MINI - PROJET: ECONOMETRIE DES VARIABLES QUALITATIVES

Combla Sunday Ezechiel

2022-08-04

Contents

0.1 INTRODUCTION

Le dataset (actuaNV1.csv) contient les données de 2765 individus enregistrées par une société d'assurance IARD (assurance non vie), assurance auto qui veut modéliser la survenance des accidents. Quinze variables sont mesurées, comme le Coût du sinistre,l'Age du véhicule,le nombre de chauffeurs du véhicule, la Durée du la signature du contrat, etc.... Les informations relatives à la variable *accident* sont consignées dans la variable *nbre* qui devra être binarisée ou dichotomiser et nommée « accident ».

0.2 IMPORTATION DU JEU DE DONNEES ET PREPARATION DE LA BASE DE DONNEES

Affichage des six premières lignes

```
setwd("C:/Users/DELL/Documents/R/Base de données")
actuarNV1<-read.table("actuarNV1.csv",header=TRUE,sep=";",check.names=FALSE)
head(actuarNV1)
```

```
##   nocontrat exposition zone puissance agevehicule ageconducteur bonus marque
## 1       217       0.74    A         5           4            31    64      3
## 2       709       0.18    B         7           8            22   100      2
## 3       714       0.48    C         9           0            32    61     12
## 4       852       0.27    F         7           5            39   100     12
## 5      1083       0.51    E         4           0            49    50     12
## 6      1545       0.64    D        10           0            58    50     12
##   carburant densite region nbre    no garantie   cout
## 1         D      21      8    1 17001      1RC   0.00
## 2         E      26      0    1 17419      1RC   0.00
## 3         E      41     13    1 15851      4BG 687.82
## 4         E      11      0    1 21407      2DO  96.64
## 5         E      31     13    1 15589      2DO  70.88
## 6         D      72     13    2   772      1RC   0.00
```

Description du jeu de données

```
str(actuarNV1)
```

```
## 'data.frame':    2765 obs. of  15 variables:
##  $ nocontrat    : int  217 709 714 852 1083 1545 1545 1870 1870 1963 ...
##  $ exposition   : num  0.74 0.18 0.48 0.27 0.51 0.64 0.64 0.11 0.11 0.1 ...
##  $ zone         : chr  "A" "B" "C" "F" ...
##  $ puissance    : int  5 7 9 7 4 10 10 5 5 9 ...
##  $ agevehicule  : int  4 8 0 5 0 0 0 0 0 0 ...
##  $ ageconducteur: int  31 22 32 39 49 58 58 52 52 78 ...
##  $ bonus        : int  64 100 61 100 50 50 50 50 50 50 ...
##  $ marque       : int  3 2 12 12 12 12 12 12 12 12 ...
##  $ carburant    : chr  "D" "E" "E" "E" ...
##  $ densite      : int  21 26 41 11 31 72 72 73 73 72 ...
##  $ region       : int  8 0 13 0 13 13 13 13 13 13 ...
##  $ nbre         : int  1 1 1 1 1 2 2 2 2 1 ...
##  $ no           : int  17001 17419 15851 21407 15589 772 762 17219 17111 16336 ...
##  $ garantie     : chr  "1RC" "1RC" "4BG" "2DO" ...
##  $ cout         : num  0 0 687.8 96.6 70.9 ...
```

Quelques statistiques de base Notre jeu de données contient à la fois des variables qualitatives et quantitatives.

```
summary(actuarNV1)
```

```
##    nocontrat         exposition          zone             puissance
##  Min.   :    217   Min.   :0.008219   Length:2765        Min.   : 4.000
##  1st Qu.: 108786   1st Qu.:0.500000   Class :character   1st Qu.: 5.000
##  Median :1041406   Median :0.870000   Mode  :character   Median : 6.000
##  Mean   : 778477   Mean   :0.737445                      Mean   : 6.374
##  3rd Qu.:1128643   3rd Qu.:1.000000                      3rd Qu.: 7.000
##  Max.   :2054297   Max.   :1.300000                      Max.   :15.000
##   agevehicule     ageconducteur       bonus            marque
##  Min.   : 0.000   Min.   :18.00   Min.   : 50.00   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.:33.00   1st Qu.: 50.00   1st Qu.: 2.000
##  Median : 5.000   Median :43.00   Median : 50.00   Median : 2.000
##  Mean   : 6.209   Mean   :44.09   Mean   : 61.21   Mean   : 4.437
##  3rd Qu.: 9.000   3rd Qu.:53.00   3rd Qu.: 68.00   3rd Qu.: 6.000
##  Max.   :35.000   Max.   :99.00   Max.   :165.00   Max.   :14.000
##   carburant            densite          region            nbre
##  Length:2765        Min.   :11.00   Min.   :-1.000   Min.   :1.000
##  Class :character   1st Qu.:24.00   1st Qu.: 7.000   1st Qu.:1.000
##  Mode  :character   Median :52.00   Median :13.000   Median :1.000
##                     Mean   :49.11   Mean   : 9.998   Mean   :1.712
##                     3rd Qu.:82.00   3rd Qu.:13.000   3rd Qu.:2.000
##                     Max.   :94.00   Max.   :13.000   Max.   :7.000
##        no           garantie              cout
##  Min.   :   148   Length:2765        Min.   : -3811.2
##  1st Qu.: 28810   Class :character   1st Qu.:   132.7
##  Median : 38506   Mode  :character   Median :   405.6
##  Mean   : 44531                      Mean   :  1069.3
##  3rd Qu.: 49055                      3rd Qu.:  1128.1
##  Max.   :104152                      Max.   :152449.0
```

Identification du nombre d'individus ayant des données manquantes

```
actuarNV1[!complete.cases(actuarNV1),]
```

```
##  [1] nocontrat     exposition    zone          puissance     agevehicule
##  [6] ageconducteur bonus         marque        carburant     densite
## [11] region        nbre          no            garantie      cout
## <0 lignes> (ou 'row.names' de longueur nulle)
```

```
nrow(actuarNV1[!complete.cases(actuarNV1),])
```

```
## [1] 0
```

Détection d'éventuelles valeurs maquantes

```
which(is.na(actuarNV1),arr.ind=TRUE)
```

```
##      row col
```

La base ne contient pas de valeurs manquantes

0.3 MODELISONS PAR REGRESSION LOGISTIQUE LA PROBABILITE DE SURVENANCE DE L'ACCIDENT EN FONCTION DU PROFIL DE L'ASSURE. (La variable – cible « nombre » sera binarisée ou dichotomisée et nommée « accident » pour nos besoins d'analyse.)

Nous allons procéder à une régression logistique binaire qui est la méthode appliquée pour les variables d'intérêt qualitatives, nominales et à deux modalités. Dans ce cas de figure, une modalité de la variable d'intérêt sera choisie comme référence. Les odds ratio seront donc exprimés par rapport à cette dernière.

```
actuarNV1$nbre<-as.numeric(actuarNV1$nbre)
```

0.3.1 Creation de la colonne dichotomique accident à partir de la colonne nbre de conducteurs du véhicule

Nous partons de l'hypothèse que pour un nombre maximum de 7 chauffeurs on peut estimer à 0 le nombre d'accidents pour tous les véhicules avec un maximum de 3 chauffeurs, etpour tous les autres nos estimons à 1 ce nombre. on crée ainsi une nouvele colonne accident à partir des valeurs del la colonne nbre.

```
accident<-ifelse(actuarNV1$nbre > "3", "1", "0")
```

Soit actuarNV2 la nouvelle base obtenue

```
actuarNV2<-cbind(actuarNV1,accident)
head(actuarNV2)
```

```
##   nocontrat exposition zone puissance agevehicule ageconducteur bonus marque
## 1       217       0.74    A         5           4            31    64      3
## 2       709       0.18    B         7           8            22   100      2
## 3       714       0.48    C         9           0            32    61     12
## 4       852       0.27    F         7           5            39   100     12
## 5      1083       0.51    E         4           0            49    50     12
## 6      1545       0.64    D        10           0            58    50     12
##   carburant densite region nbre    no garantie   cout accident
## 1         D      21      8    1 17001      1RC   0.00        0
## 2         E      26      0    1 17419      1RC   0.00        0
## 3         E      41     13    1 15851      4BG 687.82        0
## 4         E      11      0    1 21407      2DO  96.64        0
## 5         E      31     13    1 15589      2DO  70.88        0
## 6         D      72     13    2   772      1RC   0.00        0
```

```
nrow(actuarNV2)
```

```
## [1] 2765
```

suppression de la variable nbre

```
library(dplyr)
jeusansnbre = select(actuarNV2, -12)
```

0.3.2 Normalisons les variables de notre jeu de données excepté la variable accident

extraction de la variable accident

```
jeusansnbre1 = select(actuarNV2, -16)
```

Extraction des variables qualitatives

zone, carburant, garantie et region

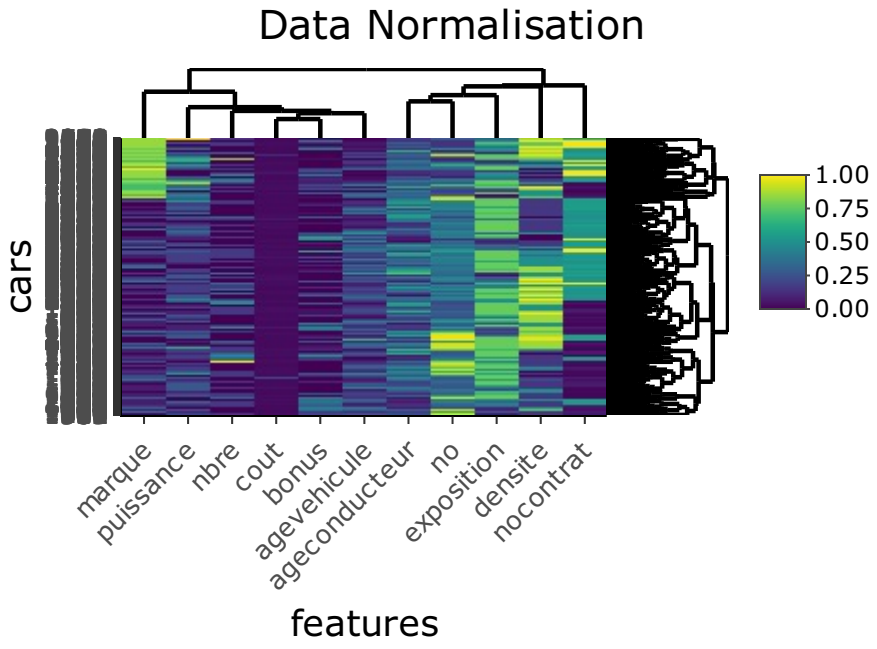
```
jeusansqali = select(jeusansnbre1, -3,-9,-11,-14)
```

Utilisation de la fonction de normalisation Min-Max

Lorsque les variables des données proviennent de distributions éventuellement différentes (et non normales), d'autres transformations peuvent être nécessaires. Une autre possibilité consiste à normaliser les variables pour amener les données sur l'échelle de 0 à 1 en soustrayant le minimum et en divisant par le maximum de toutes les observations. Cela préserve la forme de la distribution de chaque variable tout en les rendant facilement comparables sur la même "échelle".

```
library(heatmaply)
```

```
heatmaply(
  normalize(jeusansqali),
  xlab = "features",
  ylab = "cars",
  main = "Data Normalisation"
)
```



0.3.3 Ramenons les variables qualitatives et la cible

0.3.4 Construction une nouvelle base

Rajout de variables dans la base NVB

NVB pour Nouvelle Base

```
zone<-select(actuarNV2,zone)
carburant<-select(actuarNV2,carburant)
region<-select(actuarNV2,region)
garantie<-select(actuarNV2,garantie)
accident<-select(actuarNV2,accident)
NVB<-cbind(jeusansqali,zone,carburant,region,garantie,accident)
```

```
View(NVB)
```

```
library(dplyr)
```

```
NVB <-select(NVB,-9)
```

0.3.4.1 Suppresssion de la variable nbre

0.3.5 Scindons ensuite la base en échantillons d'apprentissage (avec les 2700 premières lignes) et en échantillons de test(à partir des 65 lignes restantes)

NVBaprt(nouvelle base de données d'apprentissage) NVBtest(nouvelle base de données de test)

```
NVBaprt<-NVB[1:2700, ]
```

```
NVBtest<-NVB[2701:2765, ]
```

```
NVBtest<-as.data.frame(NVBtest)
NVBaprt<-as.data.frame(NVBaprt)
```

```
str(accident)
```

0.3.5.1 Une des manières de sélectionner les N premières lignes d'un bloc de données consiste à utiliser la syntaxe d'indexation de base R :

```
## 'data.frame':    2765 obs. of  1 variable:
##  $ accident: chr  "0" "0" "0" "0" ...
```

```
head(NVBtest)
```

```
##      nocontrat exposition puissance agevehicule ageconducteur bonus marque
## 2701   2028214       0.21         7           0            36    72     12
## 2702   2028229       0.47         4           0            55    50     12
## 2703   2028229       0.47         4           0            55    50     12
## 2704   2029445       0.35         6          10            56    50     12
## 2705   2029445       0.35         6          10            56    50     12
## 2706   2035885       0.40         6           0            34   100     12
##      densite    no    cout zone carburant region garantie accident
## 2701      43 15643 1580.48    C         D      6      2DO        0
## 2702      91 16171  876.87    D         D     13      1RC        0
## 2703      91 16169  710.79    D         D     13      2DO        0
## 2704      93 16217    0.00    C         E     13      1RC        0
## 2705      93 17092  813.93    C         E     13      1RC        0
## 2706      11 16186 1749.16    E         D      0      2DO        0
```

```
str(NVBtest)
```

```
## 'data.frame':    65 obs. of  15 variables:
##  $ nocontrat    : int  2028214 2028229 2028229 2029445 2029445 2035885 2036077 2036127 2036127 20361
##  $ exposition   : num  0.21 0.47 0.47 0.35 0.35 0.4 0.32 0.07 0.07 0.07 ...
##  $ puissance    : int  7 4 4 6 6 6 6 11 11 11 ...
##  $ agevehicule  : int  0 0 0 10 10 0 0 0 0 0 ...
##  $ ageconducteur: int  36 55 55 56 56 34 74 36 36 36 ...
##  $ bonus        : int  72 50 50 50 50 100 73 50 50 50 ...
##  $ marque       : int  12 12 12 12 12 12 12 12 12 12 ...
##  $ densite      : int  43 91 91 93 93 11 31 74 74 74 ...
##  $ no           : int  15643 16171 16169 16217 17092 16186 16933 15942 15943 15941 ...
##  $ cout         : num  1580 877 711 0 814 ...
##  $ zone         : chr  "C" "D" "D" "C" ...
##  $ carburant    : chr  "D" "D" "D" "E" ...
##  $ region       : int  6 13 13 13 13 0 5 13 13 13 ...
##  $ garantie     : chr  "2DO" "1RC" "2DO" "1RC" ...
##  $ accident     : chr  "0" "0" "0" "0" ...
```

0.3.6 Construction du modèle

On considère le modèle logistique (identifiable) permettant d'expliquer une variable binaire Y par p variables X1.

Dans le cadre d'un modèle logistique, généralement on ne présente pas les coefficients du modèle mais leur valeur exponentielle, cette dernière correspondant en effet à des odds ratio, également appelés rapports des cotes. L'odds ratio diffère du risque relatif. Cependent son interprétation est similaire. Un odds ratio de 1 signifie l'absence d'effet. Un odds ratio largement supérieur à 1 correspond à une augmentation du phénomène étudié et un odds ratio largement inféieur à 1 correspond à une diminution du phénomène étudié.

Transformation de la variable accident en facteur

```
NVBaprt$accident <- factor(NVBaprt$accident)
```

Nous choisissons la modalité 0 accident comme référence.

```
NVBaprt$accident <- relevel(NVBaprt$accident, "0")
```

```
levels(NVBaprt$accident)
```

```
## [1] "0" "1"
```

La modalité 0 accident a effectivement été prise comme référence.

```
Reg <- glm(accident ~ bonus + marque + densite+nocontrat+ no
           + cout+zone+carburant+region+garantie+puissance+agevehicule+exposition
           +ageconducteur,
           data = NVBaprt,family ="binomial")
```

En effet, le modèle est identifiable car dans notre cas il est possible d''apprendre la vraie valeur des paramètres à partir d"un nombre infini d'observations.

```
summary(Reg)
```

```
##
## Call:
## glm(formula = accident ~ bonus + marque + densite + nocontrat +
##     no + cout + zone + carburant + region + garantie + puissance +
##     agevehicule + exposition + ageconducteur, family = "binomial",
##     data = NVBaprt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3793  -0.3734  -0.2608  -0.1733   3.0712
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.423e+00  1.651e+00  -0.862  0.38891
## bonus         -1.037e-02  1.577e-02  -0.658  0.51083
## marque         6.302e-02  2.361e-02   2.669  0.00760 **
## densite        8.140e-05  3.236e-03   0.025  0.97993
## nocontrat     -3.060e-07  1.496e-07  -2.046  0.04076 *
## no            -1.034e-05  4.177e-06  -2.475  0.01332 *
## cout           1.548e-06  2.258e-05   0.069  0.94535
## zoneB          5.342e-01  3.799e-01   1.406  0.15969
## zoneC          4.726e-01  3.221e-01   1.467  0.14238
## zoneD          7.591e-01  3.245e-01   2.340  0.01930 *
## zoneE          6.795e-01  3.341e-01   2.034  0.04199 *
## zoneF          5.606e-02  6.519e-01   0.086  0.93146
## carburantE    -9.856e-01  1.922e-01  -5.129 2.92e-07 ***
## region        -1.027e-01  6.303e-02  -1.630  0.10310
## garantie2DO    1.046e-01  2.020e-01   0.518  0.60465
## garantie3VI    3.171e-01  3.152e-01   1.006  0.31439
## garantie4BG   -1.133e+00  2.883e-01  -3.931 8.47e-05 ***
## garantie5CO    3.618e+00  7.791e-01   4.644 3.42e-06 ***
## garantie6CL   -1.140e+01  5.077e+02  -0.022  0.98208
## puissance      2.951e-02  4.813e-02   0.613  0.53976
## agevehicule   -6.588e-02  2.381e-02  -2.767  0.00566 **
## exposition     1.413e+00  3.425e-01   4.127 3.68e-05 ***
## ageconducteur -8.240e-03  7.394e-03  -1.114  0.26508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1203.5  on 2699  degrees of freedom
## Residual deviance: 1055.1  on 2677  degrees of freedom
## AIC: 1101.1
##
## Number of Fisher Scoring iterations: 13
```

On peut remarquer une forte contribution des varibles exposition,agevehicule,garantie,carburant, zone et marques à la réalisaton des valeurs de la cible (accident).

```
library(gtsummary)
tbl_regression(Reg, exponentiate = TRUE)
```

Characteristic	OR	95% CI	p-value
bonus	0.99	0.96, 1.02	0.5
marque	1.07	1.02, 1.12	0.008
densite	1.00	0.99, 1.01	>0.9
nocontrat	1.00	1.00, 1.00	0.041
no	1.00	1.00, 1.00	0.013
cout	1.00	1.00, 1.00	>0.9
zone			
A			
B	1.71	0.81, 3.63	0.2
C	1.60	0.87, 3.11	0.14
D	2.14	1.16, 4.15	0.019
E	1.97	1.04, 3.90	0.042
F	1.06	0.25, 3.46	>0.9
carburant			
D			
E	0.37	0.25, 0.54	<0.001
region	0.90	0.79, 1.02	0.10
garantie			
1RC			
2DO	1.11	0.75, 1.65	0.6
3VI	1.37	0.72, 2.49	0.3
4BG	0.32	0.18, 0.55	<0.001
5CO	37.3	8.01, 181	<0.001
6CL	0.00		>0.9
puissance	1.03	0.93, 1.13	0.5
agevehicule	0.94	0.89, 0.98	0.006
exposition	4.11	2.14, 8.21	<0.001
ageconducteur	0.99	0.98, 1.01	0.3

Selon les résultats de notre modèle, la survenance des accidents se fait avec l'importance de la densité de la population; on remarque aussi que les véhicules utilisant le type de carburant E ont tandence à enregistrer plus d'accidents.on constate par ailleurs que le nombre d'accidents est relativement important dans la zone D, le même constat est fait pour les garanties de type 5CO et 3VI.En outre, la puissance du moteur, les ages des vehicules et des chauffeurs ainsi que l'exposition (Durée du la signature du contrat) constituent des variables avec une influence remarquable sur la survenance de l'accident.

0.3.7 Construction de modèle sur la base des variables a forte contribution

```
Reg1 <- glm(accident ~ exposition+agevehicule+marque+zone+carburant+garantie,
            data = NVBaprt,family ="binomial")
```

```
summary(Reg1)
```

```
##
## Call:
```

```
## glm(formula = accident ~ exposition + agevehicule + marque +
##     zone + carburant + garantie, family = "binomial", data = NVBaprt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4995  -0.3768  -0.2816  -0.1904   2.8667
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.76973    0.43822  -8.602  < 2e-16 ***
## exposition    1.19001    0.32729   3.636 0.000277 ***
## agevehicule  -0.06235    0.02309  -2.700 0.006938 **
## marque        0.06916    0.02035   3.399 0.000675 ***
## zoneB         0.58071    0.37662   1.542 0.123100
## zoneC         0.55759    0.31639   1.762 0.078014 .
## zoneD         0.84087    0.31624   2.659 0.007838 **
## zoneE         0.73039    0.32465   2.250 0.024462 *
## zoneF         0.09488    0.64476   0.147 0.883014
## carburantE   -0.98175    0.18792  -5.224 1.75e-07 ***
## garantie2DO   0.08870    0.19816   0.448 0.654434
## garantie3VI   0.24518    0.30963   0.792 0.428449
## garantie4BG  -1.17823    0.28435  -4.144 3.42e-05 ***
## garantie5CO   3.46011    0.73363   4.716 2.40e-06 ***
## garantie6CL -11.65323  506.17959  -0.023 0.981633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1203.5  on 2699  degrees of freedom
## Residual deviance: 1083.9  on 2685  degrees of freedom
## AIC: 1113.9
##
## Number of Fisher Scoring iterations: 13
```

0.3.8 Modèle avec interactions

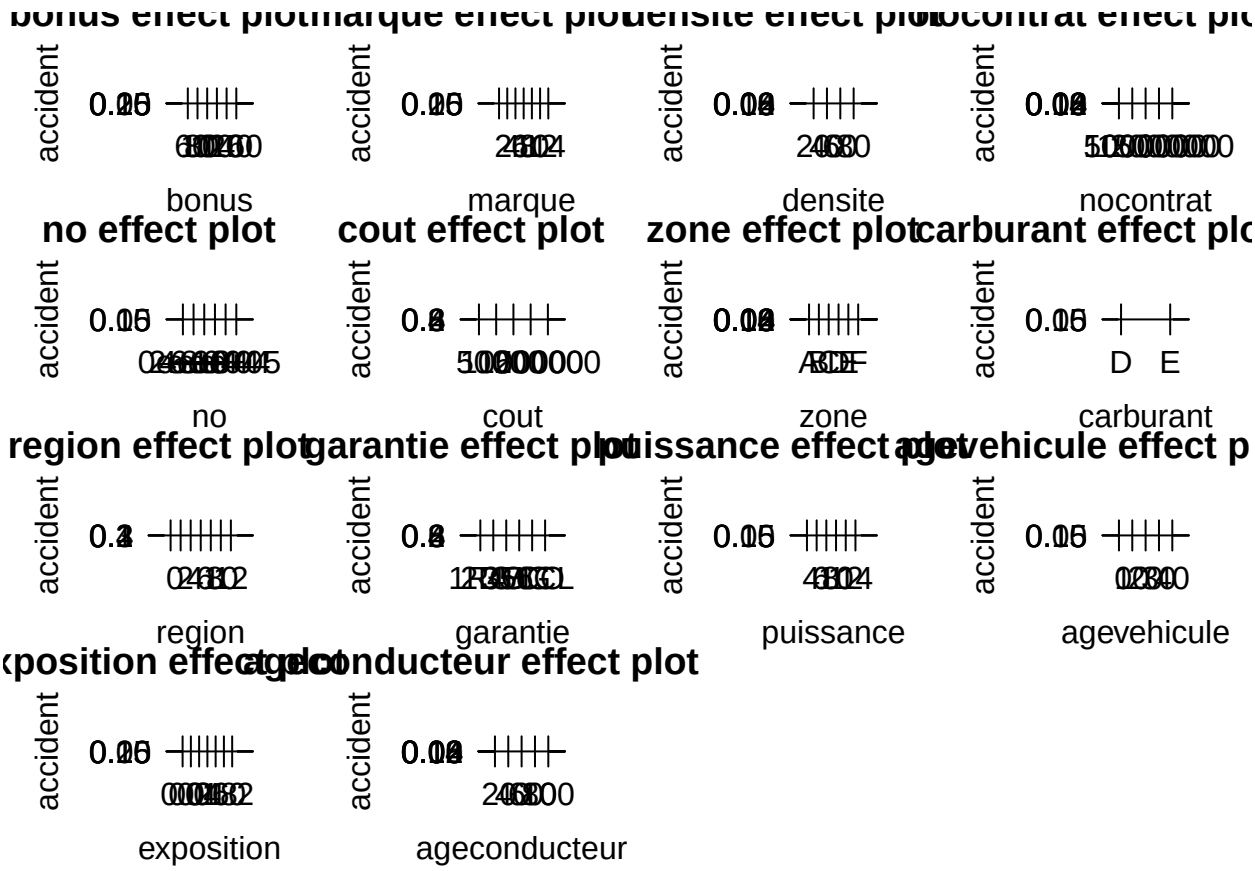
Dans un modèle statistique classique, on fait l'hypothèse implicite que chaque variable explicative est indépendante des autres. Cependant, cela ne se vérifie pas toujours. Par exemple, l'effet de l'âge peut varier en fonction du sexe. Il est dès lors nécessaire de prendre en compte dans son modèle les effets d'interaction

Pour représenter les effets différentes variables, on peut avoir recours à la fonction allEffects de l'extension effects.

Test d'Effets d'interaction dans le modèle Reg

```
library(effects)
```

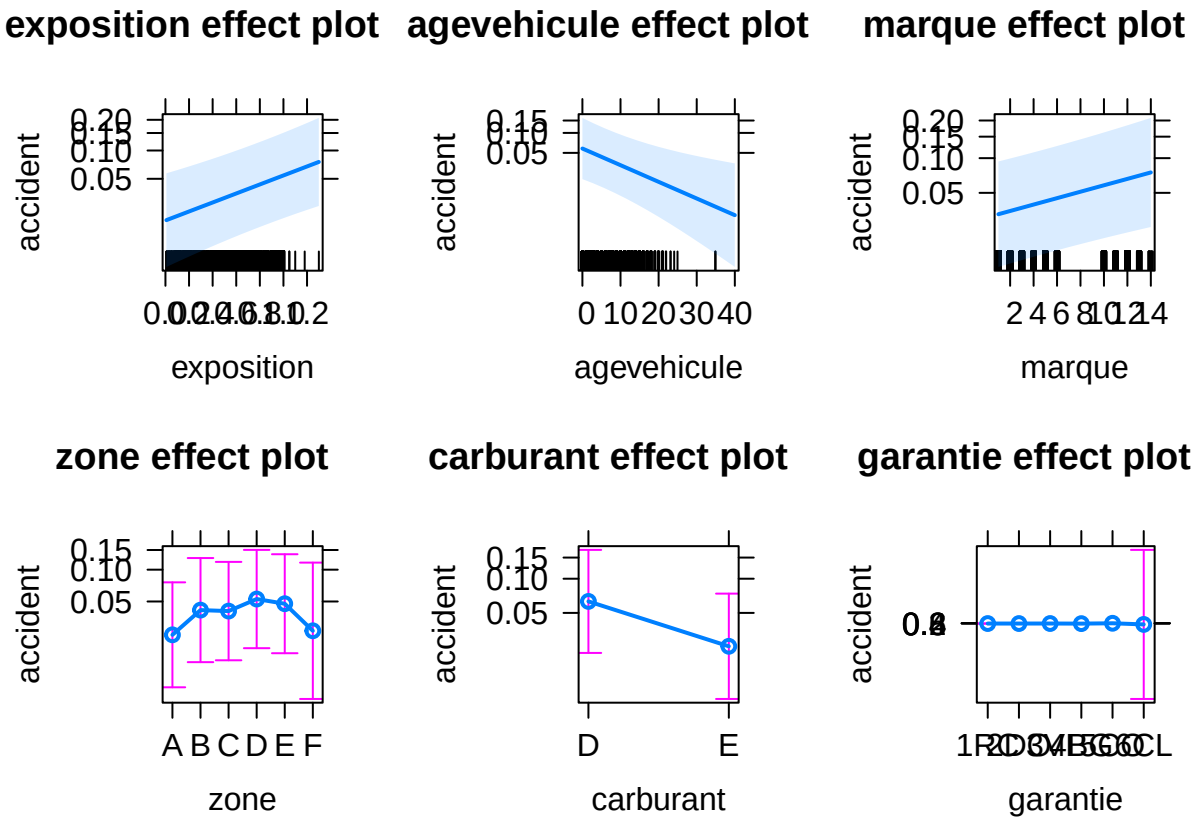
```
plot(allEffects(Reg))
```



On peut conclure à une faible interraction (ou absence d'effet d'interraction) entre les différentes variables explicatives du modèle Reg.

Test d'Effets d'interaction dans le modèle Reg1

```
plot(allEffects(Reg1))
```



En observant le graphique plusieurs hypothèses nous viennent à l'esprit parmis lesquelles celle de savoir si l'exposition est ou ou n'est pas en interraction avec l'age du véhicule ? Nous allons donc introduire une interaction entre l'âge du véhicule et l'exposition dans notre modèle (Reg1), ce qui sera représenté par exposition * agevehicule dans l'équation du modèle.

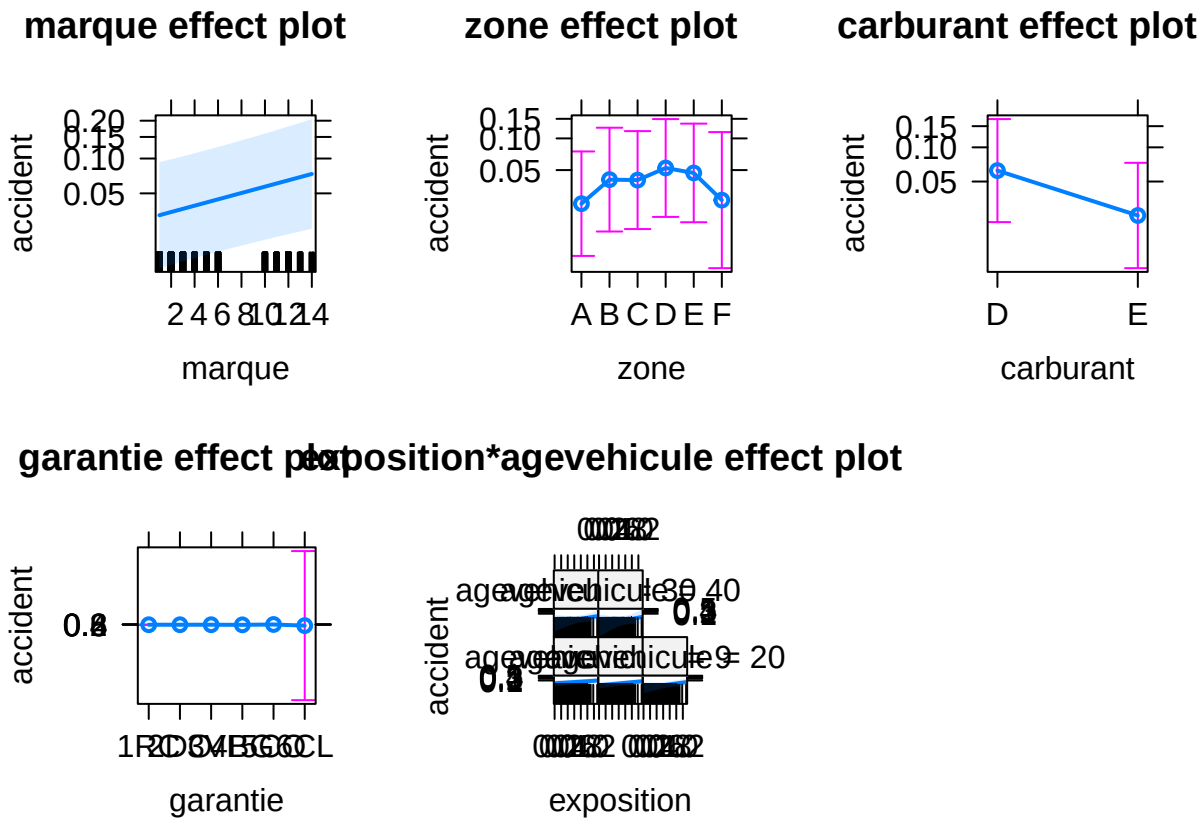
```
Reg1x <- glm(accident ~ exposition*agevehicule+marque+zone+carburant+garantie,
             data = NVBaprt,family ="binomial")
tbl_regression(Reg1x, exponentiate = TRUE)
```

Characteristic	**OR**	**95% CI**	**p-value**
exposition	2.64	1.05, 6.91	0.043
agevehicule	0.90	0.78, 1.03	0.14
marque	1.07	1.03, 1.11	<0.001
zone			
A			
B	1.76	0.84, 3.73	0.13
C	1.75	0.96, 3.34	0.078
D	2.32	1.28, 4.45	0.008
E	2.07	1.11, 4.02	0.025
F	1.09	0.26, 3.49	0.9
carburant			
D			
E	0.38	0.26, 0.54	<0.001
garantie			

Characteristic	OR	95% CI	p-value
1RC			
2DO	1.09	0.74, 1.61	0.7
3VI	1.28	0.68, 2.29	0.4
4BG	0.31	0.17, 0.52	<0.001
5CO	31.7	7.43, 143	<0.001
6CL	0.00		>0.9
exposition * agevehicule	1.05	0.90, 1.25	0.5

Observons les effets du modèle.

```
plot(allEffects(Reg1x))
```



Sur ce graphique, on voit que l'effet de l'exposition sur l'âge du vehicule est différent selon que l'on soit en présence de vieux vehicule ou de vehicules recents.

On peut tester si l'ajout de l'interaction améliore significativement le modèle avec anova

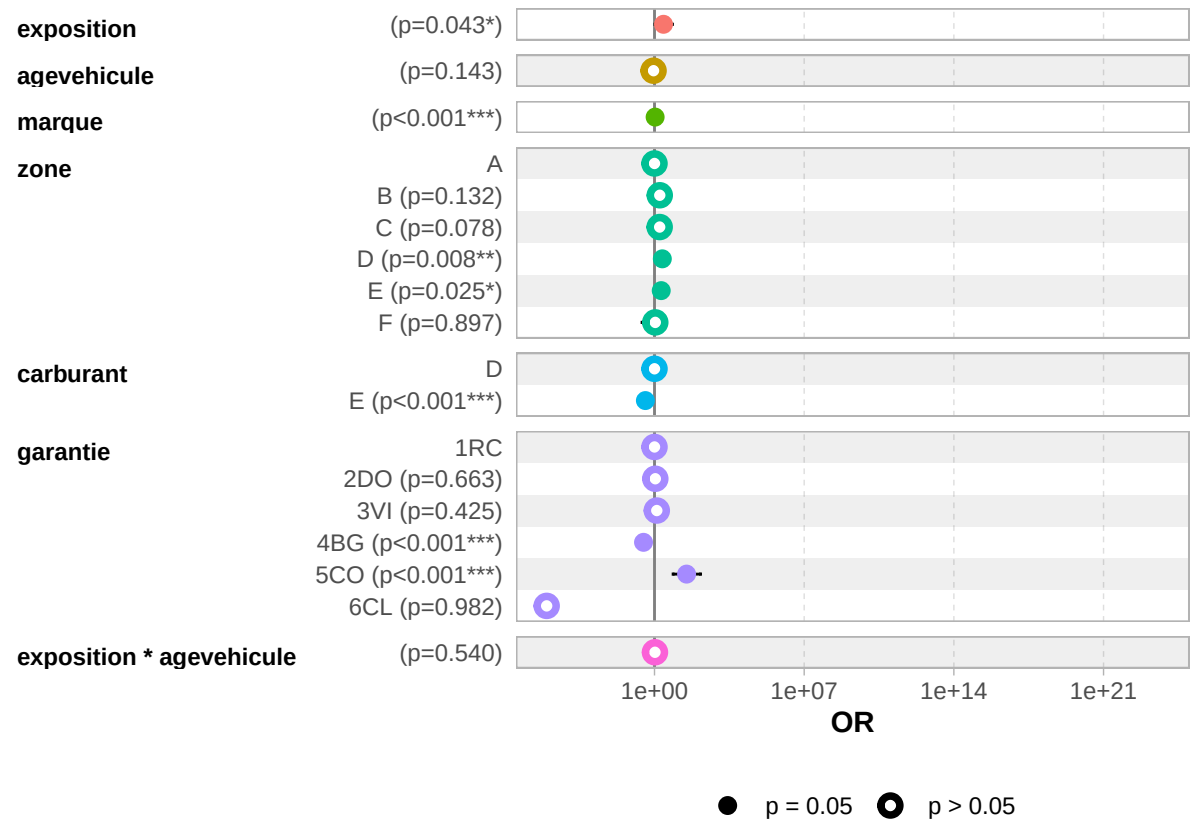
```
anova(Reg1x, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
```

```
##
## Response: accident
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                      2699     1203.5
## exposition                1    4.380      2698     1199.1  0.036371 *
## agevehicule               1   23.934      2697     1175.2 9.968e-07 ***
## marque                    1   10.086      2696     1165.1  0.001494 **
## zone                      5    7.346      2691     1157.8  0.196179
## carburant                 1   24.269      2690     1133.5 8.376e-07 ***
## garantie                  5   49.626      2685     1083.9 1.653e-09 ***
## exposition:agevehicule    1    0.385      2684     1083.5  0.535161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Jetons maintenant un oeil aux coefficients du modèle. Pour rendre les choses plus visuelles, nous aurons recours à ggcoef_model de l'extension GGally.

```
library(GGally)
ggcoef_model(Reg1x, exponentiate = TRUE)
```



Représentation graphique des coefficients du modèle avec interaction entre l'âge du véhicule et l'exposition

Concernant l'âge du véhicule et l'exposition, nous avons un coefficient avec une probabilité de 0.54 avec un *Odds Ratio*=1 qui correspondent à l'effet global de la variable âge du véhicule.p=0,540Les indique que l'odd ratio n'est pas significativement différent de 1. Cela signifie l'absence d'effet (ou faible effet) de la variable âge du véhicule sur l'exposition.

0.3.9 Matrice de confusion

Une manière de tester la qualité d'un modèle est le calcul d'une matrice de confusion, c'est-à-dire le tableau croisé des valeurs observées et celles des valeurs prédites en appliquant le modèle aux données d'origine.

La méthode predict avec l'argument type="response" permet d'appliquer notre modèle logistique à un tableau de données et renvoie pour chaque individu la probabilité qu'il ait vécu le phénomène étudié.

```
library(stats)
```

```
accident.pred1 <- predict(Reg, type = "response", newdata = NVBtest)
head(accident.pred1)
```

```
##       2701       2702       2703       2704       2705       2706
## 0.11688610 0.09871174 0.10839341 0.01387848 0.01377232 0.22373973
```

```
accident.pred2 <- predict(Reg1, type = "response", newdata = NVBtest)
head(accident.pred2)
```

```
##       2701       2702       2703       2704       2705       2706
## 0.11470349 0.17659571 0.18986520 0.02736132 0.02736132 0.16183107
```

```
accident.pred3 <- predict(Reg1x, type = "response", newdata = NVBtest)
head(accident.pred3)
```

```
##       2701       2702       2703       2704       2705       2706
## 0.12655758 0.18534075 0.19875376 0.02391555 0.02391555 0.17099022
```

Or notre variable étudiée est de type binaire. Usuellement, les probabilités prédites seront réunies en deux groupes selon qu'elles soient supérieures ou inférieures à la moitié. La matrice de confusion est alors égale à :

```
table(accident.pred1 > 0.5, NVBtest$accident)
```

```
##
##          0  1
##   FALSE 62  3
```

```
table(accident.pred2 > 0.5, NVBtest$accident)
```

```
##
##          0  1
##   FALSE 62  3
```

```
table(accident.pred3 > 0.5, NVBtest$accident)
```

```
##
##          0  1
##   FALSE 62  3
```

Pour tous les trois modeles,nous avons donc 3 prédictions incorrectes sur un total de 65, soit un taux de mauvais classement de 04,62 %.

0.3.10 Sélection de modèles

Il est toujours tentant lorsque l'on recherche les facteurs associés à un phénomène d'inclure un nombre important de variables explicatives potentielles dans un mmodèle logistique. Cependant, un tel modèle n'est pas forcément le plus efficace et certaines variables n'auront probablement pas d'effet significatif sur la variable d'intérêt. La technique de sélection descendante pas à pas est une approche visant à améliorer son modèle explicatif Il faut également définir un critère pour déterminer la qualité d'un modèle. L'un des plus utilisés est le Akaike Information Criterion ou AIC. Plus l'AIC sera faible, meilleure sera le modèle.

Au terme de ces analyses,la fonction step permet de sélectionner le meilleur modèle par une procédure pas à pas descendante basée sur la minimisation de l'AIC. La fonction affiche à l'écran les différentes étapes de la sélection et renvoie le modèle final.

```
step1 <- step(Reg)
```

```
## Start:  AIC=1101.08
## accident ~ bonus + marque + densite + nocontrat + no + cout +
##     zone + carburant + region + garantie + puissance + agevehicule +
##     exposition + ageconducteur
##
##                 Df Deviance    AIC
## - zone           5   1062.2 1098.2
## - densite        1   1055.1 1099.1
## - cout           1   1055.1 1099.1
## - puissance      1   1055.5 1099.5
## - bonus          1   1055.5 1099.5
## - ageconducteur  1   1056.3 1100.3
## <none>               1055.1 1101.1
## - region         1   1057.9 1101.9
## - nocontrat      1   1059.3 1103.3
## - no             1   1061.8 1105.8
## - marque         1   1062.0 1106.0
## - agevehicule    1   1063.3 1107.3
## - exposition     1   1074.2 1118.2
## - carburant      1   1083.8 1127.8
## - garantie       5   1102.2 1138.2
##
## Step:  AIC=1098.23
## accident ~ bonus + marque + densite + nocontrat + no + cout +
##     carburant + region + garantie + puissance + agevehicule +
##     exposition + ageconducteur
##
##                 Df Deviance    AIC
```

```
## - cout           1   1062.2 1096.2
## - puissance      1   1062.5 1096.5
## - bonus          1   1062.6 1096.6
## - densite        1   1062.8 1096.8
## - ageconducteur  1   1063.3 1097.3
## <none>               1062.2 1098.2
## - region         1   1065.3 1099.3
## - nocontrat      1   1066.7 1100.7
## - no             1   1069.0 1103.0
## - marque         1   1070.0 1104.0
## - agevehicule    1   1070.2 1104.2
## - exposition     1   1081.1 1115.1
## - carburant      1   1090.8 1124.8
## - garantie       5   1109.1 1135.1
##
## Step:  AIC=1096.23
## accident ~ bonus + marque + densite + nocontrat + no + carburant +
##     region + garantie + puissance + agevehicule + exposition +
##     ageconducteur
##
##                 Df Deviance    AIC
## - puissance      1   1062.5 1094.5
## - bonus          1   1062.6 1094.6
## - densite        1   1062.8 1094.8
## - ageconducteur  1   1063.3 1095.3
## <none>               1062.2 1096.2
## - region         1   1065.3 1097.3
## - nocontrat      1   1066.7 1098.7
## - no             1   1069.0 1101.0
## - marque         1   1070.0 1102.0
## - agevehicule    1   1070.2 1102.2
## - exposition     1   1081.2 1113.2
## - carburant      1   1090.8 1122.8
## - garantie       5   1109.4 1133.4
##
## Step:  AIC=1094.51
## accident ~ bonus + marque + densite + nocontrat + no + carburant +
##     region + garantie + agevehicule + exposition + ageconducteur
##
##                 Df Deviance    AIC
## - bonus          1   1062.9 1092.9
## - densite        1   1063.1 1093.1
## - ageconducteur  1   1063.5 1093.5
## <none>               1062.5 1094.5
## - region         1   1065.6 1095.6
## - nocontrat      1   1067.1 1097.1
## - no             1   1069.2 1099.2
## - agevehicule    1   1070.2 1100.2
## - marque         1   1072.2 1102.2
## - exposition     1   1081.7 1111.7
## - carburant      1   1092.0 1122.0
## - garantie       5   1109.4 1131.4
##
## Step:  AIC=1092.91
```

```
## accident ~ marque + densite + nocontrat + no + carburant + region +
##     garantie + agevehicule + exposition + ageconducteur
##
##                 Df Deviance    AIC
## - densite        1   1063.5 1091.5
## - ageconducteur  1   1064.0 1092.0
## <none>               1062.9 1092.9
## - nocontrat      1   1067.5 1095.5
## - no             1   1069.5 1097.5
## - agevehicule    1   1070.7 1098.7
## - marque         1   1072.6 1100.6
## - region         1   1074.4 1102.4
## - exposition     1   1081.8 1109.8
## - carburant      1   1092.4 1120.4
## - garantie       5   1109.4 1129.4
##
## Step:  AIC=1091.54
## accident ~ marque + nocontrat + no + carburant + region + garantie +
##     agevehicule + exposition + ageconducteur
##
##                 Df Deviance    AIC
## - ageconducteur  1   1064.5 1090.5
## <none>               1063.5 1091.5
## - nocontrat      1   1068.1 1094.1
## - no             1   1070.1 1096.1
## - agevehicule    1   1071.3 1097.3
## - marque         1   1073.7 1099.7
## - region         1   1074.8 1100.8
## - exposition     1   1081.9 1107.9
## - carburant      1   1092.8 1118.8
## - garantie       5   1110.5 1128.5
##
## Step:  AIC=1090.49
## accident ~ marque + nocontrat + no + carburant + region + garantie +
##     agevehicule + exposition
##
##               Df Deviance    AIC
## <none>             1064.5 1090.5
## - nocontrat    1   1069.0 1093.0
## - no           1   1071.2 1095.2
## - agevehicule  1   1072.6 1096.6
## - marque       1   1074.1 1098.1
## - exposition   1   1082.2 1106.2
## - region       1   1082.4 1106.4
## - carburant    1   1095.2 1119.2
## - garantie     5   1111.2 1127.2
```

```
step2 <- step(Reg1)
```

```
## Start:  AIC=1113.87
## accident ~ exposition + agevehicule + marque + zone + carburant +
##     garantie
##
##               Df Deviance    AIC
```

```
## - zone         5   1093.2 1113.2
## <none>             1083.9 1113.9
## - agevehicule  1   1091.8 1119.8
## - marque       1   1095.0 1123.0
## - exposition   1   1098.5 1126.5
## - carburant    1   1113.7 1141.7
## - garantie     5   1133.5 1153.5
##
## Step:  AIC=1113.19
## accident ~ exposition + agevehicule + marque + carburant + garantie
##
##               Df Deviance    AIC
## <none>             1093.2 1113.2
## - agevehicule  1   1100.7 1118.7
## - marque       1   1106.1 1124.1
## - exposition   1   1106.6 1124.6
## - carburant    1   1122.0 1140.0
## - garantie     5   1142.6 1152.6
```

```
step3 <- step(Reg1x)
```

```
## Start:  AIC=1115.48
## accident ~ exposition * agevehicule + marque + zone + carburant +
##     garantie
##
##                           Df Deviance    AIC
## - exposition:agevehicule  1   1083.9 1113.9
## - zone                    5   1092.8 1114.8
## <none>                        1083.5 1115.5
## - marque                  1   1094.2 1124.2
## - carburant               1   1113.3 1143.3
## - garantie                5   1133.2 1155.2
##
## Step:  AIC=1113.87
## accident ~ exposition + agevehicule + marque + zone + carburant +
##     garantie
##
##               Df Deviance    AIC
## - zone         5   1093.2 1113.2
## <none>             1083.9 1113.9
## - agevehicule  1   1091.8 1119.8
## - marque       1   1095.0 1123.0
## - exposition   1   1098.5 1126.5
## - carburant    1   1113.7 1141.7
## - garantie     5   1133.5 1153.5
##
## Step:  AIC=1113.19
## accident ~ exposition + agevehicule + marque + carburant + garantie
##
##               Df Deviance    AIC
## <none>             1093.2 1113.2
## - agevehicule  1   1100.7 1118.7
## - marque       1   1106.1 1124.1
## - exposition   1   1106.6 1124.6
```

```
## - carburant    1   1122.0 1140.0
## - garantie     5   1142.6 1152.6
```

Considérons le modèle ci dessous (Regfin)

```
Regfin <- glm(accident ~ marque + no + nocontrat + carburant + region +
                garantie +
    agevehicule + exposition,data = NVBaprt,family ="binomial")
```

Regfin a été choisi comme modèle final car il affiche le plus faible AIC (AIC=1090.49)

0.3.11 Faisons l'anova du modèle

```
print(anova(Regfin, test="Chisq"))
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: accident
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         2699     1203.5
## marque       1   15.225      2698     1188.3 9.545e-05 ***
## no           1    8.815      2697     1179.5 0.0029870 **
## nocontrat    1    4.127      2696     1175.3 0.0421933 *
## carburant    1   26.824      2695     1148.5 2.228e-07 ***
## region       1   14.668      2694     1133.8 0.0001282 ***
## garantie     5   45.016      2689     1088.8 1.440e-08 ***
## agevehicule  1    6.583      2688     1082.2 0.0102959 *
## exposition   1   17.756      2687     1064.5 2.511e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La fonction *coef* permet d'obtenir les coefficients d'un modèle, *confint* leurs intervalles de confiance et *exp* de calculer l'exponentiel. Les odds ratio et leurs intervalles de confiance s'obtiennent ainsi :

Calculons les odds ratios

```
exp(coef(Regfin))
```

```
##  (Intercept)       marque           no    nocontrat   carburantE       region
## 1.461676e-01 1.071373e+00 9.999897e-01 9.999997e-01 3.686979e-01 9.260887e-01
##  garantie2DO  garantie3VI  garantie4BG  garantie5CO  garantie6CL  agevehicule
## 1.080393e+00 1.480896e+00 3.298064e-01 3.154502e+01 1.178084e-05 9.373967e-01
##   exposition
## 3.807708e+00
```

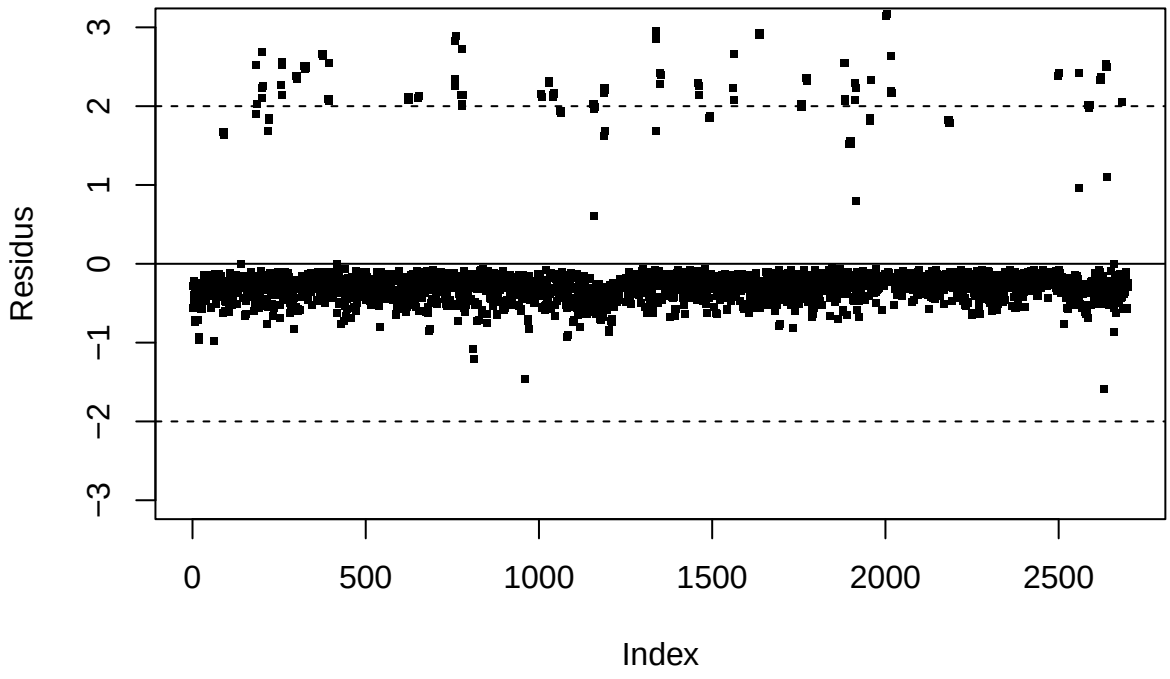
Calculons les intervalles de confiance

```
exp(confint(Regfin))
```

```
##                  2.5 %       97.5 %
## (Intercept) 0.0624661 3.330436e-01
## marque      1.0260795 1.117688e+00
## no          0.9999813 9.999976e-01
## nocontrat   0.9999994 1.000000e+00
## carburantE  0.2525807 5.296401e-01
## region      0.8947403 9.590720e-01
## garantie2DO 0.7290044 1.603024e+00
## garantie3VI 0.7798178 2.663833e+00
## garantie4BG 0.1833195 5.641119e-01
## garantie5CO 7.1747097 1.478655e+02
## garantie6CL        NA 6.479708e+23
## agevehicule 0.8937313 9.806316e-01
## exposition  2.0096079 7.487941e+00
```

0.3.12 Analyse des résidus

```
res.m<-rstudent(Regfin)
plot(res.m,pch=15,cex=.5,ylab="Residus",ylim=c(-3,3))
abline(h=c(-2,0,2),lty=c(2,1,2))
```



En théorie 95% des résidus studentisés se trouvent dans l'intervalle [-2;2].

```
res.m<-rstudent(Regfin)
sum(as.numeric(abs(res.m)<=3))/nrow(NVBaprt)*100
```

```
## [1] 99.85185
```

Ici on a visuellement beaucoup de résidus qui se trouvent dans cet intervalle soit 99.85/100. Ce qui est acceptable.

0.3.13 Estimons le taux de mauvais classement

On calcule d'abord les probabilités prédites sur le modèle sélectionné construit a partir de l'échantillon d'apprentissage et on affiche les 6 premieres valeurs.

```
accident.pred4 <- predict(Regfin,NVBaprt, type = "response")
head(accident.pred4)
```

```
##          1          2          3          4          5          6
## 0.14500323 0.03774028 0.02360979 0.09988104 0.07635454 0.22328358
```

Or notre variable étudiée est de type binaire. Usuellement, les probabilités prédites seront réunies en deux groupes selon qu'elles soient supérieures ou inférieures à la moitié. La matrice de confusion est alors égale à :

```
table(accident.pred4 > 0.5, NVBaprt$accident)
```

```
##
##             0    1
##   FALSE 2539  154
##   TRUE     3    4
```

On calcule le taux de mauvais classement avec l'échantillon d'apprentissage

```
MC<-(154+3)/(2539+4)
MC
```

```
## [1] 0.0617381
```

Soit un taux de mauvais classement de 06,17%

0.3.14 Faisons la prévision

Nous allons faire la prévision avec l'échantillon de test.

```
 NVBtest
```

```
##      nocontrat exposition puissance agevehicule ageconducteur bonus marque
## 2701   2028214       0.21         7           0             36    72     12
## 2702   2028229       0.47         4           0             55    50     12
## 2703   2028229       0.47         4           0             55    50     12
## 2704   2029445       0.35         6          10             56    50     12
## 2705   2029445       0.35         6          10             56    50     12
## 2706   2035885       0.40         6           0             34   100     12
## 2707   2036077       0.32         6           0             74    73     12
## 2708   2036127       0.07        11           0             36    50     12
## 2709   2036127       0.07        11           0             36    50     12
## 2710   2036127       0.07        11           0             36    50     12
## 2711   2038159       0.24         7           0             71    58     12
## 2712   2042109       0.04        12           0             55    50     12
## 2713   2045934       0.38         6           8             67    50      2
## 2714   2046317       1.00         7           7             66    50      2
## 2715   2046317       1.00         7           7             66    50      2
## 2716   2046574       1.00         5           6             72    50      2
## 2717   2046802       0.41         6          14             48    50      1
## 2718   2046802       0.41         6          14             48    50      1
## 2719   2046986       1.00         6           6             48    50      1
## 2720   2047157       0.67         5           6             50    50      2
## 2721   2047295       1.00        12           5             64    50     12
## 2722   2047295       1.00        12           5             64    50     12
## 2723   2047390       0.32         6           5             51    50      2
## 2724   2047422       0.75         6           4             53    50      3
## 2725   2047433       1.00         7           5             83    50      2
## 2726   2047728       1.00         6           9             51    50      1
## 2727   2047995       0.23         6          11             60    50      1
## 2728   2048330       1.00         6           3             48    50     10
## 2729   2048731       0.26         7           7             74    50      2
## 2730   2048888       1.00         5           7             64    50      1
## 2731   2049031       1.00         7          18             59    50      2
## 2732   2049035       0.33         6          13             60    50      2
## 2733   2049159       1.00        10           2             60    62     11
## 2734   2049367       1.00         7           9             43    50      2
## 2735   2049413       1.00         5           8             47    50      3
## 2736   2049584       1.00         4          10             32    50      2
## 2737   2049795       1.00         4          12             80    62      1
## 2738   2049795       1.00         4          12             80    62      1
## 2739   2050230       1.00         6           7             79    50      2
## 2740   2050294       0.24         5          10             51    50      2
## 2741   2050539       1.00         5           9             41    50      1
## 2742   2050663       1.00         4          19             77    50      2
## 2743   2050708       1.00         4          10             42    50      2
## 2744   2050848       0.83         9           1             59    50      2
## 2745   2051208       1.00         6          10             37    57      1
## 2746   2051208       1.00         6          10             37    57      1
## 2747   2051208       1.00         6          10             37    57      1
## 2748   2051383       1.00         6           9             48    50      2
## 2749   2051383       1.00         6           9             48    50      2
## 2750   2051733       1.00         5           5             64    50      6
## 2751   2051740       0.72         7           2             33    83      2
## 2752   2051935       1.00         7          13             36    72      1
## 2753   2051988       0.37         6          15             47    50      6
```

```
## 2754   2052543       0.31         6           9            34    50       2
## 2755   2052551       0.69        10           2            40    50      13
## 2756   2052551       0.69        10           2            40    50      13
## 2757   2052958       1.00         5          11            47    50       2
## 2758   2053013       0.88        10           8            37    50      12
## 2759   2053375       1.00         7          13            33   118       4
## 2760   2053375       1.00         7          13            33   118       4
## 2761   2053424       0.24         7           8            31    54       1
## 2762   2054095       0.91         5           4            32    57       1
## 2763   2054136       0.16         5          17            44    50       2
## 2764   2054184       0.07         5           7            48    50       4
## 2765   2054297       1.00         6           4            45    50       1
##       densite    no    cout zone carburant region garantie accident
## 2701       43 15643 1580.48    C         D      6      2DO        0
## 2702       91 16171  876.87    D         D     13      1RC        0
## 2703       91 16169  710.79    D         D     13      2DO        0
## 2704       93 16217    0.00    C         E     13      1RC        0
## 2705       93 17092  813.93    C         E     13      1RC        0
## 2706       11 16186 1749.16    E         D      0      2DO        0
## 2707       31 16933  556.14    D         D      5      2DO        0
## 2708       74 15942    0.00    A         D     13      1RC        0
## 2709       74 15943  874.65    A         D     13      1RC        0
## 2710       74 15941   69.67    A         D     13      2DO        0
## 2711       25 16883  382.91    C         E     13      2DO        0
## 2712       23 17520  540.94    E         D     13      2DO        0
## 2713       24 65504  272.00    A         D     13      2DO        0
## 2714       24 63984 1001.77    C         E     13      1RC        0
## 2715       24 64144  122.70    C         E     13      2DO        0
## 2716       24 58343 1172.00    C         E     13      4BG        0
## 2717       25 51652    0.00    C         D     13      1RC        0
## 2718       25 51653  108.62    C         D     13      2DO        0
## 2719       24 65565  229.06    C         D     13      4BG        0
## 2720       24 62169    0.00    D         E     13      4BG        0
## 2721       53 58941 1172.00    C         D     13      2DO        0
## 2722       53 58940  359.73    C         D     13      1RC        0
## 2723       24 62102    0.00    A         D     13      3VI        0
## 2724       52 50087  757.68    E         D     13      1RC        0
## 2725       24 52680   67.22    C         E     13      1RC        0
## 2726       24 54029   75.56    A         D     13      1RC        0
## 2727       24 54521  410.73    A         D     13      1RC        0
## 2728       53 64178 4367.81    C         D     13      4BG        0
## 2729       24 57255 2110.67    A         E     13      1RC        0
## 2730       24 55691 1172.00    A         E     13      2DO        0
## 2731       24 67167  125.84    A         D     13      1RC        0
## 2732       24 66983 3446.39    A         D     13      4BG        0
## 2733       53 53763  250.11    E         D     13      2DO        0
## 2734       52 61619  243.34    D         E     13      4BG        0
## 2735       24 50399 1125.86    B         D     13      4BG        0
## 2736       82 56511 1173.95    C         D     13      2DO        0
## 2737       24 54662 1172.00    B         E     13      2DO        0
## 2738       24 54661   37.75    B         E     13      1RC        0
## 2739       24 65499 2140.94    A         D     13      2DO        0
## 2740       24 55445  514.10    B         D     13      2DO        0
## 2741       53 63377 1056.64    C         E     13      2DO        0
```

```
## 2742      53 63455  331.65    C         E     13      1RC        0
## 2743      52 61108 1021.69    C         E     13      1RC        0
## 2744      82 59449   12.36    E         E     13      2DO        0
## 2745      23 54010   63.76    A         D     10      4BG        1
## 2746      23 54033  150.92    A         D     10      1RC        1
## 2747      23 54009   96.05    A         D     10      3VI        1
## 2748      24 53973  698.42    A         E     13      1RC        0
## 2749      24 53972  553.51    A         E     13      3VI        0
## 2750      53 55285 9476.18    D         E     13      3VI        0
## 2751      53 55345   81.33    A         E      3      2DO        0
## 2752      53 55261 1172.00    C         E      6      4BG        0
## 2753      24 57292  744.95    B         E     13      1RC        0
## 2754      24 67147  114.06    B         D     13      4BG        0
## 2755      52 67177  682.28    B         D     13      4BG        0
## 2756      52 66986 1034.89    B         D     13      4BG        0
## 2757      74 65278 1881.46    A         D     13      1RC        0
## 2758      74 65309    0.00    A         D     13      2DO        0
## 2759      52 59261 1172.00    C         E     -1      4BG        0
## 2760      52 59260  464.81    C         E     -1      1RC        0
## 2761      52 59231  156.23    B         E     11      4BG        0
## 2762      24 60566   74.15    C         D     10      1RC        0
## 2763      24 60518  325.57    A         E     13      4BG        0
## 2764      24 60515  387.04    C         D     13      4BG        0
## 2765      24 55667  231.20    A         D     13      4BG        0
```

On calcule d'abord les probabilités prédites sur le modèle sélectionné construit a partir de l'échantillon de Test et on affiche les 6 premieres valeurs.

```
accident.pred5 <- predict(Regfin, type = "response",newdata = NVBtest)
head(accident.pred5)
```

```
##       2701       2702       2703       2704       2705       2706
## 0.11961640 0.09374945 0.10053041 0.01672058 0.01657346 0.21597003
```

Or notre variable étudiée est de type binaire. Usuellement, les probabilités prédites seront réunies en deux groupes selon qu'elles soient supérieures ou inférieures à la moitié. La matrice de confusion est alors égale à :

```
accident.pred5<-as.numeric(accident.pred5>0.5)
accident.pred5
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [39] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

On peut ainsi dresser le tableau de contingence entre les labels prédits et les vrais labels.

```
table(accident.pred5, NVBtest$accident)
```

```
##
## accident.pred5  0  1
##              0 62  3
```

On calcule le taux de mauvais classement avec l'échantillon test

```
MCtest<-(3)/(65)
MCtest
```

```
## [1] 0.04615385
```

Soit un taux de mauvais classement de 04,61%

0.3.14.1 Calcul de l'AUC L'AUC correspond à la probabilité pour qu'un événement positif soit classé comme positif par le test sur l'étendue des valeurs seuil possibles. Pour un modèle idéal, on a AUC=1 (ci-dessus en bleu), pour un modèle aléatoire, on a AUC=0.5 (ci-dessus en rouge). On considère habituellement que le modèle est bon dès lors que la valeur de l'AUC est supérieure à 0.7. Un modèle bien discriminant doit avoir une AUC entre 0.87 et 0.9. Un modèle ayant une AUC supérieure à 0.9 est excellent.

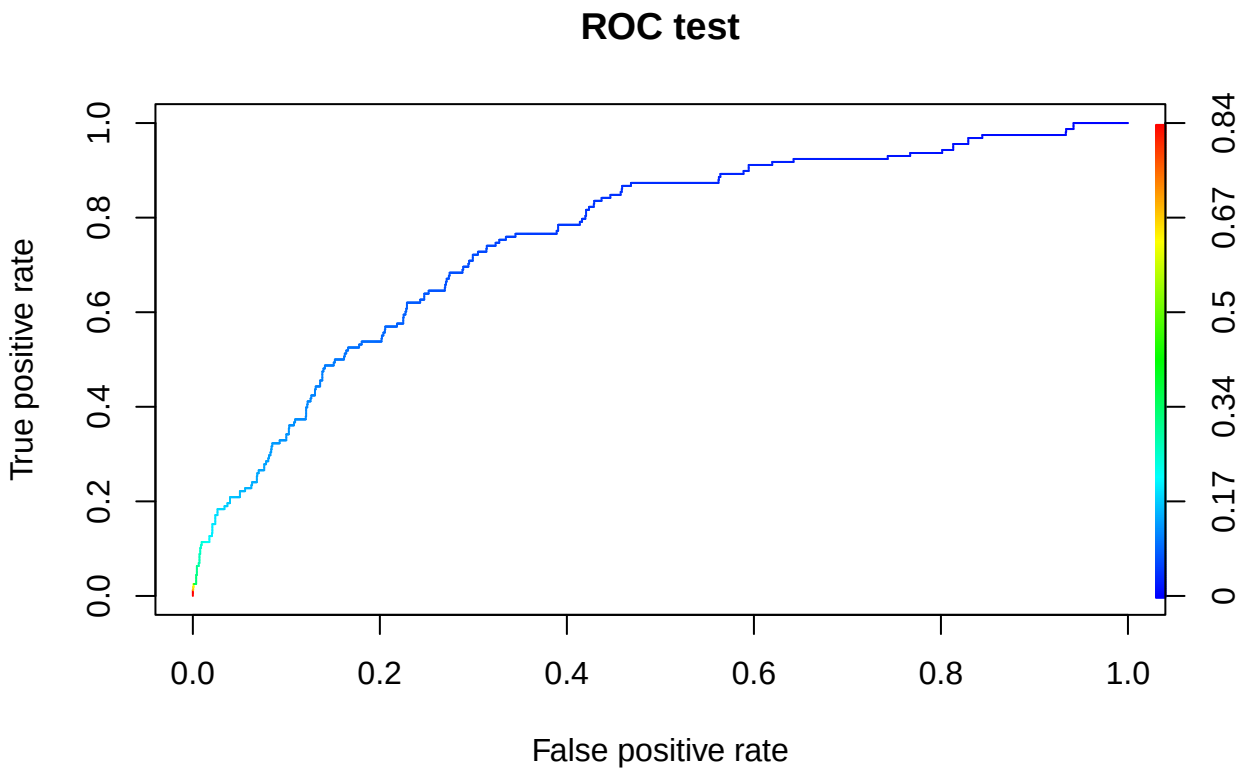
```
library(ROCR)
proba<-predict(Regfin,type='response',NVBaprt)
pred<-prediction(proba,NVBaprt$accident)
perf <- performance(pred,"tpr","fpr")
AUC <- performance(pred,measure="auc")
AUCa<-AUC@y.values[[1]]
AUCa
```

```
## [1] 0.7600564
```

Dans notre cas de figure,on enregistre une augmentation de l'AUC indiquant une amélioration des capacités discriminatoires, avec pour valeur précise 76,0%. Notre modèle se distingue d'un modèle aléatoire, on a AUC=0,76 ; il peut en effet être qualifié de bon.

0.3.14.2 Courbes ROC L'aire sous la courbe ROC (ou Area Under the Curve, AUC) peut être interprétée comme la probabilité que, parmi deux sujets choisis au hasard, un malade et un non-malade, la valeur du marqueur soit plus élevée pour le malade que pour le non-malade. Par conséquent, une AUC de 0,5 (50%) indique que le marqueur est non-informatif. Une augmentation de l'AUC indique une amélioration des capacités discriminatoires, avec un maximum de 1,0 (100%).

```
plot(perf, colorize = TRUE, main = "ROC test")
```



0.4 AVEC LE MEME JEU DE DONNEES INITIAL, CONSTRUISONS UN MODELE POISSONIEN PERMETTANT DE MODELISER LE NOMBRE D'ACCIDENTS

Le nombre d'accdent est caractérisé par la variable nbre.

La régression de Poisson est un outil puissant pour l'analyse des taux d'incidence dans les études de cohorte, et facilite les analyses de tendances temporelles qui peuvent être difficiles à évaluer avec d'autres méthodes.

0.4.1 Construisons le modèle

0.4.2 Importation du jeu de données

Affichage des six premières lignes

```
setwd("C:/Users/DELL/Documents/R/Base de données")
actuarNV1<-read.table("actuarNV1.csv",header=TRUE,sep=";",check.names=FALSE)
head(actuarNV1)
```

```
##   nocontrat exposition zone puissance agevehicule ageconducteur bonus marque
## 1       217       0.74    A         5           4            31    64      3
```

```
## 2       709       0.18    B         7           8           22   100      2
## 3       714       0.48    C         9           0           32    61     12
## 4       852       0.27    F         7           5           39   100     12
## 5      1083       0.51    E         4           0           49    50     12
## 6      1545       0.64    D        10           0           58    50     12
##   carburant densite region nbre    no garantie   cout
## 1         D      21      8    1 17001      1RC   0.00
## 2         E      26      0    1 17419      1RC   0.00
## 3         E      41     13    1 15851      4BG 687.82
## 4         E      11      0    1 21407      2DO  96.64
## 5         E      31     13    1 15589      2DO  70.88
## 6         D      72     13    2   772      1RC   0.00
```

```
str(actuarNV1)
```

```
## 'data.frame':    2765 obs. of  15 variables:
##  $ nocontrat    : int  217 709 714 852 1083 1545 1545 1870 1870 1963 ...
##  $ exposition   : num  0.74 0.18 0.48 0.27 0.51 0.64 0.64 0.11 0.11 0.1 ...
##  $ zone         : chr  "A" "B" "C" "F" ...
##  $ puissance    : int  5 7 9 7 4 10 10 5 5 9 ...
##  $ agevehicule  : int  4 8 0 5 0 0 0 0 0 0 ...
##  $ ageconducteur: int  31 22 32 39 49 58 58 52 52 78 ...
##  $ bonus        : int  64 100 61 100 50 50 50 50 50 50 ...
##  $ marque       : int  3 2 12 12 12 12 12 12 12 12 ...
##  $ carburant    : chr  "D" "E" "E" "E" ...
##  $ densite      : int  21 26 41 11 31 72 72 73 73 72 ...
##  $ region       : int  8 0 13 0 13 13 13 13 13 13 ...
##  $ nbre         : int  1 1 1 1 1 2 2 2 2 1 ...
##  $ no           : int  17001 17419 15851 21407 15589 772 762 17219 17111 16336 ...
##  $ garantie     : chr  "1RC" "1RC" "4BG" "2DO" ...
##  $ cout         : num  0 0 687.8 96.6 70.9 ...
```

Recodons les variables maruqe et region

```
actuarNV1 $marque <- factor(actuarNV1$marque)
actuarNV1$region<- factor(actuarNV1$region)
```

```
str(actuarNV1)
```

```
## 'data.frame':    2765 obs. of  15 variables:
##  $ nocontrat    : int  217 709 714 852 1083 1545 1545 1870 1870 1963 ...
##  $ exposition   : num  0.74 0.18 0.48 0.27 0.51 0.64 0.64 0.11 0.11 0.1 ...
##  $ zone         : chr  "A" "B" "C" "F" ...
##  $ puissance    : int  5 7 9 7 4 10 10 5 5 9 ...
##  $ agevehicule  : int  4 8 0 5 0 0 0 0 0 0 ...
##  $ ageconducteur: int  31 22 32 39 49 58 58 52 52 78 ...
##  $ bonus        : int  64 100 61 100 50 50 50 50 50 50 ...
##  $ marque       : Factor w/ 11 levels "1","2","3","4",..: 3 2 9 9 9 9 9 9 9 9 ...
##  $ carburant    : chr  "D" "E" "E" "E" ...
##  $ densite      : int  21 26 41 11 31 72 72 73 73 72 ...
##  $ region       : Factor w/ 15 levels "-1","0","1","2",..: 10 2 15 2 15 15 15 15 15 15 ...
##  $ nbre         : int  1 1 1 1 1 2 2 2 2 1 ...
```

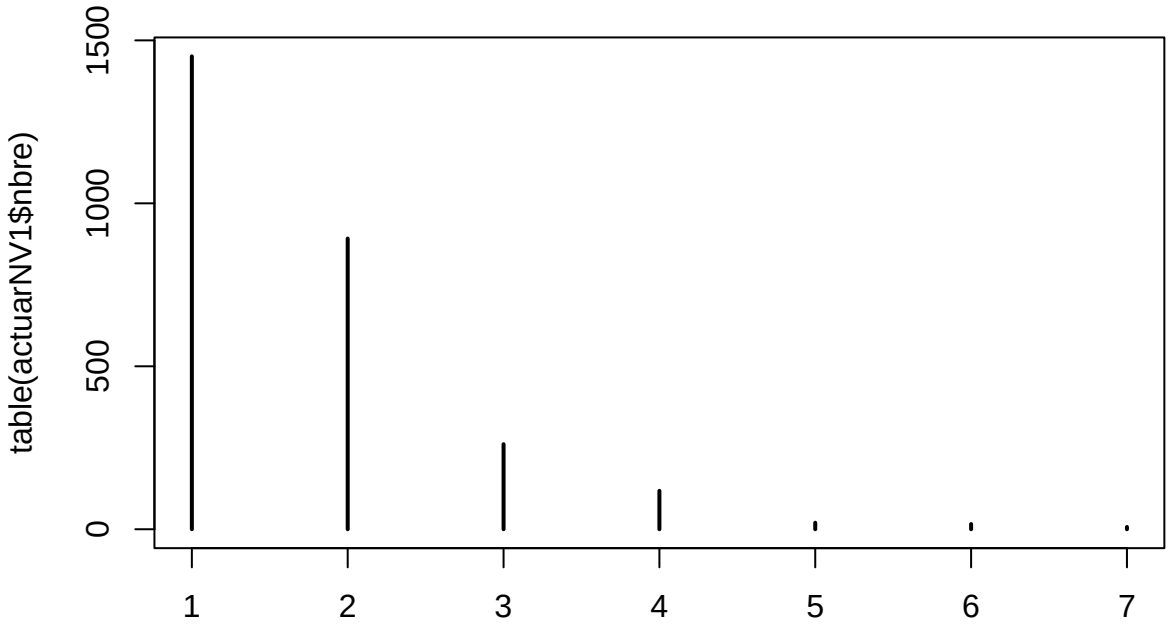
```
##  $ no           : int  17001 17419 15851 21407 15589 772 762 17219 17111 16336 ...
##  $ garantie     : chr  "1RC" "1RC" "4BG" "2DO" ...
##  $ cout         : num  0 0 687.8 96.6 70.9 ...
```

```
summary(actuarNV1)
```

```
##    nocontrat         exposition          zone             puissance
##  Min.   :    217   Min.   :0.008219   Length:2765        Min.   : 4.000
##  1st Qu.: 108786   1st Qu.:0.500000   Class :character   1st Qu.: 5.000
##  Median :1041406   Median :0.870000   Mode  :character   Median : 6.000
##  Mean   : 778477   Mean   :0.737445                      Mean   : 6.374
##  3rd Qu.:1128643   3rd Qu.:1.000000                      3rd Qu.: 7.000
##  Max.   :2054297   Max.   :1.300000                      Max.   :15.000
##
##   agevehicule     ageconducteur       bonus            marque
##  Min.   : 0.000   Min.   :18.00   Min.   : 50.00   2      :755
##  1st Qu.: 2.000   1st Qu.:33.00   1st Qu.: 50.00   1      :683
##  Median : 5.000   Median :43.00   Median : 50.00   12     :376
##  Mean   : 6.209   Mean   :44.09   Mean   : 61.21   3      :289
##  3rd Qu.: 9.000   3rd Qu.:53.00   3rd Qu.: 68.00   5      :188
##  Max.   :35.000   Max.   :99.00   Max.   :165.00   6      :128
##                                                    (Other):346
##   carburant            densite          region          nbre
##  Length:2765        Min.   :11.00   13     :1682   Min.   :1.000
##  Class :character   1st Qu.:24.00   5      : 114   1st Qu.:1.000
##  Mode  :character   Median :52.00   4      : 102   Median :1.000
##                     Mean   :49.11   7      : 100   Mean   :1.712
##                     3rd Qu.:82.00   6      :  88   3rd Qu.:2.000
##                     Max.   :94.00   3      :  82   Max.   :7.000
##                                     (Other): 597
##        no           garantie             cout
##  Min.   :   148   Length:2765        Min.   : -3811.2
##  1st Qu.: 28810   Class :character   1st Qu.:   132.7
##  Median : 38506   Mode  :character   Median :   405.6
##  Mean   : 44531                      Mean   :  1069.3
##  3rd Qu.: 49055                      3rd Qu.:  1128.1
##  Max.   :104152                      Max.   :152449.0
##
```

```
plot(table(actuarNV1$nbre))
```

0.4.2.1 Faisons la représentation graphique



La distribution de Poisson possède deux éléments remarquables : L'espérance (ou moyenne) d'une variable aléatoire distribuée selon une loi de poisson est égale à Lambda :E(y)=GAMA

La variance d'une variable aléatoire distribuée selon une loi de poisson est aussi égale à Lambda : Var(y)=GAMA La distribution a une allure en « L », calculons maintenant la moyenne et la variance. En effet dans une distribution de poisson la logique voudrait qu'il y ait une égalité entre l'espérance (ou moyenne) et la variance .

```
mean(actuarNV1$nbre)
```

```
## [1] 1.712477
```

```
var(actuarNV1$nbre)
```

```
## [1] 0.9285165
```

On remarque qu'il existe une légère différence entre ces deux valeurs, on peut donc partir sur le principe d'égalité de la moyenne avec la variance.

On dit qu'il y a surdispersion lorsque la variance réelle est supérieure à cette variance théorique. Cela est problématique car dans cette situation, l'erreur standard des paramètres des modèles de régression de Poisson sera sous estimée. Ceci peut conduire à une p-value excessivement faible, et donc aboutir à une conclusion erronée sur la significativité de la liaison entre les comptages observés et la ou les variables explicatives.

```
library(fitdistrplus)
```

0.4.2.2 Test d'adéquation à la loi de poisson 1. Estimation des paramètres par la méthode du maximum de vraisemblance

```
fpois <- fitdist(actuarNV1$nbre, "pois")
summary(fpois)
```

```
## Fitting of the distribution ' pois ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## lambda 1.712477 0.02488655
## Loglikelihood:  -3909.49   AIC:  7820.98   BIC:  7826.904
```

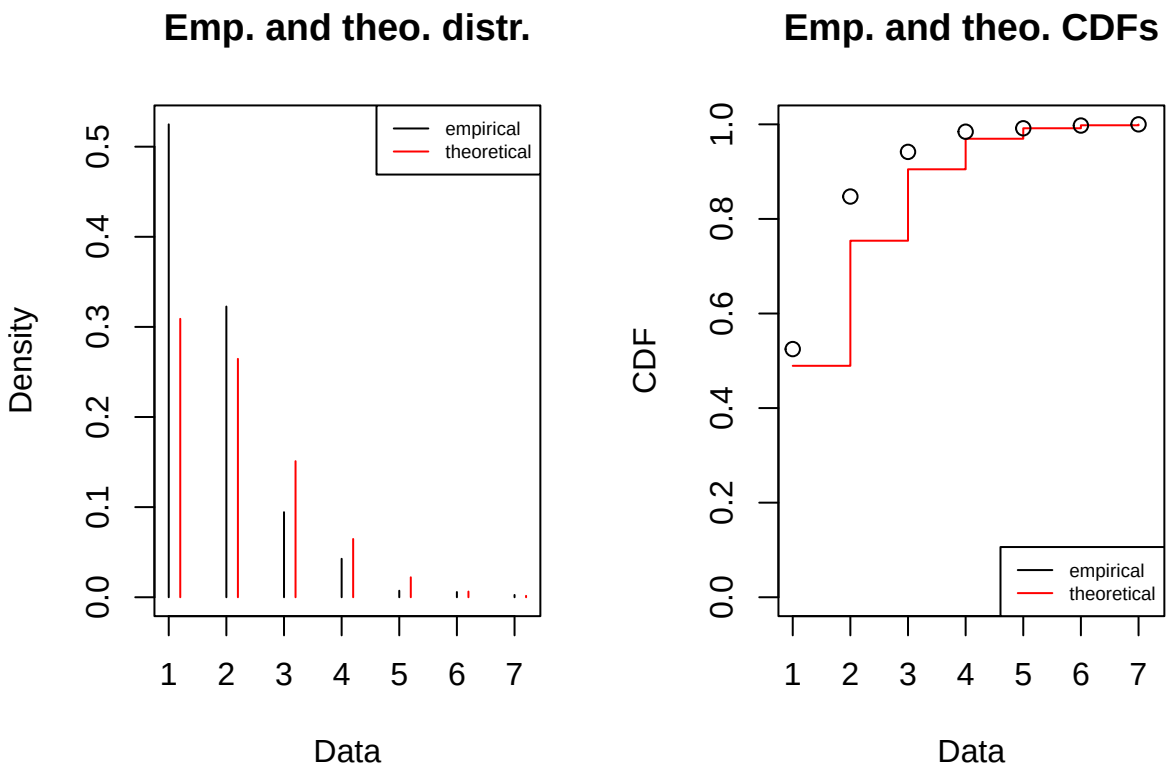
2. Tests d'adéquation

```
fpois <- fitdist(actuarNV1$nbre, "pois")
gofstat(fpois)
```

```
## Chi-squared statistic:  141.7319
## Degree of freedom of the Chi-squared distribution:  3
## Chi-squared p-value:  1.599609e-30
## Chi-squared table:
##      obscounts theocounts
## <= 1 1451.0000  1353.1370
## <= 2  892.0000   731.4680
## <= 3  261.0000   417.5408
## <= 4  118.0000   178.7573
## > 4    43.0000    84.0968
##
## Goodness-of-fit criteria
##                                1-mle-pois
## Akaike's Information Criterion   7820.980
## Bayesian Information Criterion   7826.904
```

Ici, la p.value est inférieur à 5%, les données ne suivent pas une loi de poisson. on peut le voir à travers les graphisques ci dessous.

```
plot(fpois)
```



0.4.2.3 Test d'adéquation à la loi Binomiale négative la loi binomiale négative est le nombre d'échecs avant l'obtention de n succès dans une expérience où la probabilité de succès est p. Elle peut aussi s'interpréter comme un mélange de lois de Poisson lorsque le paramètre lambda suit une loi gamma, ce qui s'interprète comme la prise en compte d'une hétérogénéité non observable.

1. Estimation des paramètres par la méthode du maximum de vraisemblance

```
fnbinom <- fitdist(actuarNV1$nbre, "nbinom")
summary(fnbinom)
```

```
## Fitting of the distribution ' nbinom ' by maximum likelihood
## Parameters :
##          estimate Std. Error
## size 1.286622e+07 3.66549831
## mu   1.712424e+00 0.02488578
## Loglikelihood:  -3909.49   AIC:  7822.98   BIC:  7834.829
## Correlation matrix:
##               size            mu
## size  1.000000e+00 -1.037037e-08
## mu   -1.037037e-08  1.000000e+00
```

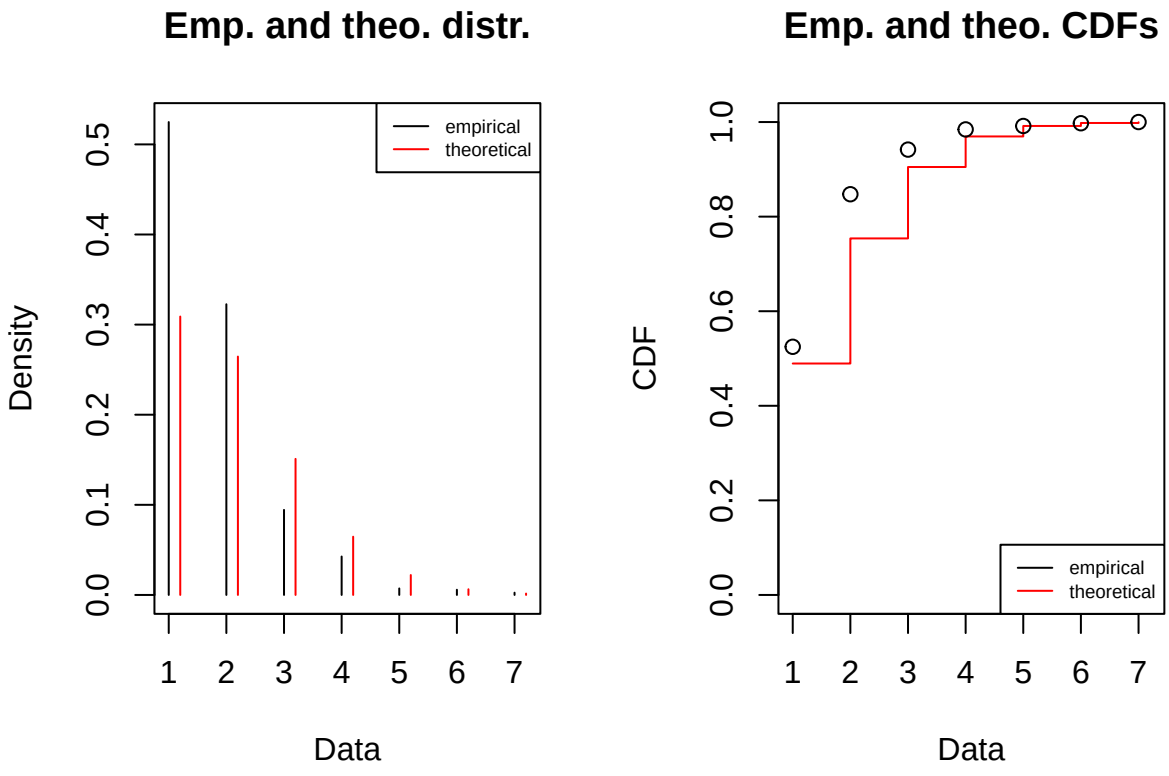
2. Tests d'adéquation

```r
fnbinom <- fitdist(actuarNV1$nbre, "nbinom")
gofstat(fnbinom)
```

```
## Chi-squared statistic:  141.7038
## Degree of freedom of the Chi-squared distribution:  2
## Chi-squared p-value:  1.695899e-31
## Chi-squared table:
##       obscounts theocounts
## <= 1 1451.00000 1353.18268
## <= 2  892.00000  731.46142
## <= 3  261.00000  417.52404
## <= 4  118.00000  178.74457
## > 4    43.00000   84.08729
##
## Goodness-of-fit criteria
##                                1-mle-nbinom
## Akaike's Information Criterion     7822.980
## Bayesian Information Criterion     7834.829
```

3. Visualisation graphique de l'adéquation

```r
plot(fnbinom)
```



0.4.2.4 Construisons le modèle

Extraction des numéros des contrats car ces variables ne sont d'aucune utilité pourles analyses

```
actuarNV1$nocontrat <- NULL
actuarNV1$no <- NULL
```

```
head(actuarNV1)
```

```
##   exposition zone puissance agevehicule ageconducteur bonus marque carburant
## 1       0.74    A         5           4            31    64      3         D
## 2       0.18    B         7           8            22   100      2         E
## 3       0.48    C         9           0            32    61     12         E
## 4       0.27    F         7           5            39   100     12         E
## 5       0.51    E         4           0            49    50     12         E
## 6       0.64    D        10           0            58    50     12         D
##   densite region nbre garantie   cout
## 1      21      8    1      1RC   0.00
## 2      26      0    1      1RC   0.00
## 3      41     13    1      4BG 687.82
## 4      11      0    1      2DO  96.64
## 5      31     13    1      2DO  70.88
## 6      72     13    2      1RC   0.00
```

```
regres= glm(nbre ~ .,data=actuarNV1,family = poisson)
summary(regres)
```

0.4.2.5 Estimation du modèle

```
##
## Call:
## glm(formula = nbre ~ ., family = poisson, data = actuarNV1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2765  -0.5232  -0.1965   0.2698   2.6439
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.343e-01  5.122e-01   0.457 0.647366
## exposition     1.833e-01  5.234e-02   3.502 0.000462 ***
## zoneB          2.746e-02  5.831e-02   0.471 0.637696
## zoneC          2.265e-02  4.744e-02   0.477 0.633089
## zoneD          7.190e-02  4.978e-02   1.444 0.148686
## zoneE          1.735e-02  5.183e-02   0.335 0.737823
## zoneF         -2.793e-02  9.967e-02  -0.280 0.779341
## puissance     -1.798e-03  8.416e-03  -0.214 0.830796
## agevehicule   -8.878e-03  3.588e-03  -2.475 0.013341 *
## ageconducteur -5.785e-04  1.253e-03  -0.462 0.644248
## bonus          3.233e-03  4.231e-03   0.764 0.444711
## marque2        1.468e-03  4.224e-02   0.035 0.972286
```

```
## marque3        9.143e-02  5.399e-02   1.694 0.090331 .
## marque4        2.398e-01  6.953e-02   3.448 0.000565 ***
## marque5        2.829e-02  6.417e-02   0.441 0.659346
## marque6        4.509e-02  7.527e-02   0.599 0.549143
## marque10       1.334e-01  8.947e-02   1.491 0.135893
## marque11       3.355e-01  9.443e-02   3.553 0.000382 ***
## marque12       7.512e-02  5.429e-02   1.384 0.166443
## marque13      -1.185e-01  1.209e-01  -0.980 0.327003
## marque14       2.190e-01  1.797e-01   1.219 0.222964
## carburantE    -1.153e-01  3.105e-02  -3.714 0.000204 ***
## densite       -6.477e-05  5.713e-04  -0.113 0.909730
## region0       -1.117e-01  1.510e-01  -0.739 0.459639
## region1        5.256e-03  1.556e-01   0.034 0.973046
## region2        1.057e-01  1.661e-01   0.637 0.524296
## region3        9.337e-02  1.802e-01   0.518 0.604320
## region4        7.830e-02  1.936e-01   0.404 0.685940
## region5        5.815e-02  2.075e-01   0.280 0.779243
## region6        2.587e-02  2.252e-01   0.115 0.908535
## region7        4.992e-03  2.385e-01   0.021 0.983301
## region8       -1.888e-01  2.670e-01  -0.707 0.479531
## region9        1.611e-01  2.684e-01   0.600 0.548457
## region10       2.323e-01  2.795e-01   0.831 0.405818
## region11       2.380e-01  2.940e-01   0.810 0.418168
## region12       3.978e-02  3.097e-01   0.128 0.897789
## region13       8.082e-02  2.891e-01   0.280 0.779860
## garantie2DO    6.931e-02  3.651e-02   1.899 0.057624 .
## garantie3VI    4.729e-02  6.009e-02   0.787 0.431338
## garantie4BG   -2.578e-01  4.029e-02  -6.399 1.57e-10 ***
## garantie5CO    7.095e-01  1.818e-01   3.904 9.48e-05 ***
## garantie6CL   -2.159e-01  5.020e-01  -0.430 0.667190
## cout           1.103e-06  3.567e-06   0.309 0.757092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1264.5  on 2764  degrees of freedom
## Residual deviance: 1081.7  on 2722  degrees of freedom
## AIC: 7722.1
##
## Number of Fisher Scoring iterations: 4
```

Conclusion :les varibles exerçant une influence significative sur le nombre d'accident sont: l'exposition, l'âge du véhicule,le carburant,la marque et la garantie.

Procédons à un test d'anova pour déterminer les variables les plus liées à la cible (ici le nombre d'accidents)

```
anova(regres,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: nbre
```

```
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           2764     1264.5
## exposition      1    3.046      2763     1261.5 0.0809405 .
## zone            5    4.337      2758     1257.2 0.5019386
## puissance       1    0.135      2757     1257.0 0.7135413
## agevehicule     1   25.016      2756     1232.0 5.685e-07 ***
## ageconducteur   1    2.064      2755     1229.9 0.1507797
## bonus           1    6.766      2754     1223.2 0.0092900 **
## marque         10   29.814      2744     1193.4 0.0009187 ***
## carburant       1   11.432      2743     1181.9 0.0007218 ***
## densite         1    0.134      2742     1181.8 0.7145070
## region         14   14.854      2728     1166.9 0.3882058
## garantie        5   85.194      2723     1081.8 < 2.2e-16 ***
## cout            1    0.092      2722     1081.7 0.7614334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reprenons le modèle avec ces variables

```
regresfin = glm(nbre ~ agevehicule+bonus+marque+carburant+garantie,data=actuarNV1,
                family = poisson)
summary(regresfin)
```

0.4.2.6 Construction du nouveau modèle

```
##
## Call:
## glm(formula = nbre ~ agevehicule + bonus + marque + carburant +
##     garantie, family = poisson, data = actuarNV1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2252  -0.5378  -0.2020   0.2563   2.7316
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.5257147  0.0710943   7.395 1.42e-13 ***
## agevehicule -0.0091624  0.0035383  -2.589 0.009612 **
## bonus        0.0017177  0.0008548   2.009 0.044497 *
## marque2      0.0044795  0.0417930   0.107 0.914644
## marque3      0.0996267  0.0534559   1.864 0.062362 .
## marque4      0.2303945  0.0691702   3.331 0.000866 ***
## marque5      0.0452754  0.0637149   0.711 0.477336
## marque6      0.0551129  0.0748615   0.736 0.461611
## marque10     0.1146047  0.0863349   1.327 0.184362
## marque11     0.3341156  0.0895966   3.729 0.000192 ***
## marque12     0.0437506  0.0524022   0.835 0.403774
```

```
## marque13    -0.1345766  0.1194809  -1.126 0.260020
## marque14     0.2198328  0.1767924   1.243 0.213701
## carburantE  -0.1180862  0.0302376  -3.905 9.41e-05 ***
## garantie2DO  0.0705334  0.0363018   1.943 0.052020 .
## garantie3VI  0.0528714  0.0596029   0.887 0.375046
## garantie4BG -0.2373692  0.0394879  -6.011 1.84e-09 ***
## garantie5CO  0.7099616  0.1793378   3.959 7.53e-05 ***
## garantie6CL -0.2430180  0.5015502  -0.485 0.628007
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1264.5  on 2764  degrees of freedom
## Residual deviance: 1113.9  on 2746  degrees of freedom
## AIC: 7706.4
##
## Number of Fisher Scoring iterations: 4
```

```
anova(regresfin,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: nbre
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         2764     1264.5
## agevehicule  1   23.785      2763     1240.8 1.077e-06 ***
## bonus        1    7.247      2762     1233.5 0.0071036 **
## marque      10   29.657      2752     1203.8 0.0009746 ***
## carburant    1   10.848      2751     1193.0 0.0009893 ***
## garantie     5   79.092      2746     1113.9 1.299e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On remarque que les variables sont toutes très liées à la cible

Calculons les intervalles de confiance

```
exp(confint(regresfin))
```

```
##                 2.5 %    97.5 %
## (Intercept) 1.4715525 1.9445270
## agevehicule 0.9840047 0.9977481
## bonus       1.0000296 1.0033865
## marque2     0.9255393 1.0903191
## marque3     0.9942590 1.2260947
## marque4     1.0973940 1.4393680
```

```
## marque5     0.9221405 1.1838721
## marque6     0.9102214 1.2208357
## marque10    0.9434228 1.3236944
## marque11    1.1670463 1.6585797
## marque12    0.9424583 1.1574081
## marque13    0.6858875 1.0963485
## marque14    0.8635936 1.7308512
## carburantE  0.8374413 0.9428324
## garantie2DO 0.9993536 1.1521933
## garantie3VI 0.9366891 1.1833037
## garantie4BG 0.7297876 0.8519806
## garantie5CO 1.4018507 2.8381003
## garantie6CL 0.2429055 1.8291309
```

Pour faciliter les interprétations comme dans le cas de la régression logistique, Prenons l'exponentiel des coefficients.

```
exp(regresfin$coefficients)
```

```
## (Intercept) agevehicule       bonus     marque2     marque3     marque4
##   1.6916675   0.9908794   1.0017192   1.0044896   1.1047584   1.2590966
##     marque5     marque6    marque10    marque11    marque12    marque13
##   1.0463160   1.0566599   1.1214300   1.3967046   1.0447217   0.8740859
##    marque14  carburantE garantie2DO garantie3VI garantie4BG garantie5CO
##   1.2458684   0.8886195   1.0730804   1.0542941   0.7887001   2.0339132
## garantie6CL
##   0.7842574
```

```
Exemple d'age du véhicule: Si B0=1,6916 et B1= ,9908
```

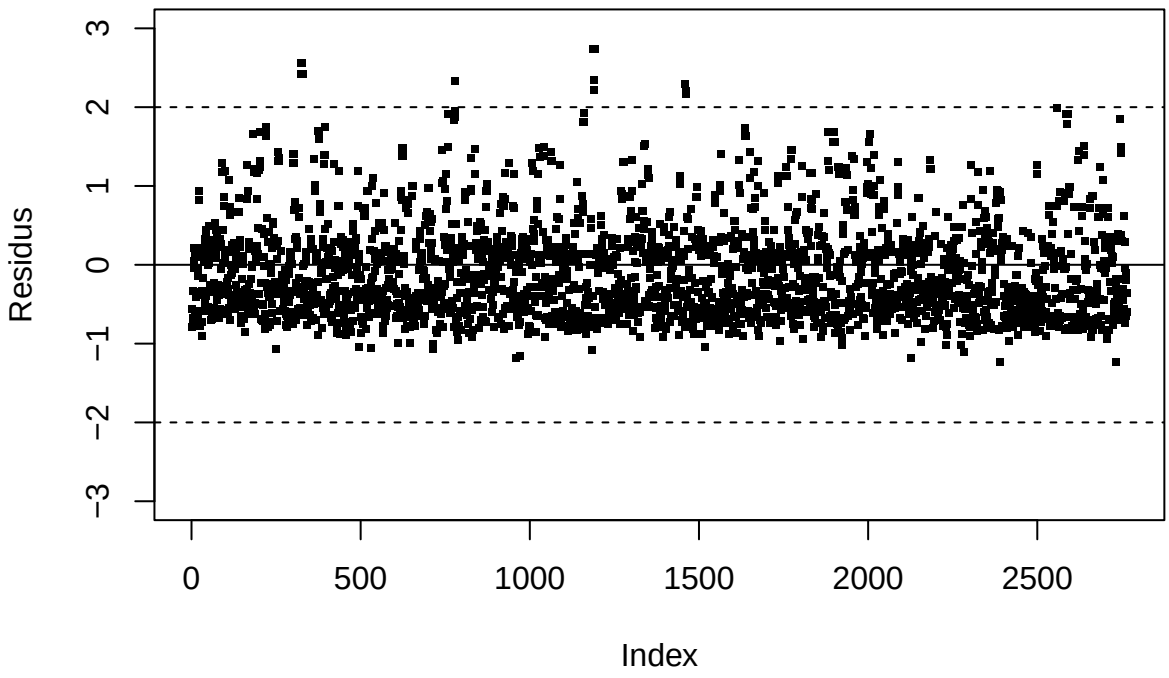
Alors en moyenne le nombre d'accident pour les véhicules ayant un an de plus est: 1,6916* ,9908= 1,676037

```
1.6916*0.9908
```

```
## [1] 1.676037
```

```
res.m1<-rstudent(regresfin)
plot(res.m1,pch=15,cex=.5,ylab="Residus",ylim=c(-3,3))
abline(h=c(-2,0,2),lty=c(2,1,2))
```

0.4.2.7 Analyse des résidus



Le résidu standardisé est asymptotiquement gaussien. Les points ne forment pas approximativement une droite, nous le vérifirons par le test de shapiro.

```
shapiro.test(res.m1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.m1
## W = 0.91742, p-value < 2.2e-16
```

On constate que p-value $<$ 2.2e-16, le test de shapiro wilk confirme que les résidus ne suivent pas une loi normale.

0.4.2.8 Faisons la prévision

On décide prédire le nombre d'accidents que fera l'individu n°2764.

```
predict.glm(regresfin, actuarNV1[2764,], type = "response")
```

```
##     2764
## 1.716843
```

Intervalle de confiance

```
lamb<-predict.glm(regresfin, actuarNV1[2764,], type = "response")
loglamb <- predict.glm(regresfin, actuarNV1[2764,], se.fit = TRUE)
icloglamb <- c(loglamb$fit - 1.96 * loglamb$se.fit, loglamb$fit + 1.96 * loglamb$se.fit)
ic <- exp(icloglamb)
ic
```

```
##     2764     2764
## 1.499226 1.966048
```

L'intervalle de confiance nous montre que le nombre d'accidents succeptible d'être réalisé avec cet individu est compris entre 1,499 et 1,966 soit 2 accident.

0.4.2.9 Méthode d'approche pas à pas avec AIC :

SELECTION DE VARIABLES

```
regresfin1 = glm(nbre ~ .,data=actuarNV1,family = poisson)
step(regresfin1, direction = "both", k = 2)
```

```
## Start:  AIC=7722.09
## nbre ~ exposition + zone + puissance + agevehicule + ageconducteur +
##     bonus + marque + carburant + densite + region + garantie +
##     cout
##
##                 Df Deviance    AIC
## - region        14   1097.7 7710.1
## - zone           5   1084.7 7715.1
## - densite        1   1081.7 7720.1
## - puissance      1   1081.7 7720.1
## - cout           1   1081.8 7720.2
## - ageconducteur  1   1081.9 7720.3
## - bonus          1   1082.2 7720.7
## <none>               1081.7 7722.1
## - agevehicule    1   1087.8 7726.3
## - marque        10   1110.1 7730.5
## - exposition     1   1094.1 7732.5
## - carburant      1   1095.5 7733.9
## - garantie       5   1166.4 7796.8
##
## Step:  AIC=7710.09
## nbre ~ exposition + zone + puissance + agevehicule + ageconducteur +
##     bonus + marque + carburant + densite + garantie + cout
##
##                 Df Deviance    AIC
## - zone           5   1100.9 7703.3
## - puissance      1   1097.7 7708.1
## - densite        1   1097.7 7708.1
## - cout           1   1097.7 7708.2
## - ageconducteur  1   1098.2 7708.7
## <none>               1097.7 7710.1
## - bonus          1   1100.7 7711.1
## - agevehicule    1   1104.4 7714.8
```

```
## - marque        10   1126.5 7719.0
## - exposition     1   1110.8 7721.2
## + region        14   1081.7 7722.1
## - carburant      1   1112.2 7722.7
## - garantie       5   1181.2 7783.6
##
## Step:  AIC=7703.34
## nbre ~ exposition + puissance + agevehicule + ageconducteur +
##     bonus + marque + carburant + densite + garantie + cout
##
##                 Df Deviance    AIC
## - puissance      1   1100.9 7701.3
## - densite        1   1101.0 7701.4
## - cout           1   1101.0 7701.4
## - ageconducteur  1   1101.4 7701.9
## <none>               1100.9 7703.3
## - bonus          1   1104.3 7704.8
## - agevehicule    1   1107.5 7707.9
## + zone           5   1097.7 7710.1
## - marque        10   1131.0 7713.4
## - exposition     1   1113.8 7714.2
## + region        14   1084.7 7715.1
## - carburant      1   1115.9 7716.3
## - garantie       5   1184.0 7776.4
##
## Step:  AIC=7701.34
## nbre ~ exposition + agevehicule + ageconducteur + bonus + marque +
##     carburant + densite + garantie + cout
##
##                 Df Deviance    AIC
## - cout           1   1101.0 7699.4
## - densite        1   1101.0 7699.4
## - ageconducteur  1   1101.4 7699.9
## <none>               1100.9 7701.3
## - bonus          1   1104.4 7702.8
## + puissance      1   1100.9 7703.3
## - agevehicule    1   1107.6 7706.0
## + zone           5   1097.7 7708.1
## - exposition     1   1113.8 7712.2
## - marque        10   1131.8 7712.3
## + region        14   1084.7 7713.2
## - carburant      1   1115.9 7714.3
## - garantie       5   1184.2 7774.7
##
## Step:  AIC=7699.42
## nbre ~ exposition + agevehicule + ageconducteur + bonus + marque +
##     carburant + densite + garantie
##
##                 Df Deviance    AIC
## - densite        1   1101.0 7697.5
## - ageconducteur  1   1101.5 7698.0
## <none>               1101.0 7699.4
## - bonus          1   1104.4 7700.9
## + cout           1   1100.9 7701.3
```

```
## + puissance      1   1101.0 7701.4
## - agevehicule    1   1107.7 7704.1
## + zone           5   1097.7 7706.2
## - exposition     1   1113.8 7710.2
## - marque        10   1132.0 7710.4
## + region        14   1084.8 7711.2
## - carburant      1   1116.0 7712.5
## - garantie       5   1184.8 7773.3
##
## Step:  AIC=7697.49
## nbre ~ exposition + agevehicule + ageconducteur + bonus + marque +
##     carburant + garantie
##
##                 Df Deviance    AIC
## - ageconducteur  1   1101.6 7696.0
## <none>               1101.0 7697.5
## - bonus          1   1104.5 7698.9
## + densite        1   1101.0 7699.4
## + cout           1   1101.0 7699.4
## + puissance      1   1101.0 7699.5
## - agevehicule    1   1107.7 7702.2
## + zone           5   1097.7 7704.2
## - exposition     1   1113.8 7708.2
## - marque        10   1132.1 7708.6
## + region        14   1084.9 7709.3
## - carburant      1   1116.1 7710.5
## - garantie       5   1185.2 7771.6
##
## Step:  AIC=7696.01
## nbre ~ exposition + agevehicule + bonus + marque + carburant +
##     garantie
##
##                 Df Deviance    AIC
## <none>               1101.6 7696.0
## + ageconducteur  1   1101.0 7697.5
## + cout           1   1101.5 7697.9
## + densite        1   1101.5 7698.0
## + puissance      1   1101.6 7698.0
## - bonus          1   1107.3 7699.8
## - agevehicule    1   1108.4 7700.9
## + zone           5   1098.3 7702.8
## - exposition     1   1113.9 7706.4
## - marque        10   1132.6 7707.0
## + region        14   1085.0 7707.5
## - carburant      1   1117.3 7709.7
## - garantie       5   1185.2 7769.7


##
## Call:  glm(formula = nbre ~ exposition + agevehicule + bonus + marque +
##     carburant + garantie, family = poisson, data = actuarNV1)
##
## Coefficients:
## (Intercept)   exposition  agevehicule        bonus      marque2      marque3
##    0.369947     0.178372    -0.009245     0.002083     0.003038     0.103030
```

```
##     marque4      marque5      marque6     marque10     marque11     marque12
##    0.241001     0.041810     0.059123     0.122969     0.340844     0.070339
##    marque13     marque14   carburantE  garantie2DO  garantie3VI  garantie4BG
##   -0.124029     0.235812    -0.119771     0.069684     0.050504    -0.246791
## garantie5CO  garantie6CL
##    0.741346    -0.207027
##
## Degrees of Freedom: 2764 Total (i.e. Null);  2745 Residual
## Null Deviance:       1265
## Residual Deviance: 1102  AIC: 7696
```

On note dans cette procédure un AIC plus faible (avec le dernier modèle) estmé à 7696 contre un AIC de 7706.4 pour le modele de régression choisi plus haut regresfin.

Considérons le nouveau modele (soit Modfin)

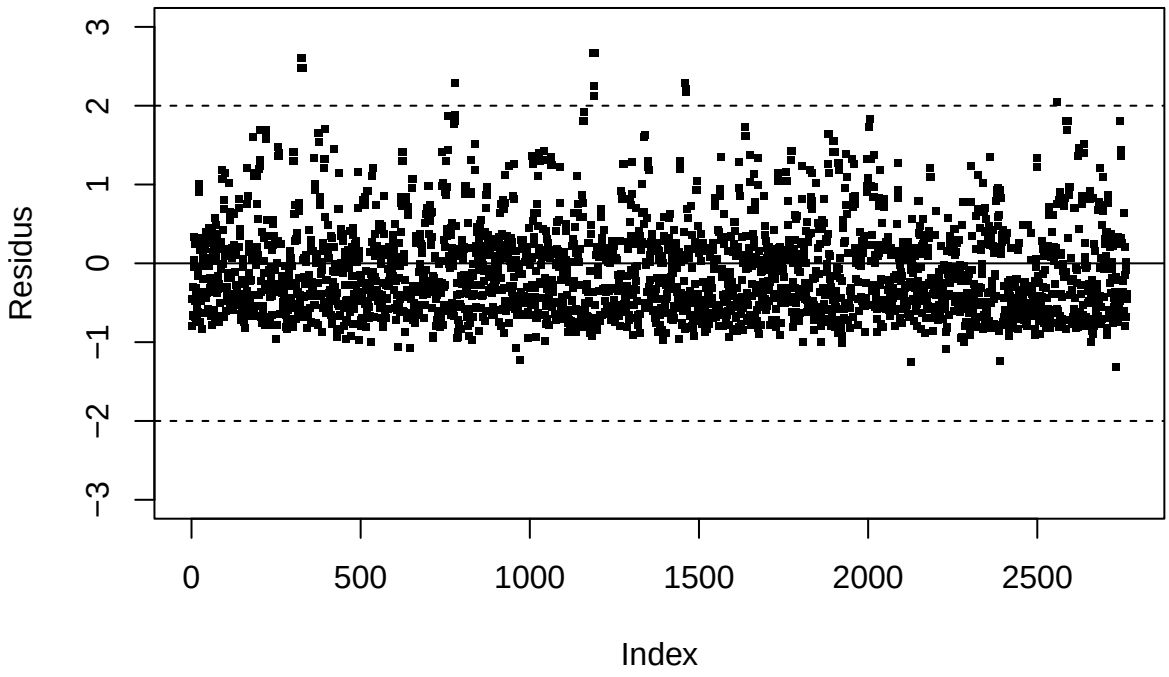
```
Modfin=glm(nbre ~ exposition + agevehicule + bonus + marque +
    carburant + garantie, family = poisson, data = actuarNV1)
summary(Modfin)
```

```
##
## Call:
## glm(formula = nbre ~ exposition + agevehicule + bonus + marque +
##     carburant + garantie, family = poisson, data = actuarNV1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2996  -0.5316  -0.2047   0.2621   2.6571
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.369947   0.084038   4.402 1.07e-05 ***
## exposition   0.178372   0.051134   3.488 0.000486 ***
## agevehicule -0.009245   0.003548  -2.605 0.009177 **
## bonus        0.002083   0.000859   2.425 0.015293 *
## marque2      0.003038   0.041796   0.073 0.942048
## marque3      0.103030   0.053479   1.927 0.054037 .
## marque4      0.241001   0.069218   3.482 0.000498 ***
## marque5      0.041810   0.063725   0.656 0.511758
## marque6      0.059123   0.074874   0.790 0.429744
## marque10     0.122969   0.086363   1.424 0.154484
## marque11     0.340844   0.089657   3.802 0.000144 ***
## marque12     0.070339   0.052919   1.329 0.183789
## marque13    -0.124029   0.119519  -1.038 0.299393
## marque14     0.235812   0.176854   1.333 0.182408
## carburantE  -0.119771   0.030234  -3.962 7.45e-05 ***
## garantie2DO  0.069684   0.036297   1.920 0.054878 .
## garantie3VI  0.050504   0.059622   0.847 0.396951
## garantie4BG -0.246791   0.039586  -6.234 4.54e-10 ***
## garantie5CO  0.741345   0.179571   4.128 3.65e-05 ***
## garantie6CL -0.207027   0.501646  -0.413 0.679829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1264.5  on 2764  degrees of freedom
## Residual deviance: 1101.6  on 2745  degrees of freedom
## AIC: 7696
##
## Number of Fisher Scoring iterations: 4
```

```
res.m2<-rstudent(Modfin)
plot(res.m2,pch=15,cex=.5,ylab="Residus",ylim=c(-3,3))
abline(h=c(-2,0,2),lty=c(2,1,2))
```

0.4.2.10 Analyse des résidus



Le résidu standardisé n'a pas la forme d'une ligne droite. Nous le vérifirons par le test de shapiro.

```
shapiro.test(res.m2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.m2
## W = 0.92033, p-value < 2.2e-16
```

On constate que p-value < 2.2e-16, le test de shapiro wilk atteste que les résidus ne suivent pas une loi normale.

0.4.2.11 Faisons la prévision On décide prédire le nombre d'accidents que fera le même individu (l'individu n°2764)

```
predict.glm(Modfin , actuarNV1[2764,], type = "response")
```

```
##     2764
## 1.516038
```

Intervalle de confiance

```
lamb1<-predict.glm(Modfin, actuarNV1[2764,], type = "response")
loglamb1 <- predict.glm(Modfin, actuarNV1[2764,], se.fit = TRUE)
icloglamb1 <- c(loglamb1$fit - 1.96 * loglamb1$se.fit, loglamb1$fit + 1.96 * loglamb1$se.fit)
IC <- exp(icloglamb1)
IC
```

```
##     2764     2764
## 1.301196 1.766354
```

On enregistre une légère baisse du risque d'accident par l'individu à travers ce modèle.