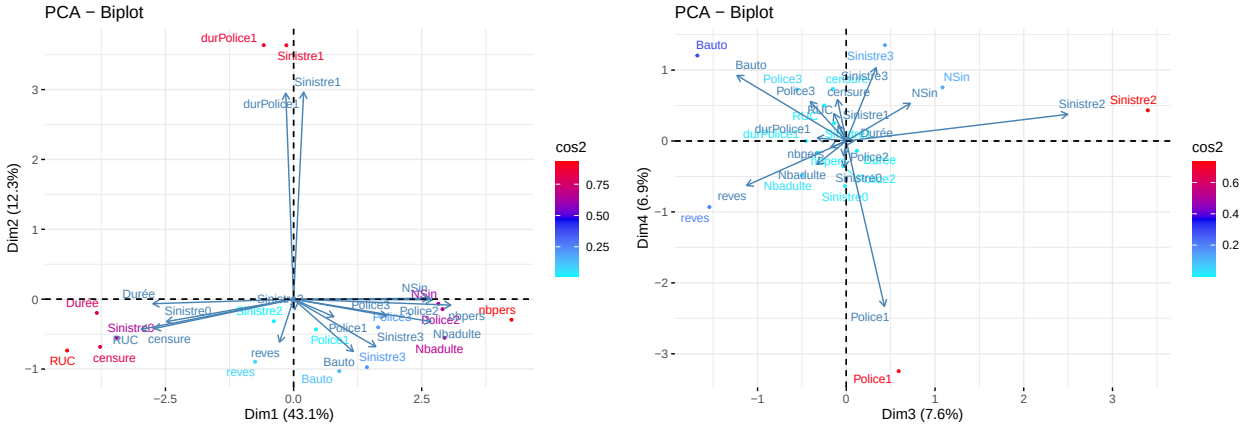

pcs
RUC
cs
reves
crevpp
region
habi
Ahabi
Atyph
agecat
Acompm
nbpers
enfants
Anat
Bauto
Nbadulte
Sinistre1
Sinistre2
Sinistre3
Police1
Police2
Police3
durPolice1
Durée
NSin
censure
Sinistre0
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1




```
## **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
```

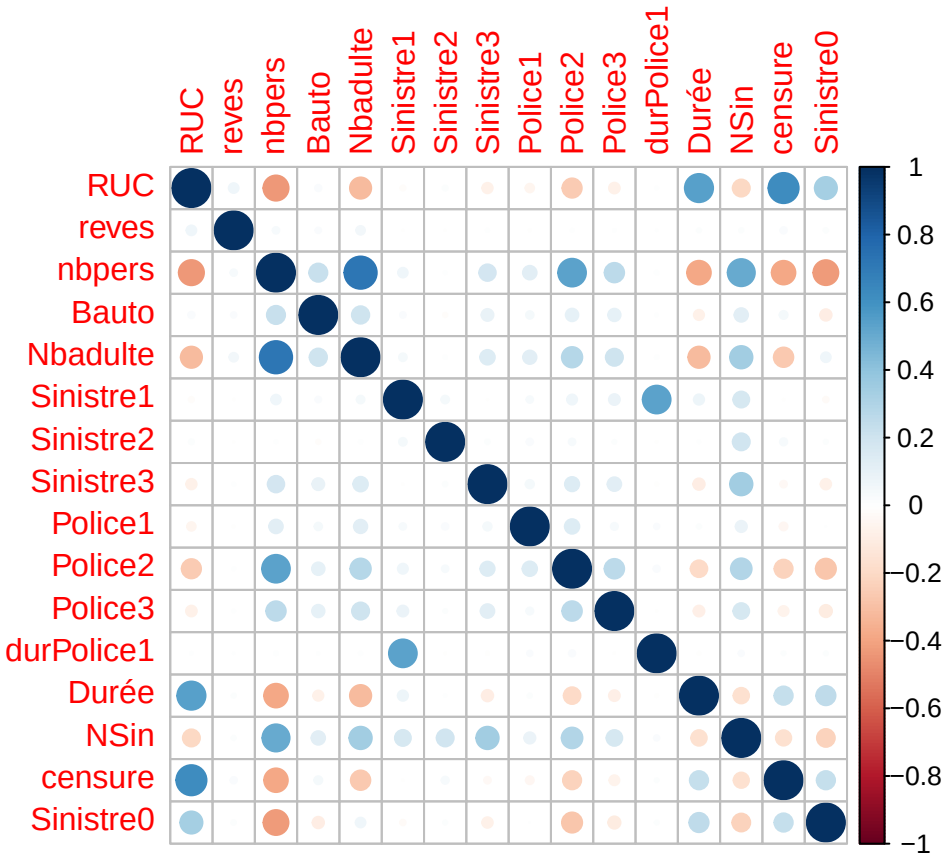


3. Modélisation des sinistres et des primes pures

3.1 Problème d'endogénéïté dans les variables

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

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



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.

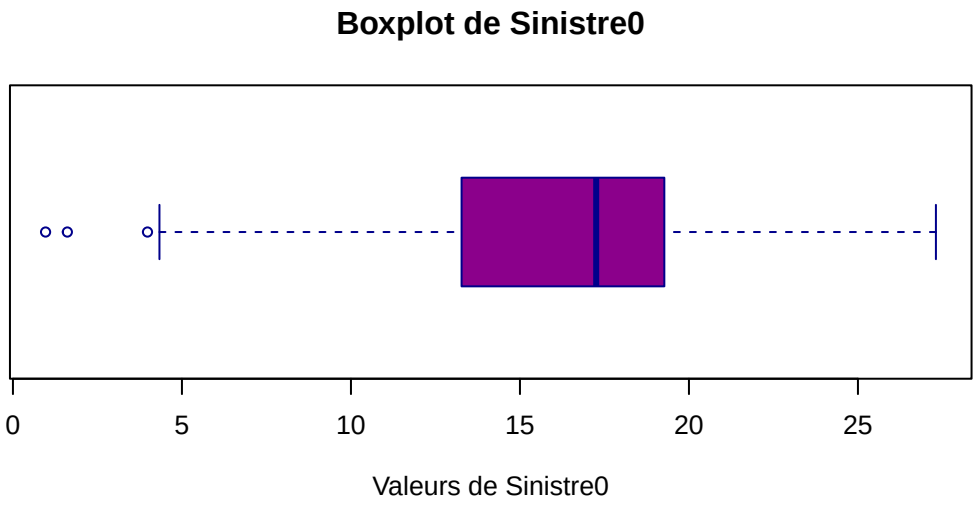
```
summary(selectionAIC)
```

```
##
## Call:
## lm(formula = Sinistre0 ~ RUC + cs + Acompm + Anat + Police2 +
##     Durée + NSin, data = data2[, vars])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5061 -1.4372  0.0079  1.4929  7.7472
##
## Coefficients:
##                                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                    7.833e+00  1.402e+00    5.589 2.39e-08 ***
## RUC                            1.270e+00  1.476e-01    8.602  < 2e-16 ***
## csModeste                     -6.831e-02  2.685e-01   -0.254  0.79919
## csMoyenne Inf                 -3.793e-01  1.832e-01   -2.070  0.03853 *
## csMoyenne Sup                 -3.766e-01  1.326e-01   -2.841  0.00452 **
## AcompmCouple avec enfant(s) -8.060e+00  7.827e-02 -102.973  < 2e-16 ***
## AcompmCouple sans enfant       5.066e-03  8.646e-02    0.059  0.95328
## AcompmPersonne seule          -8.565e-02  1.145e-01   -0.748  0.45437
```

```
## AnatMenage francais         -3.722e-01  2.039e-01   -1.826  0.06797 .
## AnatNon declare             -4.816e-01  2.343e-01   -2.055  0.03992 *
## Police2                      4.063e-03  2.507e-03    1.621  0.10518
## Durée                        1.769e-04  5.851e-05    3.024  0.00251 **
## NSin                         1.527e-02  8.675e-03    1.761  0.07837 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.157 on 5339 degrees of freedom
## Multiple R-squared:  0.7484, Adjusted R-squared:  0.7479
## F-statistic:  1324 on 12 and 5339 DF,  p-value: < 2.2e-16
```

On retient dont un modèle linéaire où Sinistre0 est expliqué par RUC, Acomp, cs, Durée

3.2 Modélisation de *Sinistre0*

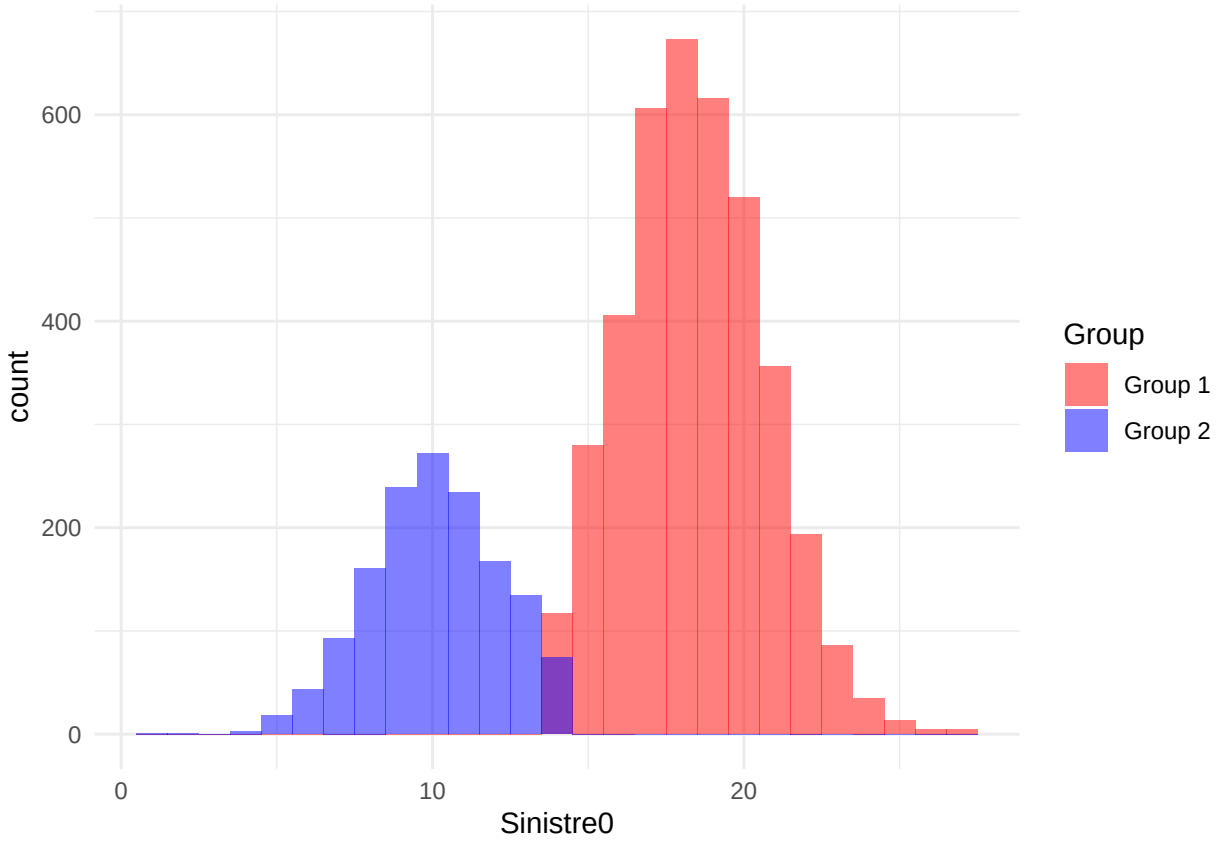


Le boxplot indique seulement 3 valeurs extrêmes dans la partie inférieure des valeurs. Nous les enlevons car ils peuvent influencer

- les paramètres de la regression (en “tirant” le paramètres de la ligne de régression vers eux),
- les residus - en augmentant la variance résiduelle et en rendant la distribution des résidus non normale
- la sensibilité de certains modèles, dont ceux de regression linéaire.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.334  13.290  17.262  16.181  19.273  27.307
```

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.

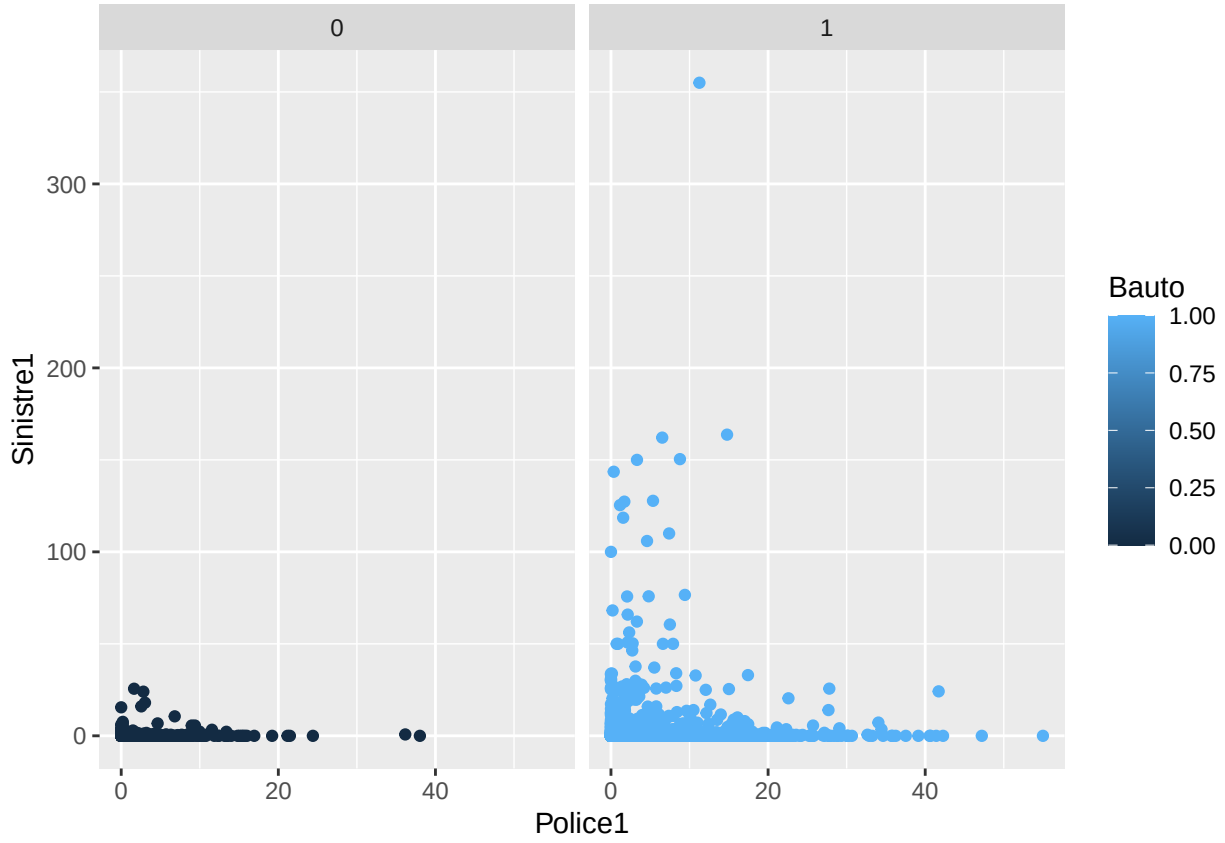
GLM: on peut essayer de modéliser. On rajoute une nouvelle var, delta: le fait que le sinistre1 soit >0. Cad j'ai une sinistre, vs. j'ai pas de sinistre. Donc j'ai une nouvelle var, que je vais modéliser par un probit: modèle lin gen, var. delta expliquée par : cs, anat, type... Fam Binomiale, avec modèle 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 Modéle pour au moins un des *Sinistre1*, *Sinistre2* ou *Sinistre3*



Ce graphique montre qu'on a plus de sinistres indemnisés parmi les souscripteurs de Police1 qui possèdent une ou plusieurs voitures.

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

3.6 Modèle retenu au final

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

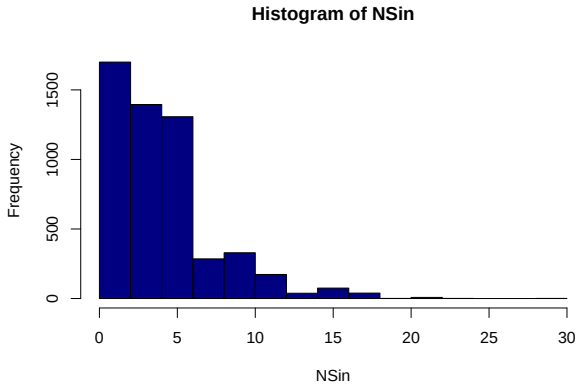
IV regressions

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