Mémoire Statistiques S6 n°étudiant: 37012370

Partie 1 : Description des données/Tests d'hypothèse :

Lire les données sous R :

```
>don<-source("C:/Users/pljea/Documents/Cours/L3 Miashs/S6/MI-Statistiques
+S6/Memoire/assurance.R")
```

1)

- **Variables quantitatives :**

Police1

```
> summary(dat$Police1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.5375  1.9500  3.7507  5.0170 54.9850 
```

Sin1

```
> summary(dat$Sin1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.967  10.427  12.117  11.960  13.508  18.967 
```

Sin2

```
> summary(dat$Sin2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.6163  7.4541  8.6644  8.6404  9.9084 16.5794 
```

- **Variables qualitatives :**

Atyph

```
> Atyph<-as.factor(dat$Atyph)
> addmargins(table(dat$Atyph))

   Locataire  Non declare Proprietaire          Sum 
        1963           73         3316         5352 
```

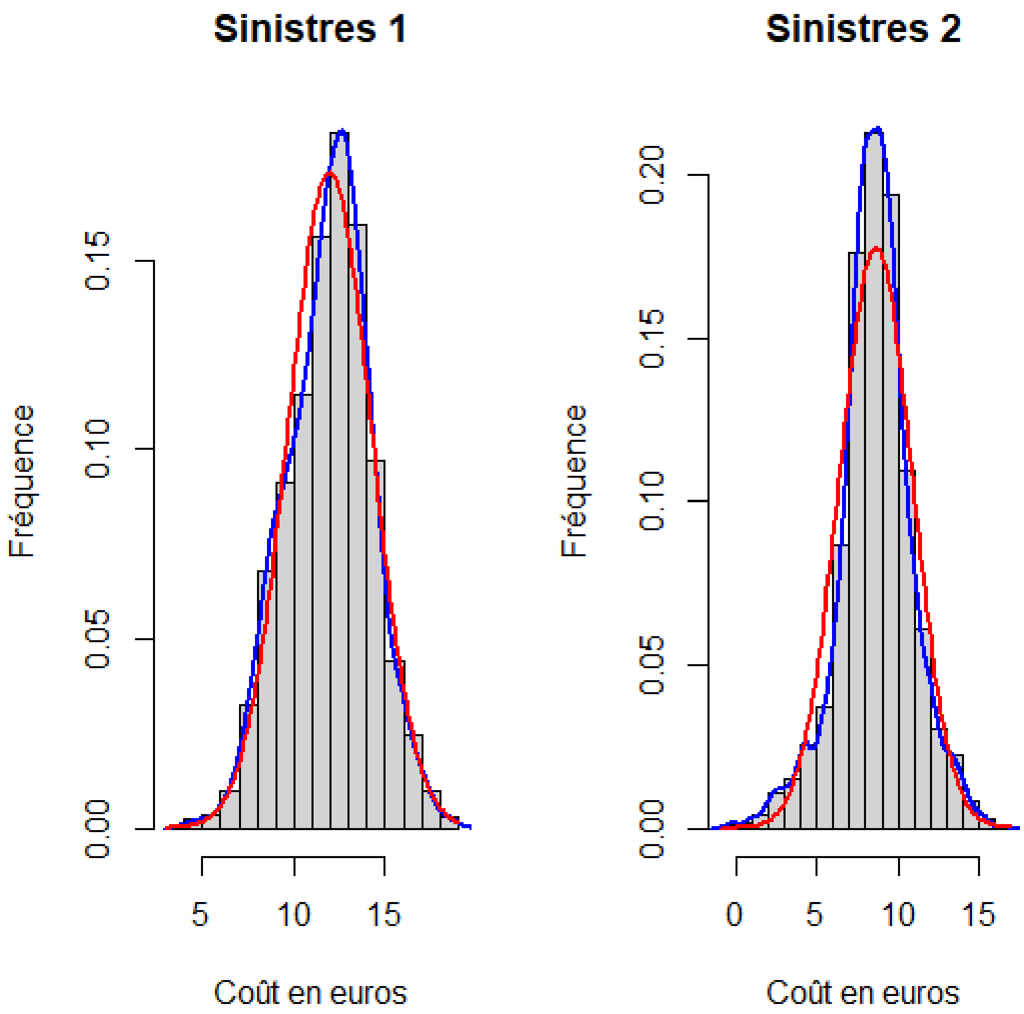
Acompm

```
> Acompm<-as.factor(dat$Acompm)
> addmargins(table(dat$Acompm))

      Autre menage Couple avec enfant(s)     Couple sans enfant 
              1921                  1379                   1293 
    Personne seule                   Sum 
               759                  5352 
```

2)

- **#Histogrammes**

```
>Sin1<-as.numeric(dat$Sin1)
>Sin2<-as.numeric(dat$Sin2)
>par(mfrow=c(1,2))
>hist(Sin1, freq=FALSE, main="Sinistres 1", xlab="Coût en euros", ylab="Fréquence")
>lines(density(Sin1, bw="nrd0", adjust=1, kernel="gaussian"), lwd = 2, col = "blue")
>curve(dnorm(x,mean=11.96042,sd=2.305368),add=TRUE, lwd = 2, col = "red")
>hist(Sin2, freq=FALSE, main="Sinistres 2", xlab="Coût en euros", ylab="Fréquence")
>lines(density(Sin2, bw="nrd0", adjust=1, kernel="gaussian"), lwd = 2, col = "blue")
>curve(dnorm(x,mean=8.640428,sd=2.245154),add=TRUE, lwd = 2, col = "red")
```



- **#Test de Kolmogorov-Smirnov**

#Sin1

```
> ks.test(Sin1, "pnorm")

        One-sample Kolmogorov-Smirnov test

data:  Sin1
D = 0.99996, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Interprétation :
Soit :

H_0 : La fonction de répartition de Sin1 est identique à la fonction de répartition issue d'une loi normale

H_a: La fonction de répartition de Sin1 est différente de la fonction de répartition issue d'une loi normale
Si l'hypothèse nulle est vraie, cela implique que la valeur de la statistique de test D (l'écart le plus grand observé entre les deux courbes) n'est pas trop éloigné de 0, 0 étant la valeur pour laquelle il n'y a aucune différence entre la fonction de répartition de Sin1 et celle issue d'une loi normale.

Pour vérifier cela, on suppose que sous l'hypothèse H_0, la statistique de test D devrait se comporter comme une valeur issue d'une loi de Kolmogorov, dont on connaît la distribution théorique.
La statistique de test D est égale à 0.99996 et la p-valeur est égale strictement inférieure à 2.2*10^(-16).
Ici, on constate que la p-valeur est très faible. Autrement dit, la valeur de D ne peut pas être considérée comme étant issue d'une loi de Kolmogorov et donc on ne peut pas accepter l'hypothèse nulle.
Ceci confirme donc que la loi de probabilité de Sin1 n'est pas celle d'une loi normale.

#Sin2

```
> ks.test(Sin2, "pnorm")

        One-sample Kolmogorov-Smirnov test

data:  Sin2
D = 0.98134, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Interprétation :
La statistique de test D est égale à 0.98134 et la p-valeur est égale strictement inférieure à 2.2*10^(-16).
Ici, on constate que la p-valeur est très faible. Autrement dit, la valeur de D ne peut pas être considérée comme étant issue d'une loi de Kolmogorov et donc on ne peut pas accepter l'hypothèse nulle.
Ceci confirme donc que la loi de probabilité de Sin2 n'est pas celle d'une loi normale.

#Sin1 et Sin2

```
> ks.test(Sin1,Sin2)

        Two-sample Kolmogorov-Smirnov test

data:  Sin1 and Sin2
D = 0.57287, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Interprétation :
Soit :

H_0 : La fonction de répartition de Sin1 est identique à la fonction de répartition de Sin2

H_a: La fonction de répartition de Sin1 est différente de la fonction de répartition de Sin2
Si l'hypothèse nulle est vraie, cela implique que la valeur de la statistique de test D (l'écart le plus grand observé entre les deux courbes) n'est pas trop éloigné de 0, 0 étant la valeur pour laquelle il n'y a aucune différence entre la fonction de répartition de Sin1 et celle de Sin2.

Pour vérifier cela, on suppose que sous l'hypothèse H_0, la statistique de test D devrait se comporter comme une valeur issue d'une loi de Kolmogorov, dont on connaît la distribution théorique.
La statistique de test D est égale à 0.57287 et la p-valeur est égale strictement inférieure à 2.2*10^(-16).
Ici, on constate que la p-valeur est très faible. Autrement dit, la valeur de D ne peut pas être considérée comme étant issue d'une loi de Kolmogorov et donc on ne peut pas accepter l'hypothèse nulle.
Ceci confirme donc que la loi de probabilité de Sin1 n'est pas celle de Sin2. Autrement dit Sin1 et Sin2 ne suivent pas une même loi normale, ne sont pas de même moyenne et de même variance.

3)

- **a) #Test d'indépendance**

```
> Ahabi<-as.factor(dat$Ahabi)
> pcs<-as.factor(dat$pcs)
> chisq.test(Ahabi,pcs,p=0.01)

        Pearson's Chi-squared test

data:  Ahabi and pcs
X-squared = 479.34, df = 28, p-value < 2.2e-16
```

Interprétation :
Soit :

H_0 : le type d'agglomération(Ahabi) dans lequel vit le ménage est indépendante de la catégorie professionnelle (pcs)

H_a : le type d'agglomération(Ahabi) dans lequel vit le ménage dépend de la catégorie professionnelle (pcs)

Sous H_0 : {Ahabi et pcs sont indépendantes}, on a trouvé que la statistique de test est égale à 479.34 en faisant de la simulation de Monte Carlo.

La p-valeur étant très faible (<2.2*10^(-16)), cela indique que la statistique de test est une valeur peu probable, et **on ne peut donc pas accepter l'hypothèse d'indépendance des deux variables au seuil 0.99**.

- **b) Forme de la région de rejet**

Voir p.24/27 b)

```
> qchisq(0.99,28)
[1] 48.27824
```

4)

```
>ecart<-Sin1-Sin2
> ks.test(ecart, "pnorm")

        One-sample Kolmogorov-Smirnov test

data:  ecart
D = 0.74502, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Sin1-Sin2 est une variable qui ne suit pas une loi normale. Le test de comparaison de moyennes le plus approprié est donc le **test de Wilcoxon.** Cependant, l'échantillon étant suffisamment grand, on peut utiliser le **test de Student**.

```
> ecart<-Sin1-Sin2
> t.test(ecart,alternative = "greater",mu=3.0)

        One Sample t-test

data:  ecart
t = 11.27, df = 5351, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 3
95 percent confidence interval:
 3.273284      Inf
sample estimates:
mean of x
 3.319996
```

```
> t.test(ecart,alternative = "less",mu=3.5)

        One Sample t-test

data:  ecart
t = -6.3395, df = 5351, p-value = 1.247e-10
alternative hypothesis: true mean is less than 3.5
95 percent confidence interval:
     -Inf 3.366708
sample estimates:
mean of x
 3.319996
```

Les test 1 montre qu'on rejette l'hypothèse que l'écart des moyennes des sinistres 1 et 2 est inférieure à 3.0 au seuil 0.95. L'intervalle de confiance à 95 % pour cet écart est [3.273284 ; +infini[.
Les test 2 montre qu'on rejette l'hypothèse que l'écart des moyennes des sinistres 1 et 2 est supérieure à 3.5 au seuil 0.95. L'intervalle de confiance à 95 % pour cet écart est

[-infini ; 3.366708[.

L'interprétation de la combinaison des deux tests montrent qu'on rejette l'hypothèse que l'écart des moyennes des sinistres 1 et 2 soit inférieure à 3.0 ou supérieure à 3.5 au seuil 0.95.
L'intervalle de confiance à 95 % pour cet écart est [3.273284 ; 3.366708]

<u>Partie 2 : Modèle linéaire simple :</u>

1)

```
> RUC<-as.numeric(dat$RUC)
> X<-log(RUC)
> Y<-Sin1
> m1 <- lm(Y ~ X)
> summary(m1)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.5217 -1.5824  0.2279  1.5787  6.9986 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.72057    0.49426   13.60   <2e-16 ***
X            0.60946    0.05737   10.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.282 on 5350 degrees of freedom
Multiple R-squared:  0.02066,	Adjusted R-squared:  0.02047 
F-statistic: 112.8 on 1 and 5350 DF,  p-value: < 2.2e-16
```

- **Test de significativité globale du modèle :**

Soit :

H_0 : absence de significativité globale des variables, i.e au moins une variable n'est pas significativement différente de zéro.
Ce test est basé sur la statistique de Fisher.

Ici, la p-valeur **est très faible (<2.2*10^(-16)) et inférieure à 0.01, donc on rejette fortement H_0. Le modèle est donc bien globalement significatif. Les variables log(RUC) et Sin1 sont donc très significatives.**

- **Qualité du modèle :**

Le coefficient de détermination se définit comme la part de variation dans la variable Sin1 (montant dommage en euros pour les sinistres de type 1) qui est expliquée par des variations dans la variable log(RUC) (log du revenu par unité de consommation du ménage) . Plus sa valeur est proche de 1, et plus l'adéquation entre le modèle et les données observées va être forte. Cependant, cette valeur est fortement influencée, par le nombre de variables explicatives incluses

dans la régression. Le R2 ajusté (Adjusted R-Squared) va alors tenir compte de ce nombre et sera donc plus correct.
Ici, le **coefficient de détermination ajusté est égal à 0.02047**. La **valeur est très proche de 0** : l'**adéquation entre le modèle et les données observées est donc très faible**.
L'erreur type des coefficients très proche de 0 témoigne de la stabilité des coefficients. De plus l'écart-type résiduel égal à 2.282, qui est une valeur assez proche de 0 , devrait témoigner d'une assez bonne capacité prédictive du modèle.

- **Interprétation des coefficients :**
- **Significativité**

  -On considère le **coefficient Intercept** :

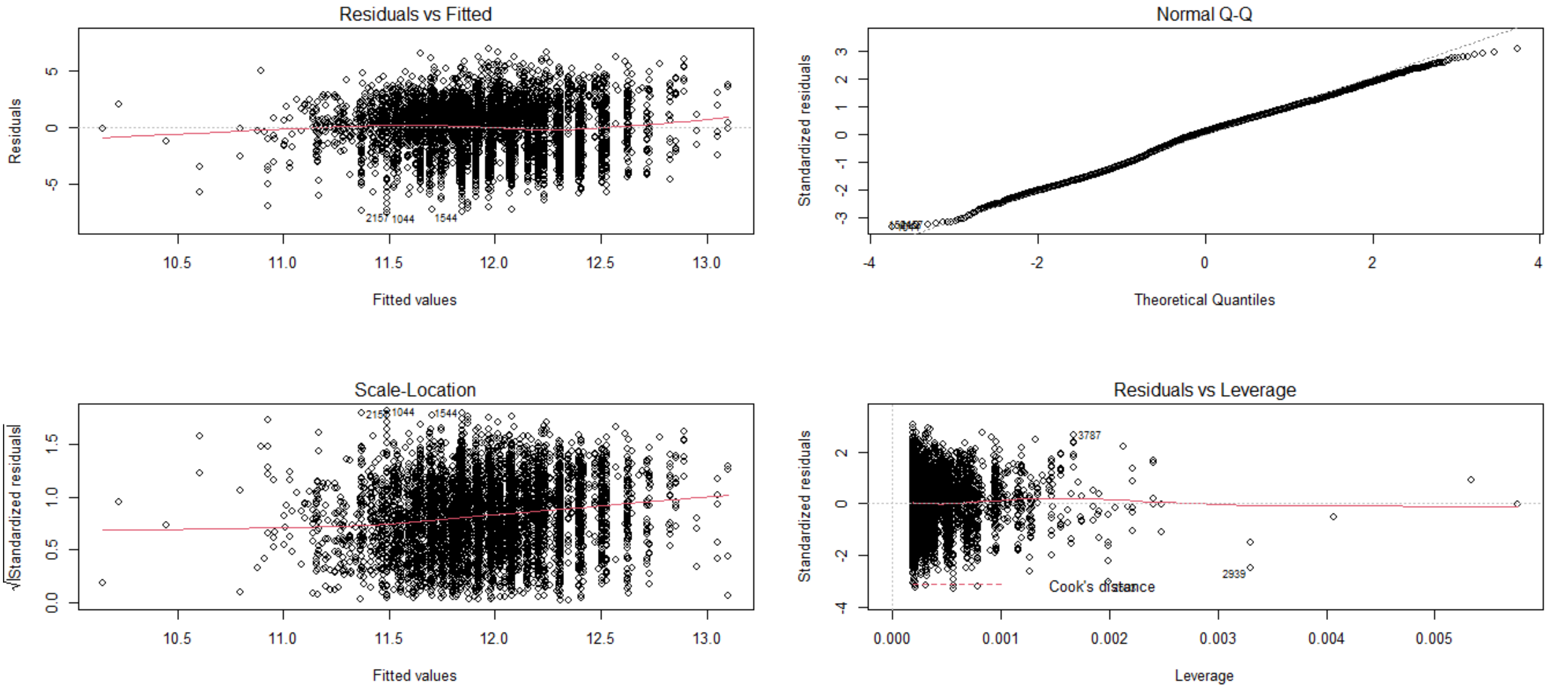
  Soit H_0 : absence de significativité du coefficient Intercept.
  La probabilité pour que la valeur t-calculée soit supérieur en valeur absolue à la valeur théorique est inférieure à 2*10(-16) donc inférieure à 0.05. Donc on rejette fortement H_0 :
  le coefficient Intercept est très significatif, on peut donc étudier son signe et sa magnitude**.**
  -On considère le coefficient de log(RUC) :

  Soit H_0 : absence de significativité du **coefficient log(RUC) :**
  La probabilité pour que la valeur t-calculée soit supérieur en valeur absolue à la valeur théorique est inférieure à 2*10(-16) donc inférieure à 0.05. Donc on rejette fortement H_0 :
  le coefficient de log(RUC) est très significatif, on peut donc étudier son signe et sa magnitude **. Le modèle linéaire est donc pertinent pour expliquer les variations du montant dommage des sinistres de type 1 sur le log du revenu par unité de consommation des ménages. Conclusion : L'ajustement linéaire est mauvais ici. Pour obtenir un meilleur pouvoir prédictif, il faudrait retirer les points aberrants de l'analyse.**
- **Signe et magnitude**

  -On considère le **coefficient Intercept :**
  Le coefficient Intercept est positif et sa magnitude est estimée à 6.72057.
  Donc **toutes choses égales par ailleurs, augmenter le revenu par unité de consommation du ménage de 1 % va approximativement augmenter le montant des sinistres de type 1 de Intercept euros.**
  -On considère le **coefficient de log(RUC) :**
  Le coefficient de log(RUC) est positif et sa magnitude est estimée à 0.60946 .
  Donc toutes choses égales par ailleurs, augmenter le revenu par unité de consommation du ménage de 1 % va approximativement augmenter le montant des sinistres de type 1 coefficient de log(RUC) euros.

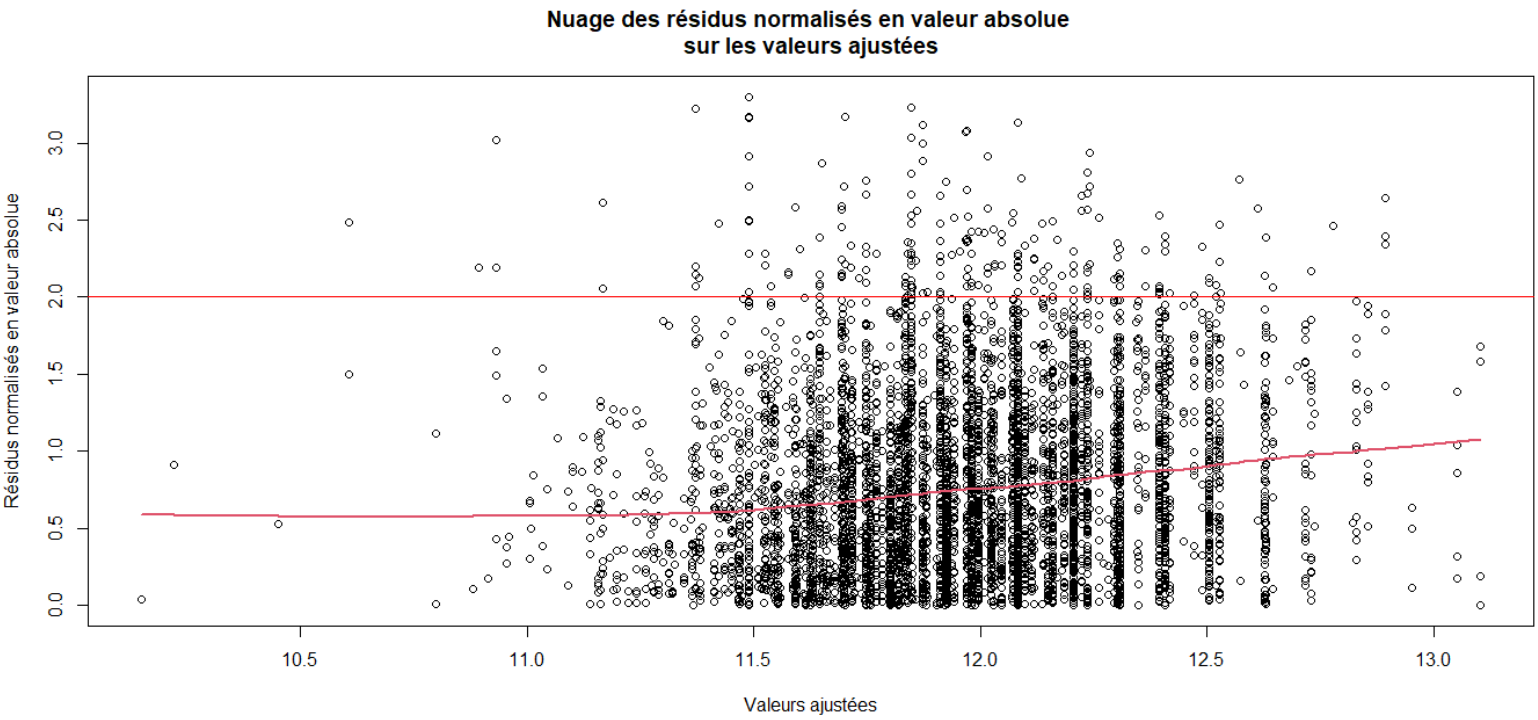
- **Analyse des résidus :**

```
>op <- par(mfrow=c(2,2))
>plot(m1)
>par(op)
```



D'après le graphique 2, **les résidus suivent globalement la droite de Henry, ils sont donc normalement distribués. L'hypothèse de normalité donc est vérifiée.**

```
>plot(m1$fitted,abs(rstudent(m1)),main="Nuage des résidus normalisés en valeur absolue
sur les valeurs ajustées",
>xlab="Valeurs ajustées",ylab="Résidus normalisés en valeur absolue")
>lines(lowess(m1$fitted,abs(rstudent(m1)),f=0.5),lwd=2,col=2)
>abline(h=2,col="red")
```



On observe que l'estimation de la **tendance est croissante à partir d'une certaine valeur (11.5)**, donc que la **variance des résidus augmente le long de l'axe des abscisses à partir d'une certaine valeur**. : **cette structure des résidus révèlent une hétéroscédasticité. Cette hétéroscédasticité peut avoir plusieurs sources:**

- **L'hétérogénéité de l'échantillon étudié pour les montants dommage en euros de sinistre de type 1 dépassant la valeur 11.5**
- **L'omission de variables explicatives dans le modèle**

L'observation de cette hétéroscédasticité est renforcée par l'analyse du graphique Résidus Vs Levier qui montre que la **propagation des résidus normalisés semble diminuer à partir d'une certaine valeur de l'effet levier (0.0022)**.

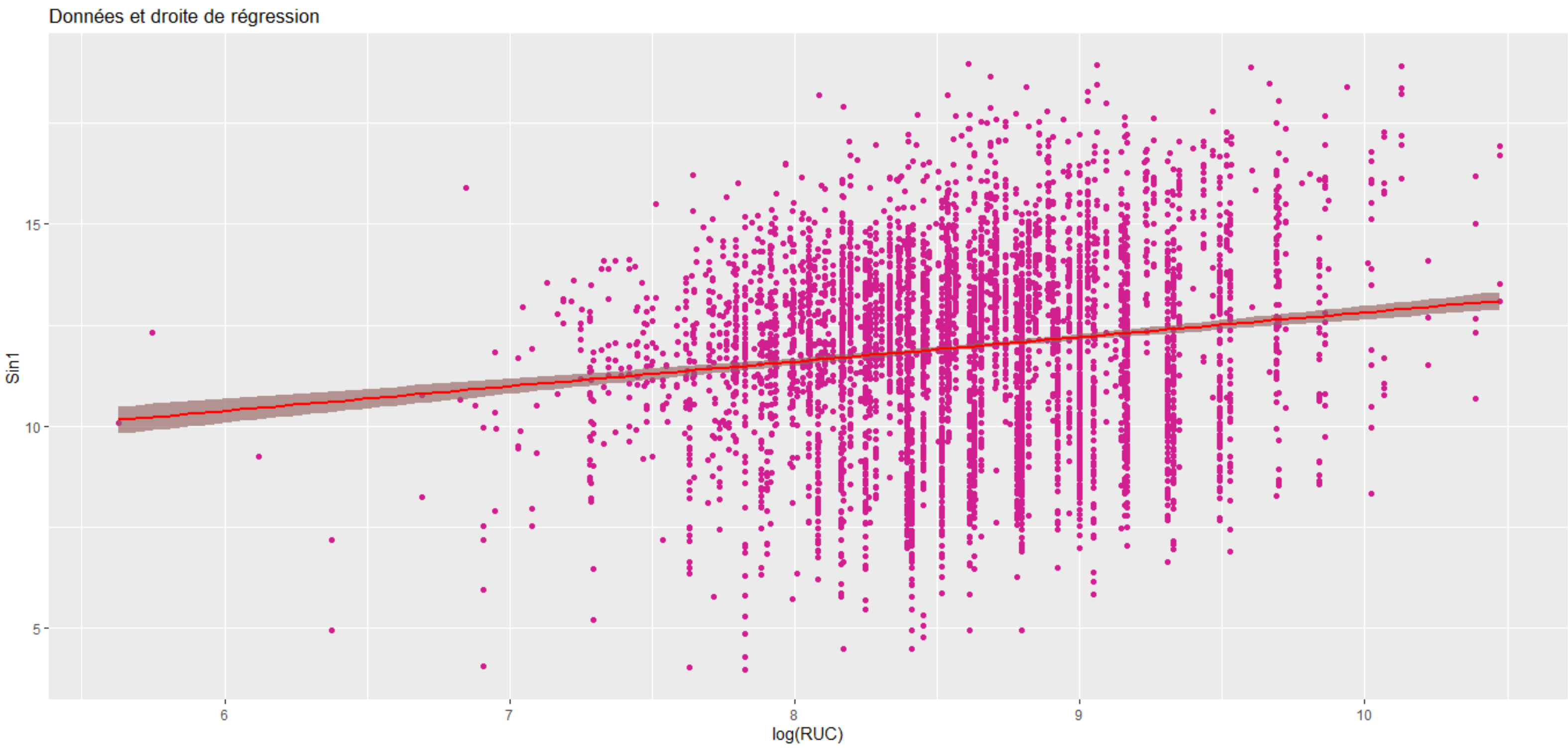
Par ailleurs, **on observe que plusieurs résidus sont supérieurs à 2 en valeur absolue : ils témoignent de la présence de points aberrants et de points contribuant à la détermination du modèle.**

<u>Conclusion :</u>

L'importante significativité de la variable log(RUC) indique qu'il s'agit d'une variable explicative qui doit être conservée dans le modèle. Cependant, son coefficient est très proche de 0, donc on pourrait penser au premier abord qu'il s'agit d'une variable non utile pour le modèle qui pourrait expliquer la faible adéquation entre le modèle observé. Cependant l'analyse des résidus a montré la présence d'hétéroscédasticité. Cette hétéroscédasticité témoigne de la forte disparité entre les types d'accidents (catastrophe naturel, attentat, accident de la route…) qui atteignent les ménages et de l'hétérogénéité de leur situation individuelle (options de contrats choisies, vie en concubinage ou non, enfants ou non etc.) pour des montants dommages en euros de sinistres de type 1 qui dépassent en valeur 11.5. Cette hétéroscédasticité témoigne aussi de l'omission de variables explicatives du modèle dont les catégories professionnelles (pcs). Il convient donc de recommencer le modèle en éliminant les valeurs aberrantes et en ajoutant les catégories professionnelles dans les variables explicatives.

2)

```
>library(ggplot2)
>ggplot(dat) +
 aes(x = log(RUC), y = Sin1) +
 ggtitle("Données et droite de régression")+
 geom_point(color="violetred") +
 geom_smooth(colour="red4", method="lm", fill="red4") +
 geom_smooth(method = "lm", formula = y ~ x, col = "red")
```



Droite de régression estimée :

```
>b1<-cov(X,Y)/var(X)
>b0<-mean(Y)-b1*mean(X)
>b0;b1
```

hat(y)=6.721+0.609x

3)

L'Analyse de la covariance (ANCOVA) est un modèle adapté pour expliquer une variable quantitative Y (dans notre exemple le montant dommage des sinistres de type 1) en fonction d'une variable quantitative x (dans notre exemple le log du revenu par unité de consommation des ménages) et d'une variable qualitative ayant I modalités (dans notre exemple les catégories professionnelles qui ont 8 modalités : les agriculteurs exploitants, les Artisans-commerçants-chefs d'entreprises, les cadres et professions intellectuelles supérieures, les Professions intermédiaires, les Employés, les Ouvriers,les Retraités et les Autres personnes sans activité professionnelle) .

```
> pcs<-as.factor(dat$pcs)
> m2 <- lm(Y ~ X + pcs)
> summary(m2)

Call:
lm(formula = Y ~ X + pcs)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.1152 -0.9449 -0.0149  0.9238  6.7971 

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              7.92560    0.36099  21.955  < 2e-16
X                                        0.47823    0.04219  11.334  < 2e-16
pcsArtisans, comm., chefs d'ent.         0.86756    0.17963   4.830 1.41e-06
pcsAutres pers. sans activite prof.     -3.27405    0.17831 -18.362  < 2e-16
pcsCadres et prof. intellectuelles sup.  3.17090    0.16225  19.544  < 2e-16
pcsEmployes                              0.51470    0.15086   3.412  0.00065
pcsOuvriers                              0.47741    0.14583   3.274  0.00107
pcsProfessions intermediaires            0.84819    0.15159   5.595 2.31e-08
pcsRetraites                            -2.39897    0.14905 -16.095  < 2e-16

(Intercept)                             ***
X                                       ***
pcsArtisans, comm., chefs d'ent.        ***
pcsAutres pers. sans activite prof.     ***
pcsCadres et prof. intellectuelles sup. ***
pcsEmployes                             ***
pcsOuvriers                             ** 
pcsProfessions intermediaires           ***
pcsRetraites                            ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.495 on 5343 degrees of freedom
Multiple R-squared:  0.5803,	Adjusted R-squared:  0.5797 
F-statistic: 923.4 on 8 and 5343 DF,  p-value: < 2.2e-16

> anova(m1,m2)
Analysis of Variance Table

Model 1: Y ~ X
Model 2: Y ~ X + pcs
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
1   5350 27852                                  
2   5343 11936  7     15915 1017.7 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Estimation du modèle**
  a)
  - **Significativité de la variable log(RUC)**

Le test basé sur la statistique de Fisher montre que la p-valeur est très faible (<2.2*10^(-16)) et inférieure à 0.01, donc on rejette fortement l’hypothèse nulle d’absence de significativité d’au moins l'une des variable du modèle. Le modèle est donc bien globalement significatif. Les variables log(RUC) et Sin1 sont donc très significatives. **La variable log(RUC) est encore très significative, de plus son coefficient est significativement différent de 0 car c'est une variable qui doit être conservée dans le modèle et donc plus on est riche au sein d'une catégorie professionnelle, plus on a de sinistres de type 1.**

- **Qualité du modèle :**

Le **coefficient de détermination ajusté est égal à 0.5803**. La **valeur est dans la moitié supérieure de l'intervalle [0;1]** : l'**adéquation entre le modèle et les données observées est donc moyennement élevé**.
L'erreur type des coefficients très proche de 0 témoigne de la stabilité des coefficients. De plus l'écart-type résiduel égal à 1.495, qui est une valeur assez proche de 0 , témoigne d'une assez bonne capacité prédictive du modèle.

- **Analyse de la covariance**

L'analyse de la covariance montre que les catégories professionnelles ont des effets significativement différents (p-valeur<2.2*10(-16)<0.05) sur le montant dommage en euros des sinistres de type 1.

b) Voir p.24/27 b)

4)

- **Modèle pcs**

```
> m3 <- lm(Y ~ pcs)
> summary(m3)

Call:
lm(formula = Y ~ pcs)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0716 -0.9681 -0.0147  0.9288  6.5988 

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              11.7025     0.1404  83.339  < 2e-16
pcsArtisans, comm., chefs d'ent.          1.1600     0.1799   6.449 1.23e-10
pcsAutres pers. sans activite prof.      -3.0690     0.1795 -17.099  < 2e-16
pcsCadres et prof. intellectuelles sup.   3.7275     0.1565  23.822  < 2e-16
pcsEmployes                               0.8205     0.1502   5.463 4.88e-08
pcsOuvriers                               0.6818     0.1464   4.656 3.30e-06
pcsProfessions intermediaires             1.2411     0.1493   8.312  < 2e-16
pcsRetraites                             -1.9856     0.1462 -13.578  < 2e-16
                                           
(Intercept)                             ***
pcsArtisans, comm., chefs d'ent.        ***
pcsAutres pers. sans activite prof.     ***
pcsCadres et prof. intellectuelles sup. ***
pcsEmployes                             ***
pcsOuvriers                             ***
pcsProfessions intermediaires           ***
pcsRetraites                            ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.512 on 5344 degrees of freedom
Multiple R-squared:  0.5702,	Adjusted R-squared:  0.5696 
F-statistic:  1013 on 7 and 5344 DF,  p-value: < 2.2e-16
```

- **Quel modèle choisir ?**
  Les analyses ont montré que les variables **log(RUC) restait très significative après l'introduction de la variable pcs qui est également très significative**, il convient donc de choisir un **modèle qui conserve ces deux variables** dans les variables explicatives. Par ailleurs, le **coefficient de détermination du modèle log(RUC) et pcs est le plus élevé** (0.5803) des trois modèles, le **modèle log(RUC) et pcs** est donc celui qui **détient l'adéquation avec les données la plus élevée**. Par ailleurs, **l'écart-type résiduel est la valeur la plus proche de 0 (1.495)** des trois modèles, ce qui témoigne que **la capacité prédictive du modèle log(RUC) et pcs est la plus élevée des trois modèles**.

  Conclusion :
  Le modèle que je choisirais est donc celui avec log(RUC) et pcs en variables explicatives.

Partie 3 : Modèle linéaire Multiple

```
>newdat<-dat[- c(5343:5352),]
>dat0<-dat[c(5343:5352),]
```

1)

```
>library(RCMDR)
```

```
> LinearModel.1 <- lm(Sin1 ~ RUC + log(RUC) + Acompm + Ahabi + Atyph + Bauto
+   + cs + habi + Nbadulte + nbpers + NSin + pcs + Police1 + Police2 + Police3
+   + region + reves + Sin2 + Sin3, data=newdat)
```

- **Sélection du modèle par étapes :**
  -Calibration d'un modèle explicatif

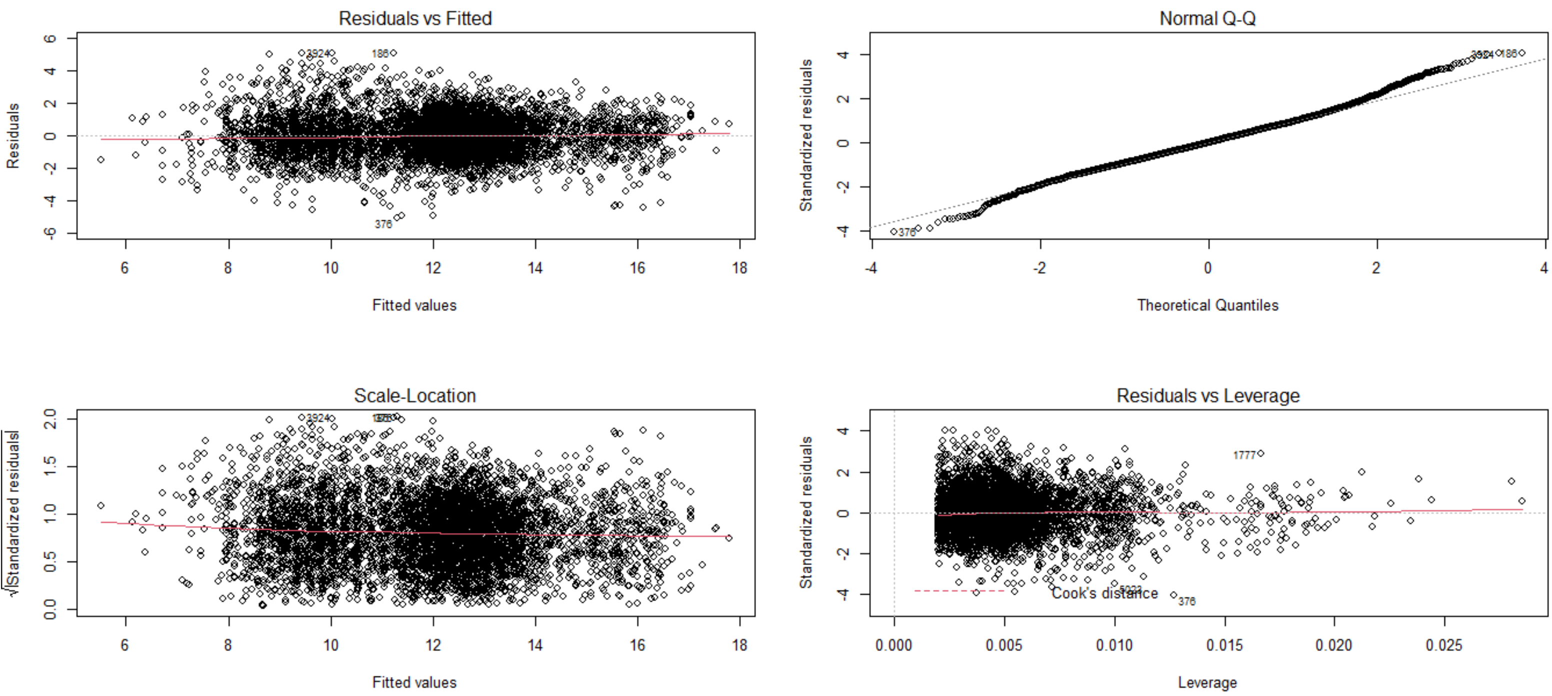
  ```
  > library(MASS, pos=16)

  > stepwise(LinearModel.1, direction='backward/forward', criterion='AIC')
  ```

  En partant du modèle le plus gros, puis en enlevant des variables et en rajoutant jusqu'à la convergence par le critère AIC, le meilleur modèle retenu pour expliquer Sin1 est celui qui retient les variables explicatives log(RUC), Acompm, Atyph, cs, habi, Nbadulte, nbpers et pcs.

```
Call:
lm(formula = Sinl ~ log(RUC) + Acompm + Atyph + cs + habi + Nbadulte +
    nbpers + pcs, data = newdat)

Coefficients:
                                (Intercept)
                                   -1.38476
                                   log(RUC)
                                    1.28848
            Acompm[T.Couple avec enfant(s)]
                                    0.28756
               Acompm[T.Couple sans enfant]
                                   -0.19306
                   Acompm[T.Personne seule]
                                   -0.26622
                       Atyph[T.Non declare]
                                    0.08131
                      Atyph[T.Proprietaire]
                                    0.08120
                             cs[T.Modeste]
                                    0.35248
                         cs[T.Moyenne Inf]
                                    0.17678
                         cs[T.Moyenne Sup]
                                    0.14683
                                habi[T.1]
                                    0.02575
                                habi[T.2]
                                    0.06008
                                habi[T.3]
                                    0.44922
                                habi[T.4]
                                    0.41567
                                habi[T.5]
                                    0.47954
                                habi[T.6]
                                    0.75877
                                habi[T.7]
                                    0.68987
                                habi[T.8]
                                    0.96360
                                 Nbadulte
                                    0.23900
                                   nbpers
                                    0.43398
       pcs[T.Artisans, comm., chefs d'ent.]
                                    0.41130
    pcs[T.Autres pers. sans activite prof.]
                                   -2.83313
pcs[T.Cadres et prof. intellectuelles sup.]
                                    2.31608
                          pcs[T.Employes]
                                    0.25438
                          pcs[T.Ouvriers]
                                    0.08039
        pcs[T.Professions intermediaires]
                                    0.35299
                         pcs[T.Retraites]
                                   -1.99968
```

- **Analyse des résidus**

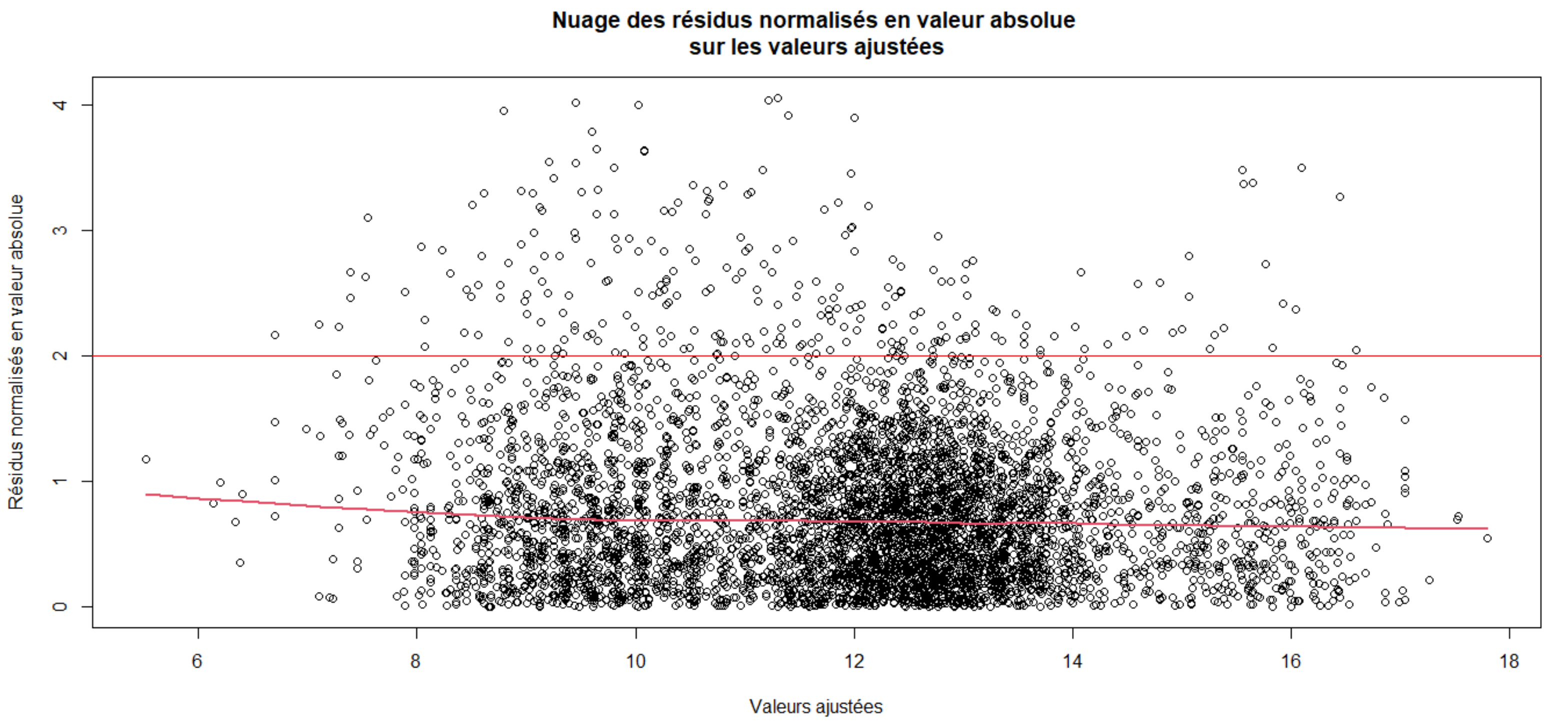
```
>m4<-lm(formula = Sin1 ~ log(RUC) + Acompm + Atyph + cs + habi + Nbadulte +
+    nbpers + pcs, data = newdat)

>op <- par(mfrow=c(2,2))
>plot(m4)
>par(op)
```



- Normalité des résidus

D'après le graphique 2, **les résidus suivent globalement la droite de Henry, ils sont donc normalement distribués. L'hypothèse de normalité donc est vérifiée**.

- **Hétéroscédasticité des résidus et points aberrants**



On observe que l'estimation de la **tendance est globalement décroissante**, donc que la **variance des résidus diminue le long de l'axe des abscisses** : **cette structure des résidus révèlent une hétéroscédasticité.**

L'observation de cette hétéroscédasticité est renforcée par l'analyse du graphique Résidus Vs Levier qui montre que la **propagation des résidus normalisés semble globalement augmenter**.

Par ailleurs, **on observe que plusieurs résidus sont supérieurs à 2 en valeur absolue : ils témoignent de la présence de points aberrants et de points contribuant fortement à la détermination du modèle.**

3) >library(dplyr)

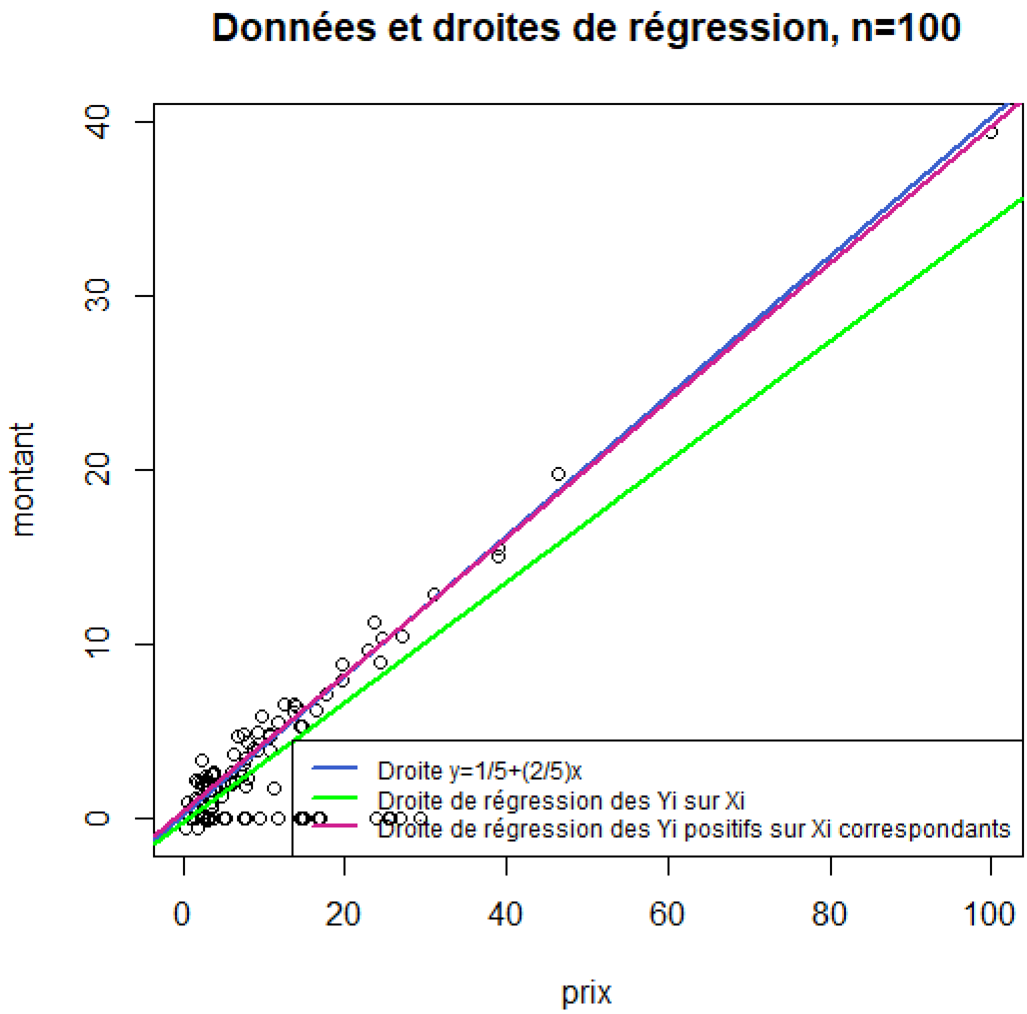
```
> pc<-predict(m4,dat0, level = 0.95, interval = "confidence")
> pc<-pc[,1]
> pc<-as.data.frame(pc)
> montantobs<-dat0[,17]
> montantobs<-as.data.frame(montantobs)
> tab<-bind_cols(pc, montantobs, id = NULL)
> colnames(tab)=c("Valeur prédite","Montant observé")
> tab
     Valeur prédite Montant observé
5343       8.833596        6.138658
5344      13.169304       12.396894
5345       9.331877        8.844001
5346      10.816640        9.028260
5347      11.101659       10.080450
5348       9.940551        7.962781
5349      11.966110       11.290157
5350      16.558954       17.207883
5351      12.999109       14.909703
5352      13.428163       13.721335
```

Partie 4 : Modélisation des zéros pour sinistre3

1)
Voir p.25-26/27

2)
```
>a<- -0.5
>b<-1
>alpha<-1/5
>beta<-2/5
>n<-100
>mb <- matrix(NA, 100, 4)

>for(i in 1:n)
+{
+eta<-rnorm(1,mean=0,sd=1)
+X<-rlnorm(1,meanlog=2,sdlog=1)
+Z<-exp(rnorm(1,mean=0,sd=1))
+dirac<-rbinom(1,1,pnorm(a+b*Z))

+if(dirac==0)
+{
+Y<-0
+}

+else
+{
+Y<-alpha+beta*X+eta
+}

+mb[i,1]<-dirac
+mb[i,2]<-Y
+mb[i,3]<-X
+mb[i,4]<-Z

+}
>mb

>selection<-mb[,1]
>montant<-mb[,2]
>prix<-mb[,3]
>vitesse<-mb[,4]

>df<-data.frame(montant,prix)
>df <- subset(df, montant > 0)
>df
>montantpos<-df[,1]
```

```
>prix2<-df[,2]

>don <- data.frame(selection,montant,prix,vitesse)
>cor(don)
>plot(montant ~ prix, don, main="Données et droites de régression, n=100")
>abline(alpha,beta,lwd = 2, col = "royalblue3")
>abline(lm(montant ~ prix, don), lwd = 2, col = "green")
>abline(lm(montantpos ~ prix2, don), lwd = 2, col = "violetred")
>legend(x="bottomright", legend=c("Droite y=1/5+(2/5)x",
+"Droite de régression des Yi sur Xi",
+"Droite de régression des Yi positifs sur Xi correspondants"),
+col=c("royalblue3","green","violetred"), lty=c(1,1,1), cex=0.75)
```



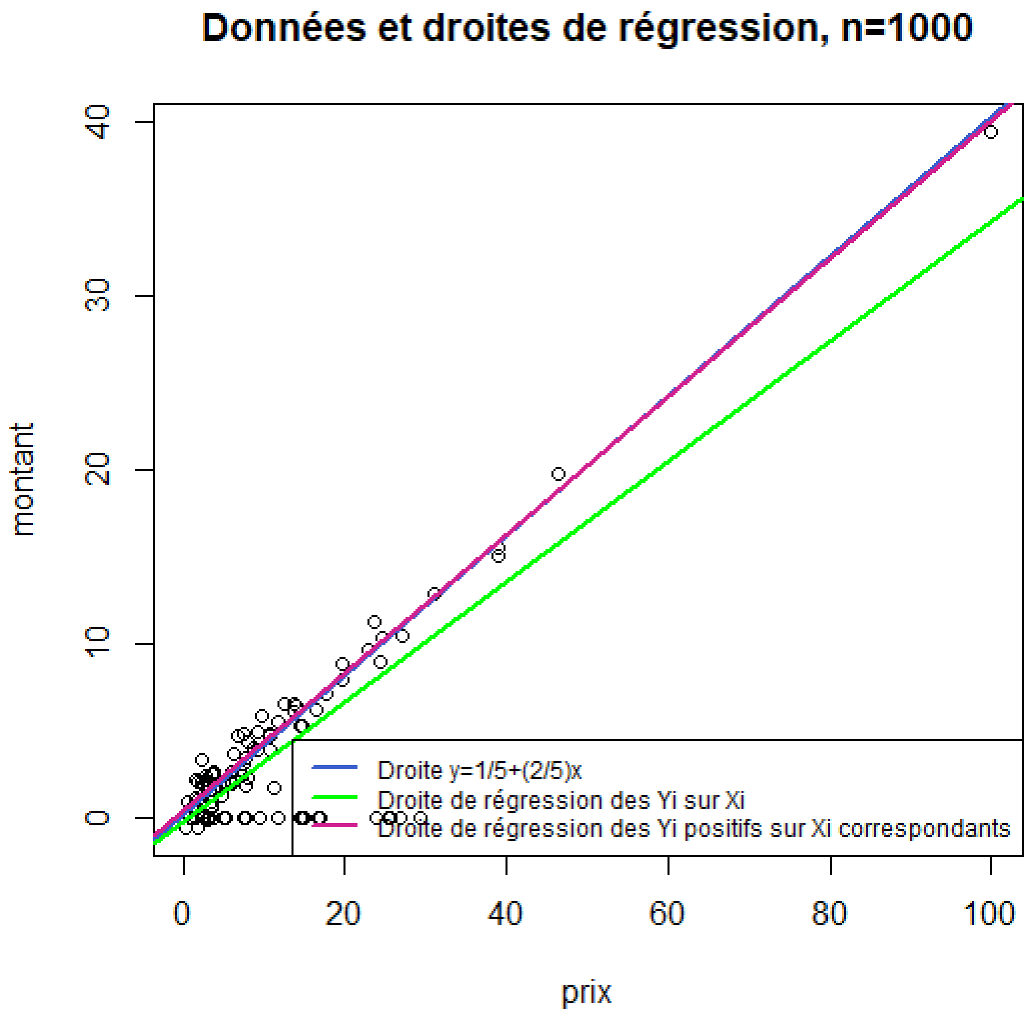
- **Commentaires**

On observe que les pentes de la droite de régression des Yi positifs sur Xi correspondants est presque égale à celle de la droite d'équation y=1/5+(2/5)x. En revanche, la pente de la droite de régression des Yi sur Xi est bien plus faible que celle de la droite d'équation y=1/5+(2/5)x. **On en déduit que les deux régressions sont significativement différentes l'une de l'autre et que la**

meilleure approximation de la droite de régression des Yi positifs sur Xi correspondants est la droite d'équation y=1/5+(2/5)x.

```
>a<- -0.5
>b<-1
>alpha<-1/5
>beta<-2/5
>n<-1000
>mb <- matrix(NA, 1000, 4)

>for(i in 1:n)
+{
+eta<-rnorm(1,mean=0,sd=1)
+X<-rlnorm(1,meanlog=2,sdlog=1)
+Z<-exp(rnorm(1,mean=0,sd=1))
+dirac<-rbinom(1,1,pnorm(a+b*Z))

+if(dirac==0)
+{
+Y<-0
+}

+else
+{
+Y<-alpha+beta*X+eta
+}

+mb[i,1]<-dirac
+mb[i,2]<-Y
+mb[i,3]<-X
+mb[i,4]<-Z

+}
>mb

>selection<-mb[,1]
>montant<-mb[,2]
>prix<-mb[,3]
>vitesse<-mb[,4]

>df<-data.frame(montant,prix)
>df <- subset(df, montant > 0)
>df

>montantpos<-df[,1]
>prix2<-df[,2]
```

```
>don <- data.frame(selection,montant,prix,vitesse)
>cor(don)
>plot(montant ~ prix, don, main="Données et droites de régression, n=1000")
>abline(alpha,beta,lwd = 2, col = "royalblue3")
>abline(lm(montant ~ prix, don), lwd = 2, col = "green")
>abline(lm(montantpos ~ prix2, don), lwd = 2, col = "violetred")
>legend(x="bottomright", legend=c("Droite y=1/5+(2/5)x",
+"Droite de régression des Yi sur Xi",
+"Droite de régression des Yi positifs sur Xi correspondants"),
+col=c("royalblue3","green","violetred"), lty=c(1,1,1), cex=0.75)
```



- **Les estimateurs des mco (sur tout l'échantillon et uniquement sur variables positives) semblent-ils convergents? Quels sont les problèmes?**
  - **Estimateurs des mco sur tout l'échantillon**
    - **hat(α)**

Les ordonnées à l'origine de la droite d'équation y=1/5+(2/5)x et de la droite de régression des Yi sur Xi se confondent. Donc **l'estimateur hat(α) semble converger vers 1/5**.

- **hat(β)**
  Au fur et à mesure que le prix augmente (sur tout l'échantillon), la droite de régression des Yi sur les Xi correspondants s'éloigne de la droite d'équation y=1/5+(2/5)x par une pente plus faible. Donc **l'estimateur hat(β) semble ne pas converger**.

  Ce problème vient du fait qu'en répétant les mesures, sans rien changer au phénomène observé , **on introduit avec les valeurs nulles de la variable expliquée des erreurs de mesure aléatoires dans les données**. **Le terme constant de l'erreur incorporé dans la variable explicative prix induit un biais qui affecte l'estimation de la constante de régression**. En l'occurrence, ici **l'hypothèse de normalité des erreurs n'est pas vérifiée** et l'estimateur **hat(β) est biaisé donc non convergent.**

- **Estimateurs des mco sur les variables positives**
- **hat(α)**
  Les ordonnées à l'origine de la droite d'équation y=1/5+(2/5)x et de la droite de régression des Yi positifs sur les Xi correspondants se confondent. Donc **l'estimateur hat(α) semble converger vers 1/5**.
- **hat(β)**
  Au fur et à mesure que le prix augmente, la droite de régression des Yi positifs sur Xi correspondants se confond avec la droite d'équation y=1/5+(2/5)x. Donc **l'estimateur hat(β) semble converger vers 2/5**.

3)

```
>a<- -0.5
>b<-1
>alpha<-1/5
>beta<-2/5
>n<-100
>N<-999
>donnees<-NULL

>for(j in 1:N)
+{

+mb <- matrix(NA, 100, 4)
+mb<-as.data.frame(mb)

+for(i in 1:n)
+{
+eta<-rnorm(1,mean=0,sd=1)
+X<-rlnorm(1,meanlog=2,sdlog=1)
+Z<-exp(rnorm(1,mean=0,sd=1))
+dirac<-rbinom(1,1,pnorm(a+b*Z))
```

```
+if(dirac==0)
+{
+Y<-0
+}

+else
+{
+Y<-alpha+beta*X+eta
+}

+mb[i,1]<-dirac
+mb[i,2]<-Y
+mb[i,3]<-X
+mb[i,4]<-Z

+}

+donnees<-rbind(donnees,mb)
+}
>colnames(donnees)=c("Dirac","Y","X","Z")
>donnees
```

```
> out<-split(donnees, factor(sort(rank(row.names(donnees))%%N)))
> out[[999]]
```

	Dirac	Y	X	Z
99801	1	3.89991847	9.5069346	0.43764083
99802	0	0.00000000	37.7353306	0.33227372
99803	1	3.25499420	6.2685907	0.91221617
99804	1	3.79090149	8.8567771	1.32515628
99805	1	3.70390452	3.8670217	1.62212906
99806	0	0.00000000	2.8237298	0.97022458
99807	1	2.71216449	7.1690336	4.48374501
99808	1	3.25066283	8.6510365	0.26739929
99809	0	0.00000000	20.2932878	0.21527958
99810	1	7.10204020	17.0586709	0.47873014
99811	1	11.41219763	31.0388006	1.66555396
99812	0	0.00000000	2.2491958	0.73070031
99813	1	3.66761414	9.7748224	1.94927240
99814	1	4.43171340	11.0735039	3.03036712
99815	0	0.00000000	2.4781189	0.29218438
99816	1	4.14910413	11.9465120	6.42128616
99817	1	1.48569771	4.9223624	0.73766820
99818	0	0.00000000	2.5858241	0.31175709
99819	1	1.86475521	4.7870713	3.66653590
99820	1	2.12519712	8.2478951	8.93756090
99821	1	14.76154129	31.4236339	1.21553502
99822	1	9.71162867	22.5616821	0.06749460
99823	1	2.79914651	10.4565207	1.60860937
99824	0	0.00000000	56.3242475	0.49972936

…

4)

- **Estimation du modèle Probit**

```
> df<-data.frame(selection,montant,prix,vitesse)
> mPROBIT<-glm(selection ~ vitesse, data=df, family=binomial(link=probit))
> summary(mPROBIT)

Call:
glm(formula = selection ~ vitesse, family = binomial(link = probit),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6007  -1.0715   0.3711   0.8917   1.3944

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.36286    0.08304   -4.37 1.25e-05 ***
vitesse      0.86448    0.07949   10.88  < 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: 1200.70  on 999  degrees of freedom
Residual deviance:  977.06  on 998  degrees of freedom
AIC: 981.06

Number of Fisher Scoring iterations: 7
```

Le modèle mProbit s'écrit :
hat(Yi)=-0.36286+0.86448Xi…

Non, les estimateurs obtenus ne sont pas convergents car â=-0.36286>>0.5 et hat(b)=0.86448<<1.

5)
Voir p.26-27/27

Partie 1 : Description des données / Test d'hypothèse

3. b) Forme de la région de rejet

La variable Ahabi a 5 modalités
La variable pcs a 8 modalités

$$W_n = \left\{ \frac{\sum_{i=1}^{5} \sum_{j=1}^{8} \left(\frac{N_{ij}}{n} - \frac{N_{i\circ} N_{\circ j}}{n} \right)^2}{\frac{N_{i\circ} N_{\circ j}}{n}} > \chi^2_{1-0.01}((5-1)(8-1)) \right\}$$

$$W_n = \left\{ \frac{\sum_{i=1}^{5} \sum_{j=1}^{8} \left(\frac{N_{ij}}{n} - \frac{N_{i\circ} N_{\circ j}}{n} \right)^2}{\frac{N_{i\circ} N_{\circ j}}{n}} > \chi^2_{0.99}(28) \right\}$$

$$W_n = \left\{ \frac{\sum_{i=1}^{5} \sum_{j=1}^{8} \left(\frac{N_{ij}}{n} - \frac{N_{i\circ} N_{\circ j}}{n} \right)^2}{\frac{N_{i\circ} N_{\circ j}}{n}} > 48.27824 \right\}$$

Partie 2 : Modèle linéaire simple

3. b) Estimation du modèle :

Le modèle de covariance s'écrit :

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \varepsilon_{ij} \quad \text{où} \quad \varepsilon_{ij} \text{ iid} \sim \mathcal{N}(0, \sigma^2)$$
$$1 \leq i \leq 8$$

- i est l'indice de la catégorie professionnelle
- j est l'indice de répétition i.e le j-ème individu pour sa catégorie professionnelle.

- Y_{ij} est le montant dommage en euros du sinistre de type 1 pour le j-ème individu de la catégorie professionnelle i
- X_{ij} est le log du revenu par unité de consommation du j-ème individu de la catégorie professionnelle i.
- α_i est la valeur du montant dommage en euros du sinistre de type 1 pour un individu de la catégorie professionnelle i au log revenu par unité de consommation nul.
- β_i est la pente de régression pour la catégorie professionnelle
- σ^2 est la variance résiduelle (identique pour tous les traitements)

Partie 4 : Modélisation des zéros pour sinistre 3

1. $(\delta_i)_{i=1..n}$ est une suite de variables aléatoires qui valent 1 si $U_i = a + bZ_i + \varepsilon_i > 0$ et 0 si $U_i = a + bZ_i + \varepsilon_i < 0$.
Donc (δ_i) suit une loi de Bernoulli de paramètre $p_i = \mathbb{P}(U_i > 0)$

$$p_i = \mathbb{P}(U_i > 0) \iff p_i = \mathbb{P}(a + bZ_i + \varepsilon_i > 0)$$
$$\iff p_i = \mathbb{P}(-\varepsilon_i < a + bZ_i)$$

Or $\forall x \in \mathbb{R}$, $-x \in \mathbb{R}$ et
$$f_{\varepsilon_i}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

De plus :
$$f_{\varepsilon_i}(-x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(-x)^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = f_{\varepsilon_i}(x).$$

La densité est paire donc ε_i et $-\varepsilon_i$ ont la même loi donc $-\varepsilon_i \sim \mathcal{N}(0,1)$

Donc $p_i = \mathbb{P}(-\varepsilon_i \leq a + b z_i)$
$= \Phi(a + b z_i)$

Finalement $\delta_i \sim Ber(1, p_i)$ avec probabilité $p_i = \Phi(a + b z_i)$

5)
• D'après la formule de l'espérance totale :
$\mathbb{E}(Z) = \mathbb{P}(Z > c) \times \mathbb{E}(Z \mid Z > c)$
Donc :
$$\mathbb{E}(Z \mid Z > c) = \frac{\mathbb{E}(Z)}{\mathbb{P}(Z > c)} = \frac{\mathbb{E}(Z)}{\mathbb{P}(-Z < -c)}$$

D'après 1) $-Z$ et Z ont la même loi donc :
$\mathbb{P}(-Z < -c) = \mathbb{P}(Z < -c) = \mathbb{P}(Z \leq -c) = \Phi(-c)$

De plus :
$$\mathbb{E}(Z) = \int_{-\infty}^{+\infty} z f_Z(z) dz$$
$$= \int_{-\infty}^{+\infty} z \times \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$
$$= \int_{-\infty}^{c} z \times \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz + \int_{c}^{+\infty} z \times \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$
Or $z > c$ donc $\int_{-\infty}^{c} z \times \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0$

Donc
$$\mathbb{E}(Z) = \int_{c}^{+\infty} z \times \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

La densité est paire donc :

$$\int_c^{+\infty} z \times \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \int_{-\infty}^{-c} -z \times \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

$$= \frac{1}{\sqrt{2\pi}} \left[e^{-\frac{z^2}{2}} \right]_{-\infty}^{-c}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{(-c)^2}{2}}$$

$$= \varphi(-c)$$

Finalement $\mathbb{E}(Z) = \varphi(-c)$ et on obtient :

$$\mathbb{E}(Z \mid Z > c) = \frac{\varphi(-c)}{\Phi(-c)}$$

- Si $X \rightsquigarrow \mathcal{N}(m, \sigma^2)$ alors d'après le théorème central limite :

$$Z = \frac{X - m}{\sigma} \rightsquigarrow \mathcal{N}(0, 1)$$

On a $Z = \frac{X-m}{\sigma}$ donc $X = \sigma Z + m$

Donc :

$$\mathbb{E}(X \mid X > c) = \sigma \mathbb{E}(Z \mid X > c) + m$$

$$= \sigma \mathbb{E}(Z \mid \sigma Z + m > c) + m$$

$$= \sigma \mathbb{E}\left(Z \mid Z > \frac{c-m}{\sigma}\right) + m$$

$$\mathbb{E}(X \mid X > c) = \sigma \frac{\varphi\left(\frac{-c+m}{\sigma}\right)}{\Phi\left(\frac{-c+m}{\sigma}\right)} + m$$

Finalement si $Z \rightsquigarrow \mathcal{N}(m\ \sigma^2)$ alors :

$$\mathbb{E}(Z \mid Z > c) = \sigma \frac{\varphi\left(\frac{-c+m}{\sigma}\right)}{\Phi\left(\frac{-c+m}{\sigma}\right)} + m$$