Université Paris Nanterre

Modélisation de Sinistres et de Montant de Police d'Assurance sur Données Réelles.

TA THI HAI CHAU	**40000165**
TOURÉ SOKHNA MOUMY	**40001675**
ZHOU LIJIA	**40001075**

L3 MIASHS

2022-2023

SOMMAIRE

- Présentation du projet
- Partie 1 : Description des données / Tests d'hypothèse
- Partie 2 : Modèle linéaire simple
- Partie 3 : Modèle linéaire multiple
- Partie 4 : Modélisation des zéros dans une régression
- Conclusion

Présentation du projet

Ce projet analyse un portefeuille d'assurés collecté par une compagnie d'assurance. Les données comprennent des informations personnelles des ménages assurés, les montants des dépenses et des sinistres. L'objectif est d'étudier les variables influençant les montants des sinistres, notamment pour le type de sinistre 1 (Sinistre1) et la police d'assurance associée (Police1). Le modèle sera calibré en utilisant toutes les variables pertinentes, avec des tests statistiques pour sélectionner les plus significatives. Les résidus seront évalués graphiquement pour vérifier la normalité et les points aberrants. Les individus avec des valeurs aberrantes seront exclus pour estimer le modèle de Police1. Les prédictions des montants de Sinistre1 pour les individus exclus et les montants observés de Police1 seront présentées dans un tableau. Ce projet permet de comprendre les facteurs influençant les sinistres et de développer un modèle de tarification pour les primes d'assurance.

Partie 1 : Description des données / Tests d'hypothèse

Question 1

```
source("F:/Miashs/L3/S6/projet_en_3/assurance.R")
dat
attach(dat)
edit(dat)
```

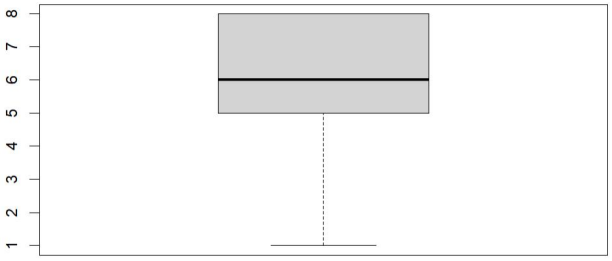
```
summary(reves)
mean(reves); mean(RUC)
sd(reves);sd(RUC)

str(dat)
summary(dat)
```

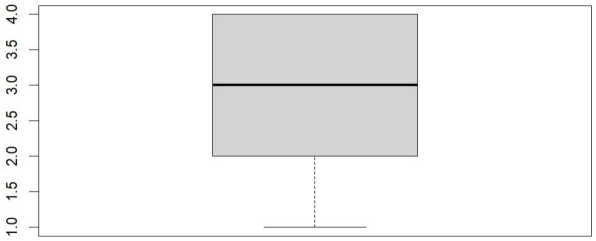
```
                                    pcs             RUC
Retraites                             :1372   Min.   :  277.8
Ouvriers                              :1327   1st Qu.: 3823.5
Professions intermediaires            : 887   Median : 5500.0
Employes                              : 806   Mean   : 6277.5
Cadres et prof. intellectuelles sup.: 480   3rd Qu.: 7812.5
Autres pers. sans activite prof.      : 183   Max.   :35294.1
(Other)                               : 297
          cs           reves                  crevpp
Aise       : 629   Min.   :   1000   1er quartile :1260
Modeste    : 901   1st Qu.:   8500   2eme quartile:1194
Moyenne Inf:2346   Median :  11250   3eme quartile:1403
Moyenne Sup:1476   Mean   :  14880   4eme quartile:1495
                   3rd Qu.:  16250
                   Max.   :3416250


   region              habi
Length:5352        Length:5352
Class :character   Class :character
Mode  :character   Mode  :character



                              Ahabi                Atyph
Communes rurales                 :1352   Locataire    :1963
Paris + Agglomeration            : 697   Non declare :  73
Un. urb. de 10 000 a 99 999 hab.:1031   Proprietaire:3316
Un. urb. de 100 000 hab. et +    :1623
Un. urb. de 2 000 a 9 999 hab.   : 649
```

```
  agecat                        Acompm         nbpers
21-40:1616   Autre menage         :1921   Min.   : 1.000
41-50:1483   Couple avec enfant(s):1379   1st Qu.: 2.000
51-60: 854   Couple sans enfant   :1293   Median : 3.000
61-96:1399   Personne seule       : 759   Mean   : 3.038
                                          3rd Qu.: 4.000
                                          Max.   :10.000

         enfants                  Bauto         Nbadulte
Au - 1 enfant:1379   Au - 1 vehicule:4909   Min.   :1.000
Pas d'enfants:3973   Pas de vehicule: 443   1st Qu.:2.000
                                            Median :2.000
                                            Mean   :2.388
                                            3rd Qu.:3.000
                                            Max.   :8.000

  Sinistre2         Sinistre3         Police1
Min.   : 0.0000   Min.   : 0.000   Min.   : 19.75
1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 31.49
Median : 0.0000   Median : 0.705   Median : 34.99
Mean   : 0.1615   Mean   : 1.837   Mean   : 35.46
3rd Qu.: 0.0000   3rd Qu.: 2.560   3rd Qu.: 38.88
Max.   :31.1000   Max.   :40.220   Max.   :231.43

   Police2           Police3           Duree
Min.   :  0.000   Min.   : 0.000   Min.   :   0.0
1st Qu.:  3.829   1st Qu.: 0.520   1st Qu.:  23.0
Median :  9.060   Median : 1.420   Median :  42.5
Mean   : 13.017   Mean   : 2.110   Mean   : 249.1
3rd Qu.: 17.950   3rd Qu.: 2.833   3rd Qu.: 233.0
Max.   :124.109   Max.   :34.743   Max.   :9877.0

     NSin           censure         Sinistre1
Min.   : 0.000   Min.   :0.0000   Min.   : 0.9652
1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:13.2769
Median : 4.000   Median :1.0000   Median :17.2558
Mean   : 4.249   Mean   :0.6179   Mean   :16.1732
3rd Qu.: 6.000   3rd Qu.:1.0000   3rd Qu.:19.2718
Max.   :30.000   Max.   :1.0000   Max.   :27.3074

      Aise     Modeste Moyenne Inf Moyenne Sup
       629         901        2346        1476
```

La variable pcs qui est une variable qualitative contient les catégories socio-professionnelles, il y en a 7 présentes et les deux plus représentées sont Retraités (1372 individus) et Ouvriers (1327)
La variable reves qui est une variable quantitative contient les revenus du foyer, en moyenne il y a 14800 euros et la valeur médiane est 11250 euros.
La variable RUC qui est une variable quantitative contient les revenu par unités de consommation, en moyenne il y a 6277,5 euros et la valeur médiane est 5500 euros.#

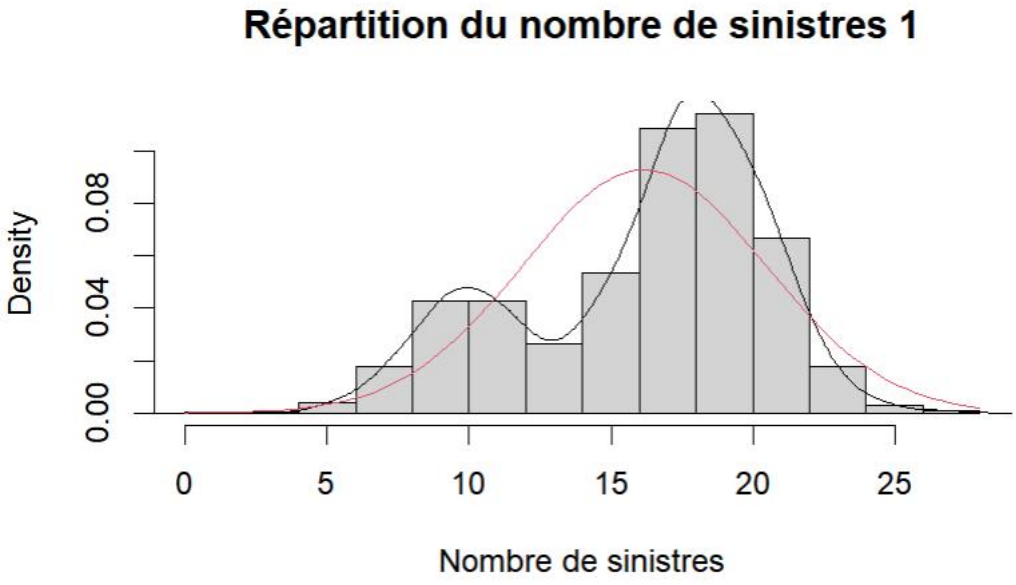
```
boxplot(pcs,col='lightgray')
boxplot(cs,col='lightgray')
```



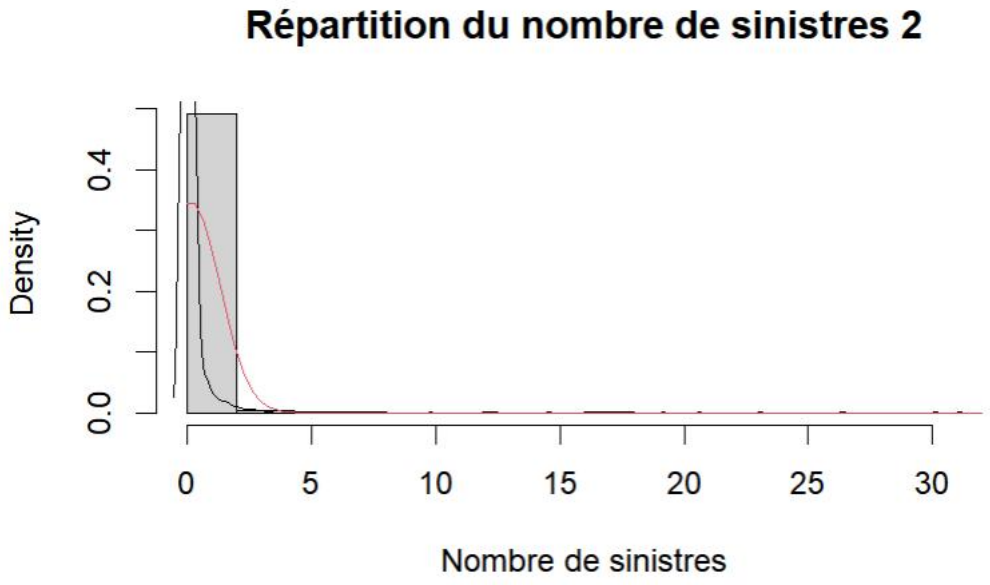


Question 2

```
plot(hist(dat$Sinistre1), freq=FALSE, xlab="Nombre de
sinistres",main="Répartition du nombre de sinistres 1")
lines(density(dat$Sinistre1))
curve(dnorm(x, mean=mean(dat$Sinistre1), sd=sd(dat$Sinistre1)),
add=TRUE, col=2)
```



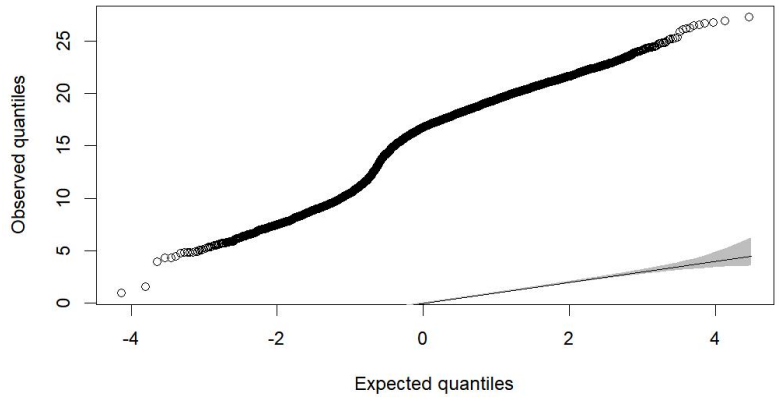
```
density.default(x = dat$Sinistre2, n = 50)
plot(hist(dat$Sinistre2), freq=FALSE, xlab="Nombre de
sinistres",main="Répartition du nombre de sinistres 2")
lines(density(dat$Sinistre2))
curve(dnorm(x, mean=mean(dat$Sinistre2), sd=sd(dat$Sinistre2)),
add=TRUE, col=2)
```



Maintenant, on va effectuer un test de Kolmogorov-Smirnov pour la normalité de sinistre1 et sinistre2:

```
x1 <- sort(Sinistre1)
esp <- mean(x1)
ecart <- sd(x1)
ks.test(x1,"pnorm", mean=esp, sd=ecart)

library(qqconf)
qq_conf_plot(x1,distribution=qnorm,dparams = list(mean=esp,sd=ecart))
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  x1
D = 0.11374, p-value < 2.2e-16
alternative hypothesis: two-sided
```



```
x2 <- sort(Sinistre2)
esp <- mean(x2)
ecart <- sd(x2)
ks.test(x2,"pnorm", mean=esp, sd=ecart)
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  x2
D = 0.45213, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Comme les p-valeurs de test1 et test2 sont tous trop petites, on peut conclure que le nombre de sinistre 1 et sinistre 2 ne suivent pas la loi normale.

Partie 2: Modèle linéaire simple :

Un étudiant décide d'expliquer le montant des sinistres de type 1 (Sin1) des ménages en fonction de la variable log(RUC) (log du revenu par unité de consommation du ménage).

Question 1

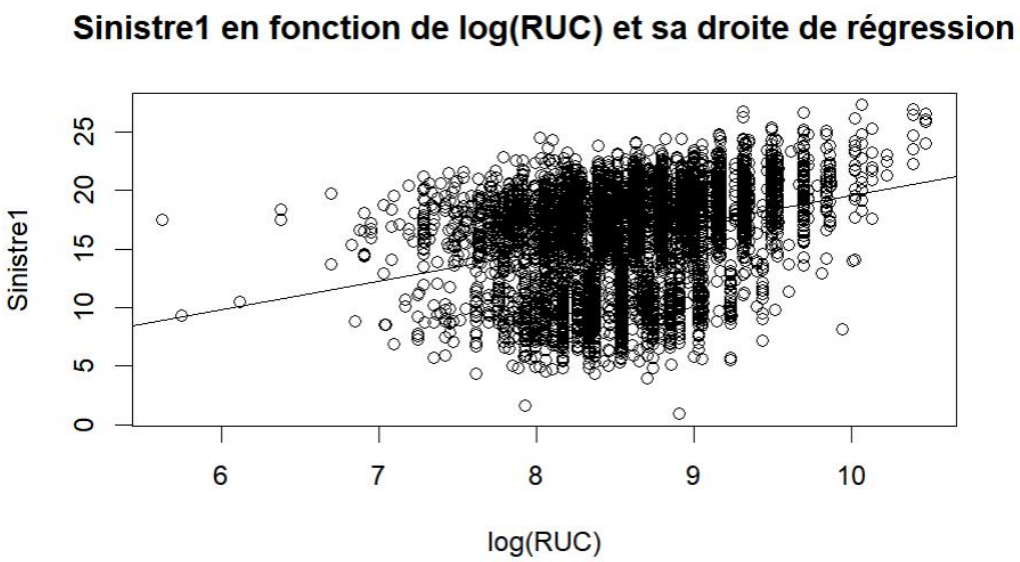
```
m1 <- lm(Sinistre1~log(RUC), data=dat)
summary(m1)
```

```
Call:
lm(formula = Sinistre1 ~ log(RUC), data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.9678  -2.7138   0.9247   3.0136   9.7131 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -4.9077     0.8847  -5.547 3.04e-08 ***
log(RUC)      2.4520     0.1027  23.876  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.084 on 5350 degrees of freedom
Multiple R-squared:  0.09629,	Adjusted R-squared:  0.09613 
F-statistic: 570.1 on 1 and 5350 DF,  p-value: < 2.2e-16
```

On remarque que la p-valeur de variable log(RUC) dans ce modèle est égale à 2e-16 et donc inférieur à 5%. La variable log(RUC) est donc significative avec un niveau d'erreur de 5%. Le log du revenu par unité de consommation des ménages est alors pertinent pour expliquer la variation du montant du sinistre 1.

Question 2

```
plot(Sinistre1~log(RUC),data = dat, main="Sinistre1 en fonction de log(RUC)
et sa droite de régression")
abline(m1)
```



Question 3

```
m2 <- lm(Sinistre1~log(RUC)+pcs)
summary(m2)
```

```
                                        Pr(>|t|)
(Intercept)                              0.26137
log(RUC)                                 < 2e-16 ***
pcsArtisans, comm., chefs d'ent.         0.76755
pcsAutres pers. sans activite prof.      0.00102 **
pcsCadres et prof. intellectuelles sup.  0.18873
pcsEmployes                              0.05811 .
pcsOuvriers                             3.19e-05 ***
pcsProfessions intermediaires            0.00927 **
pcsRetraites                            4.28e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.897 on 5343 degrees of freedom
Multiple R-squared:  0.1783,    Adjusted R-squared:  0.1771
F-statistic:   145 on 8 and 5343 DF,  p-value: < 2.2e-16
```

La variable log(RUC) est donc toujours significative. Le revenu et la catégorie socioprofessionnelle de l'individu sont alors toutes deux pertinentes pour expliquer la variabilité du montant du sinistre 1.#

Question 4

```
m1 <- lm(Sinistre1~log(RUC), data=dat)
summary(m1)
m2 <- lm(Sinistre1~log(RUC)+pcs)
summary(m2)
m3 <- lm(Sinistre1~ pcs)
summary(m3)
AIC(m1)
AIC(m2)
AIC(m3)
```

```
[1] 30253.58
[1] 29758.18
[1] 30088.58
```

On peut utiliser la fonction AIC pour comparer quel modèle est le plus meuilleur. D'après l'affichage du code, on considère le modèle avec log (RUC) et pcs comme variables explicatives car le modèle (avec constante) ayant l'AIC le plus faible est celui contenant les variables pcs et log(RUC) à la fois.

Partie 3: Modèle linéaire multiple :

Nous allons commencer par exclure les 10 derniers individus de la base de données dat et les mettre dans une base à part appelée dat0.

Question 1

```
edit(dat)
dat0 <- dat[1:(nrow(dat)-10), ]  # Exclut les 10 dernières lignes de dat
dat1 <- dat[(nrow(dat)-9):nrow(dat), ]  # Sélectionne uniquement les 10 dernières lignes de dat

library(MASS)
initial_model <- lm(Sinistre1 ~ ., data = dat0)
final_model1 <- stepAIC(initial_model, direction = "both")
summary(final_model1)
```

```
Step:  AIC=6416.44
Sinistre1 ~ RUC + agecat + Acompm + nbpers + Bauto + Police1 +
    NSin

Call:
lm(formula = Sinistre1 ~ RUC + agecat + Acompm + nbpers + Bauto +
    Police1 + NSin, data = dat0)

Residuals:
    Min      1Q  Median      3Q     Max
-53.562  -1.046   0.000   1.123   5.889

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.110e+01  1.952e-01  56.867  < 2e-16 ***
RUC                         -1.175e-04  1.053e-05 -11.158  < 2e-16 ***
agecat41-50                  2.459e-01  7.519e-02   3.271  0.00108 **
agecat51-60                  6.369e-02  8.827e-02   0.722  0.47059
agecat61-96                 -2.077e-01  8.352e-02  -2.487  0.01291 *
AcompmCouple avec enfant(s) -5.870e+00  9.113e-02 -64.416  < 2e-16 ***
AcompmCouple sans enfant     7.399e-02  9.046e-02   0.818  0.41346
AcompmPersonne seule        -6.100e-02  1.199e-01  -0.509  0.61089
nbpers                      -8.030e-01  3.825e-02 -20.993  < 2e-16 ***
BautoPas de vehicule         2.283e-01  9.738e-02   2.345  0.01907 *
Police1                      2.724e-01  5.990e-03  45.482  < 2e-16 ***

NSin                         1.134e-02  7.616e-03   1.489  0.13658
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.821 on 5330 degrees of freedom
Multiple R-squared:  0.8206,    Adjusted R-squared:  0.8202
F-statistic:  2217 on 11 and 5330 DF,  p-value: < 2.2e-16
```

Puis on construit le modèle initial en incluant toutes les variables disponibles en tant que régresseurs. Ensuite, nous utilisons la fonction StepAIC en éliminant les variables non significatives du modèle précédent pour trouver AIC le plus faible. Ce processus est répété jusqu'à ce que seules les variables significatives restent dans le modèle. D’après l’affichage de summary(final_model1), on constate que par rapport de Sinistre1, les variables explicatives suivantes sont retenues : RUC, agecat, Acompm, nbpers, Bauto, Police1 et NSin.

```r
library(MASS)
initial_model <- lm(Police1 ~ ., data = dat0)
final_model2 <- stepAIC(initial_model, direction = "both")
summary(final_model2)
```

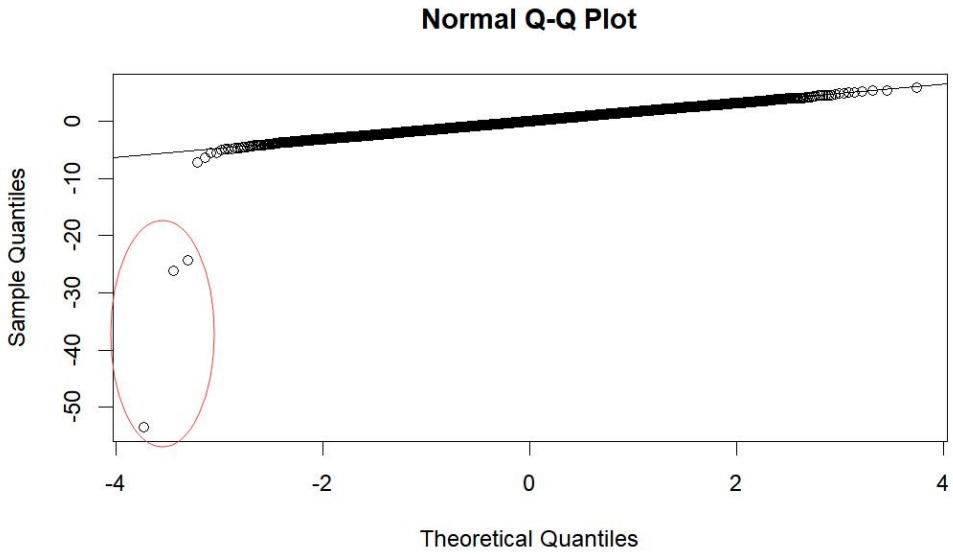
```
Step:  AIC=13493.33
Police1 ~ RUC + agecat + nbpers + Bauto + censure + Sinistre1 +
    Ahabi

Call:
lm(formula = Police1 ~ RUC + agecat + nbpers + Bauto + censure +
    Sinistre1 + Ahabi, data = dat0)

Residuals:
    Min      1Q  Median      3Q     Max
 -4.043  -0.827  -0.072   0.671 200.177

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            4.282e+00  3.205e-01  13.363  < 2e-16 ***
RUC                                    9.938e-04  1.825e-05  54.442  < 2e-16 ***
agecat41-50                           -9.218e-01  1.379e-01  -6.683 2.59e-11 ***
agecat51-60                           -4.237e-01  1.632e-01  -2.596 0.009456 **
agecat61-96                            5.777e-01  1.528e-01   3.780 0.000159 ***
nbpers                                 2.963e+00  4.643e-02  63.826  < 2e-16 ***
BautoPas de vehicule                  -9.173e-01  1.876e-01  -4.890 1.04e-06 ***
censure                               -2.153e-01  1.305e-01  -1.650 0.099025 .
Sinistre1                              1.003e+00  1.424e-02  70.408  < 2e-16 ***
AhabiParis + Agglomeration             6.159e-01  1.764e-01   3.493 0.000482 ***
AhabiUn. urb. de 10 000 a 99 999 hab.  5.570e-02  1.474e-01   0.378 0.705534
AhabiUn. urb. de 100 000 hab. et +     3.393e-02  1.325e-01   0.256 0.797981
AhabiUn. urb. de 2 000 a 9 999 hab.    5.229e-02  1.688e-01   0.310 0.756796
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.532 on 5329 degrees of freedom
Multiple R-squared:  0.7135,    Adjusted R-squared:  0.7128
F-statistic:  1106 on 12 and 5329 DF,  p-value: < 2.2e-16
```

Le même raisonnement que Sinistres de type1. D’après l’affichage de summary(final_model2), on constate que par rapport de Police1, les variables explicatives suivantes sont retenues : RUC, agecat, nbpers, Bauto, censure, Sinistre1 et Ahabi.

Question 2

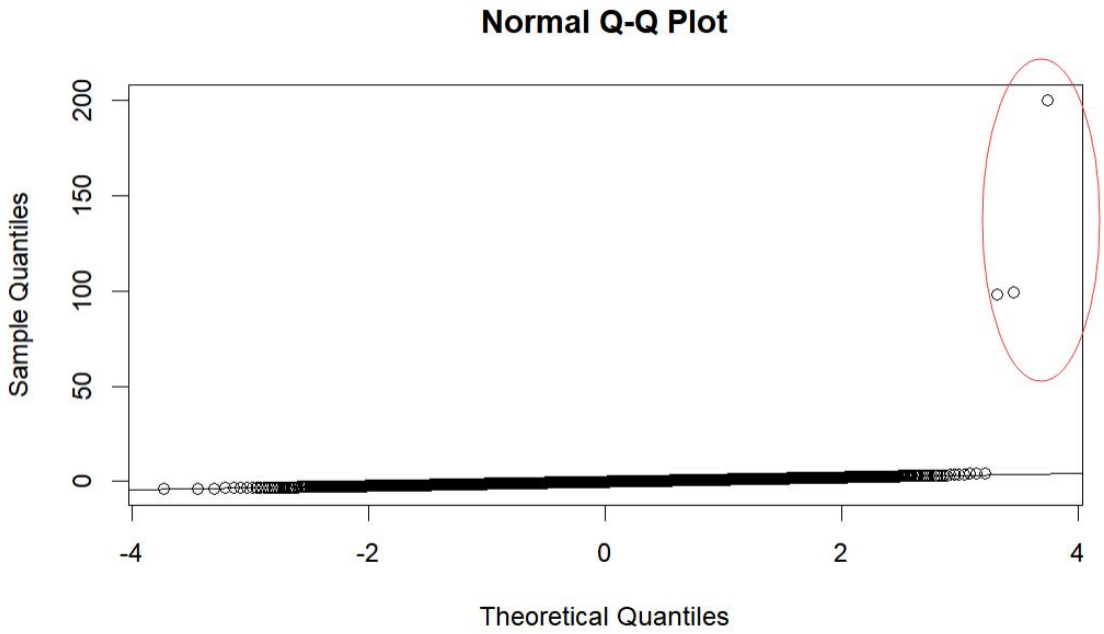
Tests graphiques:

```r
residuals <- residuals(final_model1)
qqnorm(residuals)
qqline(residuals)
```



Le graphique QQ-plot peut nous aider à évaluer visuellement la normalité des résidus. Si les résidus semblent dévier considérablement d'une distribution normale sur le graphique, cela peut indiquer une violation de l'assomption de normalité dans votre modèle linéaire.
Donc les résidus de final_modèle1 semblent être linéairement distribués, ce qui suggère une normalité. Cependant, il existe aussi quelques points abérrants (ce qui se situent très loin par rapport de la ligne droite). #

```
residuals <- residuals(final_model2)
qqnorm(residuals)
qqline(residuals)
```



Donc les résidus de final_modèle2 semblent être linéairement distribués, ce qui suggère une normalité. Cependant, il existe aussi quelques points abérrants (ce qui se situent très loin par rapport de la ligne droite).

Question 3

```
final_model2 <- lm(Police1 ~ RUC + agecat + nbpers + Bauto + censure + Sinistre1 + Ahabi,
data = dat0)
student_resid <- rstudent(final_model2)
outliers <- which(abs(student_resid) > 2)  # Sélection des observations avec des résidus
studentisés supérieurs à 2
clean_data <- dat0[-outliers, ]
clean_model <- lm(Police1 ~ RUC + agecat + nbpers + Bauto + censure + Sinistre1 + Ahabi,
data = clean_data)
summary(clean_model)
```

```
Call:
lm(formula = Police1 ~ RUC + agecat + nbpers + Bauto + censure +
    Sinistre1 + Ahabi, data = clean_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0901 -0.7235 -0.0077  0.7226  4.4601

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            4.210e+00  9.955e-02  42.291   <2e-16 ***
RUC                                    1.001e-03  5.673e-06 176.517   <2e-16 ***
agecat41-50                           -9.905e-01  4.286e-02 -23.110   <2e-16 ***
agecat51-60                           -5.137e-01  5.072e-02 -10.129   <2e-16 ***
agecat61-96                            5.247e-01  4.749e-02  11.049   <2e-16 ***
nbpers                                 3.003e+00  1.442e-02 208.188   <2e-16 ***
BautoPas de vehicule                  -1.383e+00  5.835e-02 -23.707   <2e-16 ***
censure                               -6.747e-03  4.055e-02  -0.166    0.868
Sinistre1                              9.940e-01  4.424e-03 224.674   <2e-16 ***
AhabiParis + Agglomeration             4.599e-02  5.485e-02   0.838    0.402
AhabiUn. urb. de 10 000 a 99 999 hab.  5.314e-02  4.579e-02   1.161    0.246
AhabiUn. urb. de 100 000 hab. et +     3.067e-02  4.117e-02   0.745    0.456
AhabiUn. urb. de 2 000 a 9 999 hab.    5.398e-02  5.245e-02   1.029    0.303
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.097 on 5326 degrees of freedom
Multiple R-squared:  0.9628,    Adjusted R-squared:  0.9627
F-statistic: 1.149e+04 on 12 and 5326 DF,  p-value: < 2.2e-16
```

Pour enlever des points aberrants dans ce modèle linéaire, on peut utiliser l’approche "méthode des résidus studentisés"

Question 4

On donnera les prédictions des montants des Sinistre1 pour les individus:

```
pred <- predict(final_model1, dat1)
obs <- dat1$Sinistre1
cbind(pred, obs)
```

```
          pred      obs
5343 19.274676 22.76060
5344 16.788985 16.06724
5345 18.927851 21.20237
5346 18.902990 21.57460
5347 19.669826 21.96043
5348 17.124279 15.03810
5349 17.426731 18.63927
5350 21.005666 17.59627
5351  9.300468 10.30899
5352 19.959896 20.99097
```

On donnera les prédictions des montant de Police1 pour les individus:

```
pred <- predict(final_model2, dat1)
obs <- dat1$Police1
cbind(pred, obs)
```

```
         pred      obs
5343 37.99081 36.76939
5344 35.46025 34.00404
5345 38.40452 37.97438
5346 41.98332 42.55233
5347 43.34903 44.09972
5348 32.25823 31.35385
5349 34.09855 32.34460
5350 55.07420 55.75561
5351 32.06125 30.04415
5352 44.52235 45.23877
```

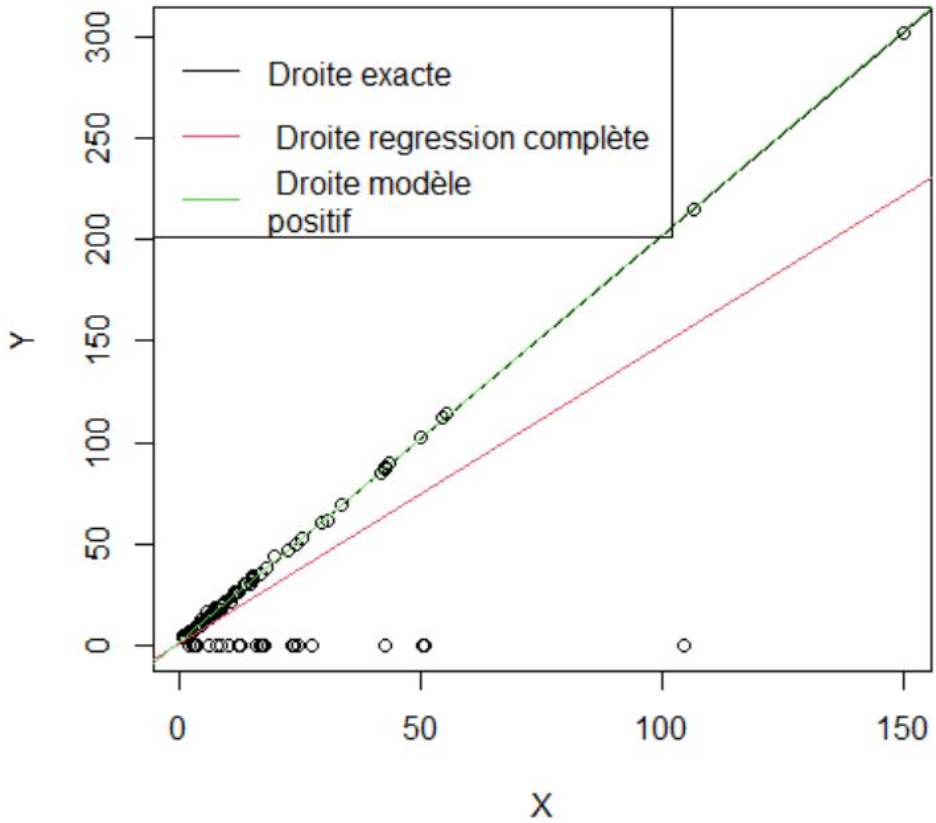
#On peut constater que les différences entre les prédictions et les observations ne sont pas grandes, donc on peut conclure qu'on a bien choisi les deux modèles finials pour les montant Sinistre1 et les montant de Police1. #

Partie 4: Modélisation des zéros dans une régression

Les variable Sinistre2 et Sinistre3 comportent de nombreuses valeurs avec des zéros, c'est à dire des individus n'ayant pas eu de sinistres de type 2 ou 3. Cette partie a pour but de montrer que inclure les zéros dans la procédure des m.c.o. ou les éliminer peut créer de nombreux biais et de proposer une méthode d'estimation adéquate.

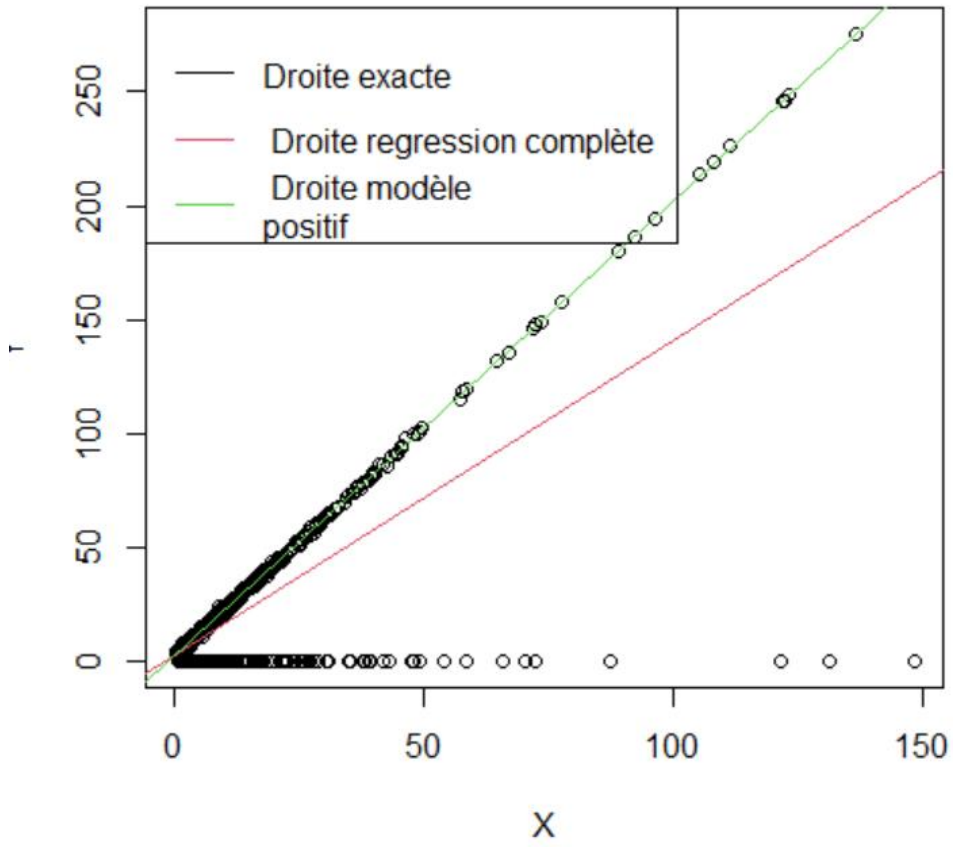
Case n=100

```
a=-0.5; b=1; alp=2; beta=2; n=100
X = rlnorm(n,2,1)
Z = rlnorm(n)
eps = rnorm(n)
delta = as.numeric(a+b*Z+eps > 0)
Y = delta * (alp+beta*X+eps)
Data1 = as.data.frame(cbind(eps = eps, Z = Z, X = X, delta = delta, Y = Y))
tail(Data1)
plot(Y ~ X, data=Data1)
#Droite y=alp+beta*x
abline(alp, beta)
#Modeliser Y Modele complet
abline(lm(Y ~ X, data= Data1), col=2)
#Modeliser Y quand Y>0
abline(lm(Y ~ X, data= Data1[Data1$Y>0,]),col=3)
legend("topleft", legend=c("Droite exacte", " Droite regression complète ", "
Droite modèle
positif "), col=c(1:3), lty=1)
```



Case n=1000

```
a=-0.5; b=1; alp= 2; beta= 2; n=1000
X = rlnorm(n,2,1)
Z = rlnorm(n)
eps = rnorm(n)
delta = as.numeric(a+b*Z+eps > 0)
Y = delta * (alp+beta*X+eps)
Data2 = as.data.frame(cbind(eps = eps, Z = Z, X = X, delta = delta, Y = Y))
tail(Data2)
plot(Y ~ X, data=Data2)
#Droite y=alp+beta*x
abline(alp, beta)
#Modeliser Y Modele complet
abline(lm(Y ~ X, data= Data2), col=2)

#Modeliser Y quand Y>0
abline(lm(Y ~ X, data= Data2[Data2$Y>0,]),col=3)
legend("topleft", legend=c("Droite exacte", " Droite regression complète ", "
Droite modèle
positif "), col=c(1:3) , lty=1)
```

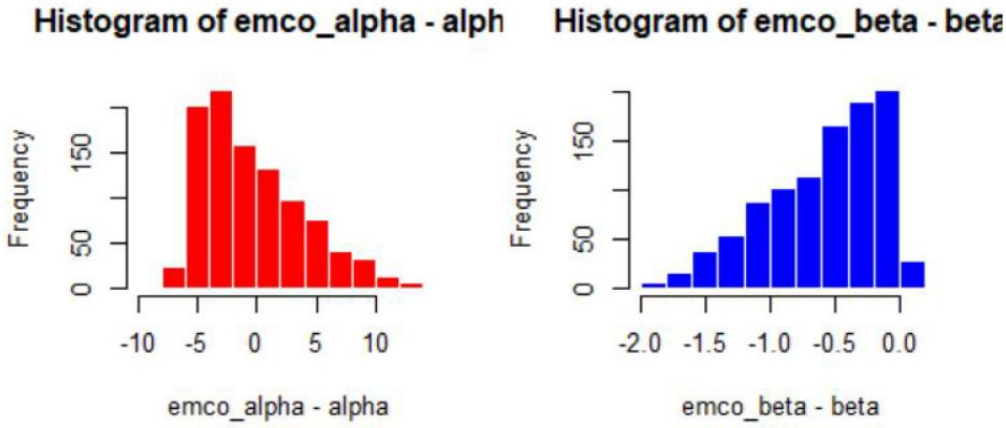


le modèle de régression positive de Y sur X fournit des estimations qui convergent vers les valeurs réelles de alpha et bêta. Les lignes de régression obtenues dans ce cas (verte et noire) sont pratiquement indiscernables. Les estimations obtenues à l'aide des Moindres Carrés Ordinaires (MCO) diffèrent significativement de celles obtenues en excluant les valeurs nulles de Y.Le biais est d'autant plus important lorsque la taille de l'échantillon, représentée par n, est grande.

Question 3

```
f4= function(N=999)
{ a=-0.5; b=1; alpha=2; beta=2; n=100
emco_alpha= rep(0,999)
emco_beta= rep(0,999)
set.seed(2400)
for(i in 1: 999){
```

```
X=exp(rnorm(n, mean=2, sd=1))
Z = rlnorm(n)
eps = rnorm(n)
delta = as.numeric(a+b*Z+eps > 0)
Y = delta * (alpha+beta*X+eps)
model4=lm(Y~X)
emco_alpha[i]= model4$coefficient[1]
emco_beta[i]= model4$coefficient[2]
}
par(mfrow=c(2,2))
hist(emco_alpha-alpha, col="red", border="white")
hist(emco_beta-beta, col=" blue", border="white")
}
f4()
```



Question 4

```
Res41=glm(delta~Z, data=Data1,family=binomial(link="probit"))
summary(Res41)
Res41$coef
```

```
Call:
glm(formula = delta ~ Z, family = binomial(link = "probit"),
    data = Data1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.12581  -0.81167   0.07818   0.77779   1.60661

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0140     0.3470  -2.922  0.00348 **
Z             2.1006     0.5133   4.093 4.26e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118.591  on 99  degrees of freedom
Residual deviance:  79.664  on 98  degrees of freedom
AIC: 83.664

Number of Fisher Scoring iterations: 8
```

```
> Res41=glm(delta~Z, data=Data1,family=binomial(link="probit"))
Message d'avis :
glm.fit: fitted probabilities numerically 0 or 1 occurred
> Res41$coef
(Intercept)           Z
  -1.014021    2.100581
```

Res42=glm(delta~Z, data=Data2,family=binomial(link="probit"))
summary(Res42)
Res42$coef

```
Call:
glm(formula = delta ~ Z, family = binomial(link = "probit"),
    data = Data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7398  -0.9731   0.2291   0.8635   1.5475

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.58670    0.08448  -6.945 3.79e-12 ***
Z            1.03987    0.08552  12.160  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1241.38  on 999  degrees of freedom
Residual deviance:  941.89  on 998  degrees of freedom
AIC: 945.89

Number of Fisher Scoring iterations: 7
```

```
> Res42=glm(delta~Z, data=Data2,family=binomial(link="probit"))
Message d'avis :
glm.fit: fitted probabilities numerically 0 or 1 occurred
> Res42$coef
(Intercept)           Z
 -0.5867022   1.0398676
```

#Lorsque la taille de l'échantillon augmente de plus en plus (de manière asymptotique), les estimateurs convergent vers les vraies valeurs des paramètres a et b. #

Question 5

Etape1 :Modeliser de la variable delta en utilisant le model probit

mprobit = glm(delta ~ Z, family=binomial(link="probit"))

Etape2 : Modeliser de la variable Y sachant delta=1 par mco

W=dnorm(mprobit$coef[2]*Z+mprobit$coef[1])/pnorm(mprobit$coef[2]*Z+mprobit$coef[1])
mlinear =lm(Y[delta==1] ~X[delta==1] +W[delta==1])
summary(mlinear)

Case n=100

```
> table(delta)
delta
 0  1
37 63
> mprobit = glm(delta ~ Z, family=binomial(link="probit"))
Message d'avis :
glm.fit: fitted probabilities numerically 0 or 1 occurred
> W = dnorm(mprobit$coef[2]*Z+mprobit$coef[1])/pnorm(mprobit$coef[2]*Z+mprobit$coef[1])
> mlinear =lm(Y[delta==1] ~X[delta==1] +W[delta==1])
> summary(mlinear)

Call:
lm(formula = Y[delta == 1] ~ X[delta == 1] + W[delta == 1])

Residuals:
    Min      1Q  Median      3Q     Max
-1.7766 -0.5352 -0.2103  0.6407  2.1096

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.773619   0.224105   7.914 6.86e-11 ***
X[delta == 1] 2.006962   0.009825 204.278  < 2e-16 ***
W[delta == 1] 1.218122   0.359175   3.391  0.00124 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8145 on 60 degrees of freedom
Multiple R-squared:  0.9986,	Adjusted R-squared:  0.9985
F-statistic: 2.1e+04 on 2 and 60 DF,  p-value: < 2.2e-16
```

Case n=1000

```
Call:
lm(formula = Y[delta == 1] ~ X[delta == 1] + W[delta == 1])

Residuals:
     Min       1Q   Median       3Q      Max 
-2.48898 -0.53152 -0.07345  0.44169  2.41751 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.933284   0.046947   41.18   <2e-16 ***
X[delta == 1] 2.000188   0.001672 1196.43   <2e-16 ***
W[delta == 1] 1.093994   0.088001   12.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7542 on 723 degrees of freedom
Multiple R-squared:  0.9995,	Adjusted R-squared:  0.9995 
F-statistic: 7.164e+05 on 2 and 723 DF,  p-value: < 2.2e-16

> table(delta)
delta
  0   1 
274 726 
>
```

Question 6

attach(dat)
delta = as.numeric(dat$Sinistre3>0)
heckit1 <- selection(selection = delta ~ Atyph+ Ahabi+ Bauto+pcs , outcome = Sinistre3 ~Atyph + cs + Nbadulte + nbpers + NSin + pcs + region, data = dat, method = "2step")
summary(heckit1)

```
Outcome equation:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           3.695525   0.699574   5.283 1.32e-07 ***
AtyphNon declare                     -0.273385   0.442702  -0.618 0.536906    
AtyphProprietaire                     0.202589   0.112736   1.797 0.072389 .  
csModeste                             0.075041   0.228321   0.329 0.742420    
csMoyenne Inf                         0.505673   0.188522   2.682 0.007334 ** 
csMoyenne Sup                         0.357697   0.180400   1.983 0.047442 *  
Nbadulte                              0.127598   0.064846   1.968 0.049153 *  
nbpers                                0.223408   0.062299   3.586 0.000339 ***
NSin                                  0.008158   0.017843   0.457 0.647553    
pcsArtisans, comm., chefs d'ent.     -1.950589   0.449567  -4.339 1.46e-05 ***
pcsAutres pers. sans activite prof.  -1.244063   0.455789  -2.729 0.006364 ** 
pcsCadres et prof. intellectuelles sup. -1.505931   0.411212  -3.662 0.000253 ***
pcsEmployes                          -2.057813   0.388267  -5.300 1.20e-07 ***
pcsOuvriers                          -2.107805   0.373658  -5.641 1.78e-08 ***
pcsProfessions intermediaires        -1.606322   0.389467  -4.124 3.77e-05 ***
pcsRetraites                         -1.250946   0.399324  -3.133 0.001742 ** 
region2                               0.233700   0.188922   1.237 0.216136    
region3                              -0.480541   0.237379  -2.024 0.042983 *  
region4                               0.932211   0.206055   4.524 6.20e-06 ***
region5                              -0.178667   0.198237  -0.901 0.367480    
region7                               0.164253   0.216516   0.759 0.448114    
region8                               0.259004   0.196965   1.315 0.188575    
region9                               0.696163   0.210183   3.312 0.000932 ***
Multiple R-Squared:0.0612,	Adjusted R-Squared:0.0551 
   Error terms:
               Estimate Std. Error t value Pr(>|t|)
invMillsRatio  -1.9325     0.8840  -2.186   0.0288 *
sigma           3.1755         NA      NA       NA  
rho            -0.6086         NA      NA       NA  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La partie démonstration

D. Montrer que δ_i suit une loi Bernoulli $Ber(1, p_i)$ avec $p_i = \Phi(a + bz_i)$

Dans l'énoncé, on suppose que :

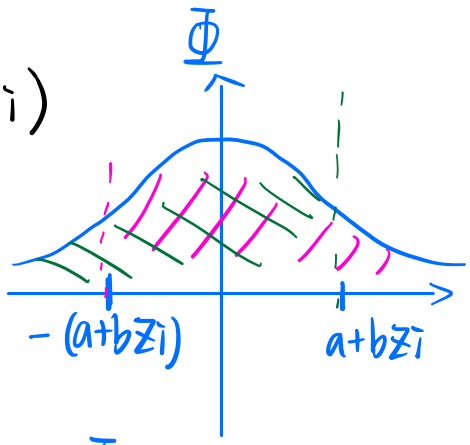
$$\delta_i = \begin{cases} 0, & \text{si } U_i = a + bz_i + \varepsilon_i < 0 \\ 1, & \text{si } U_i = a + bz_i + \varepsilon_i > 0 \end{cases} \quad \text{où } \varepsilon_i \sim N(0,1)$$

Donc $\mathbb{P}(\delta_i = 1) = \mathbb{P}(a + bz_i + \varepsilon_i > 0)$

$= \mathbb{P}(\varepsilon_i > -a - bz_i)$

$\varepsilon_i \sim N(0,1)$

$= \mathbb{P}(\varepsilon_i < a + bz_i)$



$= \Phi(a + bz_i)$, où Φ : la fonction de répartition $N(0,1)$

D'où le résultat, δ_i suit la loi Bernoulli $\left(1, p_i = \Phi(a + bz_i)\right)$

5). Montrer que si une variable aléatoire Z est $N(0,1)$, alors $\mathbb{E}[Z \mid Z > c] = \dfrac{\phi(-c)}{\Phi(-c)}$

$$\mathbb{P}(Z \leq z \mid Z > c) = \frac{\mathbb{P}(Z \leq z) \cdot \mathbb{P}(Z > c)}{\mathbb{P}(Z > c)}$$

$$= \frac{\mathbb{P}(c < Z \leq z)}{\mathbb{P}(Z > c)}$$

$$= \frac{\mathbb{P}(Z \leq z) - \mathbb{P}(Z < c)}{\mathbb{P}(Z > c)}$$

En dérivant $\mathbb{P}(Z \leq z \mid Z > c)$ par rapport à la variable z, on obtient :

ϕ : la densité de $N(0,1)$

$$f(Z \leq z \mid Z > c) = \begin{cases} \dfrac{\phi(z)}{1 - \Phi(c)} = \dfrac{\phi(z)}{\Phi(-c)}, & \text{si } z > c \\ 0, & \text{sinon} \end{cases}$$

$N(0,1)$ est centré en 0 (fonction paire)

Donc $\mathbb{E}(Z \mid Z > c) = \int_{-\infty}^{+\infty} z \cdot f(Z \leq z \mid Z > c)\, dz$

$$= \int_{c}^{+\infty} z \cdot \frac{\phi(z)}{\Phi(-c)}\, dz$$

$$= \frac{1}{\Phi(-c)} \int_{c}^{+\infty} z \cdot \phi(z)\, dz$$

$$= \frac{\phi(-c)}{\Phi(-c)}$$

- Maintenant, déterminons $\mathbb{E}[Z \mid Z > c]$, lorsque $Z \sim N(m, \sigma^2)$

Effectivement : $\frac{Z-m}{\sigma} \sim N(0,1)$

$$\mathbb{P}(Z \leq z \mid Z > c) = \frac{\mathbb{P}(c < Z \leq z)}{\mathbb{P}(Z > c)} = \frac{\mathbb{P}\left(\frac{c-m}{\sigma} < \frac{Z-m}{\sigma} \leq \frac{z-m}{\sigma}\right)}{\mathbb{P}\left(\frac{Z-m}{\sigma} > \frac{c-m}{\sigma}\right)}$$

$$f(z \mid Z > c) = \begin{cases} \dfrac{f\left(\frac{z-m}{\sigma}\right) \cdot \frac{1}{\sigma}}{\Phi\left(-\frac{c-m}{\sigma}\right)}, & \text{si } \frac{z-m}{\sigma} > \frac{c-m}{\sigma} \\ 0, & \text{sinon} \end{cases}$$

$$\mathbb{E}(Z \mid Z > c) = \int_{-\infty}^{+\infty} \frac{z \cdot f\left(\frac{z-m}{\sigma}\right) \cdot \frac{1}{\sigma}}{\Phi\left(-\frac{c-m}{\sigma}\right)} \, dz$$

$$= \frac{\phi\left(-\frac{c-m}{\sigma}\right)}{\Phi\left(-\frac{c-m}{\sigma}\right)}$$

si $\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}\right)$

vérifiant $\mathbb{E}[Y_i \mid \delta_i = 1] = \alpha + \beta X_i + \rho \cdot \frac{\phi(aZ_i + b)}{\Phi(aZ_i + b)}$

$$Y_i = \alpha + \beta X_i + \eta_i$$

$$\delta_i = 1 \Rightarrow a + bZ_i + \varepsilon_i > 0$$

$$\Rightarrow \varepsilon_i > -a - bZ_i$$

alors $\mathbb{E}[Y_i \mid \delta_i = 1] = \mathbb{E}[\alpha + \beta X_i + \eta_i \mid \delta_i = 1]$

$$= \alpha + \beta \mathbb{E}[X_i \mid \delta_i = 1] + \mathbb{E}[\eta_i \mid \delta_i = 1]$$

$$= \alpha + \beta \mathbb{E}[X_i \mid \varepsilon_i > -a - bZ_i]$$

$$+ \mathbb{E}[\eta_i \mid \varepsilon_i > -a - bZ_i]$$

$$= \alpha + \underline{\beta X_i} + \mathbb{E}[\eta_i \mid \varepsilon_i > -a - bZ_i]$$

(X_i et ε_i sont indépendants)

D'après la formule d'espérance conditionnelle dans un modèle linéaire :

$\mathbb{E}[Y|X] = mX + n$ où X est la valeur de prédicteur

donc $\mathbb{E}[\eta_i|\varepsilon_i] = m\varepsilon_i + n$ avec $\begin{pmatrix} m \\ n \end{pmatrix} \in \mathbb{R}^2$

on a $\mathbb{E}\left[\mathbb{E}[\eta_i|\varepsilon_i]\right] = \mathbb{E}[m\varepsilon_i + n]$

$= m\,\mathbb{E}[\varepsilon_i] + n$

et aussi $\mathbb{E}\left[\mathbb{E}[\eta_i|\varepsilon_i]\right] = \mathbb{E}[\eta_i]$

donc $m\,\mathbb{E}[\varepsilon_i] + n = \mathbb{E}[\eta_i]$

or $\mathbb{E}[\varepsilon_i] = \mathbb{E}[\eta_i] = 0$

donc $n = 0$, $\mathbb{E}[\eta_i] = \mathbb{E}[m\varepsilon_i] = 0$

$$\mathbb{E}[\eta_i \mid \varepsilon_i > -a - bZ_i] = \mathbb{E}[m\varepsilon_i \mid \varepsilon_i > -a - bZ_i]$$

$$= m\,\mathbb{E}[\varepsilon_i \mid \varepsilon_i > -a - bZ_i]$$

$$= m \cdot \frac{\phi(a+bZ_i)}{\Phi(a+bZ_i)}$$

Ensuite, on va trouver la valeur de m :

$$\mathrm{Cov}\Big(\underbrace{\eta_i - \mathbb{E}[\eta_i|\varepsilon_i]}_{\text{résidu}},\ \underbrace{\varepsilon_i}_{\text{prédicteur}}\Big) = 0$$

Rq : La méthode des moindres carré est basée sur l'idée que les résidus doivent être non corrélés aux valeurs des prédicteurs. Cela signifie que la covariance entre les résidus et les valeurs des prédicteurs doit être nulle.

$$\Rightarrow \quad \mathrm{cov}(\eta_i - m\varepsilon_i,\ \varepsilon_i) = 0$$

$$\Rightarrow \quad \mathrm{cov}(\eta_i, \varepsilon_i) - m\,\mathrm{cov}(\varepsilon_i) = 0$$

or dans l'énoncé :

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right)$$

$$\Rightarrow \quad \rho\sigma^2 - m = 0$$

$$\Rightarrow \quad \rho\sigma^2 = m$$

quand $\sigma^2 = 1$, on obtient :

$$\mathbb{E}[Y_i \mid \delta_i = 1] = \alpha + \beta X_i + \rho \cdot \frac{\phi(a+bZ_i)}{\Phi(a+bZ_i)}$$

CONCLUSION

En conclusion, cette analyse nous a fourni des connaissances approfondies sur les données et les modèles utilisés. Nous avons examiné la distribution et la significativité des variables, étudié les relations entre les variables explicatives et les sinistres de type 1, ainsi que les effets des catégories socio-professionnelles sur les résultats. De plus, nous avons mis en évidence les défis liés à la modélisation des zéros dans une régression et proposé des méthodes d'estimation appropriées. Ces résultats nous permettent de mieux comprendre les facteurs influençant les sinistres et d'obtenir des informations utiles pour la modélisation et la prédiction dans ce domaine.