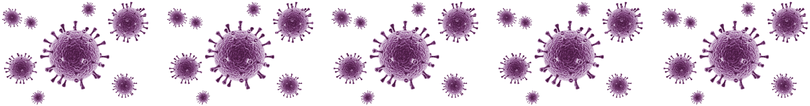
Modèles Linéaires Appliqués / Régression
Modèle de Poisson : Exemple

Arthur Charpentier

UQAM

Hiver 2020 - COVID-19 # 11



Fréquence d'Accident

```
> loc = "http://freakonometrics.free.fr/SAuto.RData"
> download.file(loc_fichier, "SinistresAuto.RData")
> load("SinistresAuto.RData")
> str(base)
'data.frame': 50000 obs. of  11 variables:
 $ nocontrat     : int  27 115 121 142 155 186 217 ...
 $ exposition    : num  0.87 0.72 0.05 0.9 0.12     ...
 $ zone          : Factor w/ 6 levels "A","B","C",  ...
 $ puissance     : int  7 5 6 10 7 5 5 5 4 4 ...
 $ agevehicule   : int  0 0 0 10 0 0 4 0 0 0 ...
 $ ageconducteur: int  56 45 37 42 59 75 31 41 42 ...
 $ bonus         : int  50 50 55 50 50 50 64 90 50 ...
 $ marque        : int  12 12 12 12 12 12 3 12 12  ...
 $ carburant     : Factor w/ 2 levels "D","E":      ...
 $ region        : int  93 54 11 93 73 42 21 11     ...
 $ nb            : num  0 0 0 0 0 0 0 0 0 0 ...
```

On veut modélier la fréquence annuelle de sinistre
à partir de `nb` survenus pendant la durée `exposition`

Fréquence d'Accident : 1

Modèle : $Y_i \sim \mathcal{P}(\lambda e_i)$ où $\lambda_i = \exp(\beta_0)$

```
> reg = glm(nb~1+offset(log(exposition)),
          data=base,family=poisson)
> summary(reg)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6201     0.0228  -114.9   <2e-16 ***

(Dispersion parameter taken to be 1)

    Null deviance: 12680  on 49999  degrees of freedom
Residual deviance: 12680  on 49999  degrees of freedom
AIC: 16353

Number of Fisher Scoring iterations: 6
```

Modèle : $Y_i \sim \mathcal{P}(\lambda e_i)$ où $\lambda_i = \exp(\beta_0)$

```
> with(base, sum(nb)/sum(exposition))
[1] 0.07279295
> coefficients(reg)
(Intercept)
  -2.620136
> exp(coefficients(reg))
(Intercept)
 0.07279295
> predict(reg,newdata=data.frame(exposition=1))
        1
-2.620136
> predict(reg,type="response",newdata=data.frame(
    exposition=1))
         1
0.07279295
```

Aussi, $\widehat{\eta}_i = \widehat{\beta}_0 = -2.6201$ et $\widehat{\mu}_i = 0.07276$.

Fréquence d'Accident : x_1

Modèle : $Y_i \sim \mathcal{P}(\lambda_i e_i)$ où $\lambda_i = \exp(\beta_0 + \beta_1 \mathbf{1}(x_i) = \text{E})$,
x_i est le carburant (diesel, essence)

```
> with(base,table(carburant,nb))
          nb
carburant     0     1     2     3     4    16
        D 23475   886    45     6     1     0
        E 24722   821    41     1     1     1
> reg1 = glm(nb~carburant+offset(log(exposition)),
+              data=base,family=poisson)
> summary(reg1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.52929    0.03165 -79.904  < 2e-16 ***
carburantE  -0.18031    0.04563  -3.952 7.76e-05 ***
```

Aussi $\widehat{\beta}_0 = -2.5293$ et $\widehat{\beta}_1 = -0.1803$.

Fréquence d'Accident : x_1

$\widehat{\beta}_0 = -2.5293$ et $\widehat{\beta}_1 = -0.1803$.

$$\widehat{\lambda}_i = \begin{cases} e^{-2.5293} = 0.079715 \text{ si } x_i = \text{D} \\ e^{-2.5293-0.1803} = 0.066563 \text{ si } x_i = \text{E} \end{cases}$$

```
> predict(reg1,newdata=data.frame(carburant=c("D","E")
   ,exposition=1),type="response")
         1          2
0.07971533 0.06656323
```

Fréquence d'Accident : x_2

Modèle : $Y_i \sim \mathcal{P}(\lambda_i e_i)$ où $\lambda_i = \exp(\beta_0 + \beta_j \mathbf{1}(x_i) = j)$,
x_i est la zone $\{A, B, C, D, E, F\}$

```
> with(base,table(zone,nb))
     nb
zone      0     1     2     3     4    16
   A   7411   231     9     0     1     0
   B   5442   170     5     0     0     0
   C  13619   462    22     2     0     0
   D  10749   385    25     2     0     1
   E   9704   401    23     3     1     0
   F   1272    58     2     0     0     0
```

La zone A est la zone de référence ici.

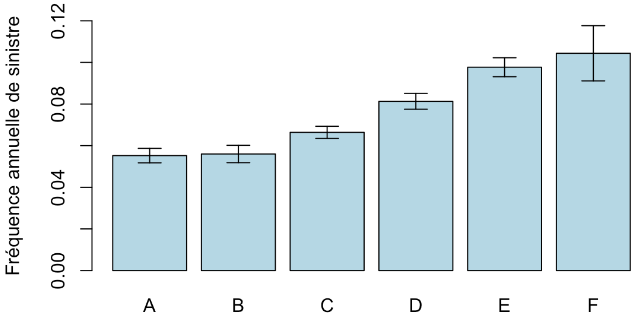
Modèle : $Y_i \sim \mathcal{P}(\lambda_i e_i)$ où $\lambda_i = \exp(\beta_0 + \beta_j \mathbf{1}(x_i) = j)$,
x_i est la zone $\{A, B, C, D, E, F\}$

```
> reg2 = glm(nb~zone+offset(log(exposition)),
+             data=base,family=poisson)
> summary(reg2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.89626    0.06287 -46.069  < 2e-16 ***
zoneB        0.01396    0.09751   0.143   0.8862
zoneC        0.18407    0.07685   2.395   0.0166 *
zoneD        0.38662    0.07836   4.934 8.07e-07 ***
zoneE        0.57025    0.07827   7.286 3.20e-13 ***
zoneF        0.63659    0.14171   4.492 7.05e-06 ***
```

Fréquence d'Accident : x_2

```
y=predict(reg2,newdata=data.frame(zone=LETTERS[1:6],
    exposition=1),type="response", se.fit =TRUE)
error.bar = function(x,y,upper, lower,length=0.1,...){
  arrows(x,y+upper, x, y-lower, angle=90, code=3,
    length=length, ...)}
barres = barplot(y$fit, beside=TRUE, col="light blue")
error.bar(barres,y$fit,y$se.fit)
```



Fréquence d'Accident : x_2

```
> M = matrix(NA,6,6)
> for(i in 1:6){
  base$zone = relevel(base$zone,LETTERS[i])
  reg2i = glm(nb~zone+offset(log(exposition)),
              data=base,family=poisson)
  p=summary(reg2i)$coefficients[2:6,4]
  names(p)=substr(names(p),5,5)
  M[i,] = p[LETTERS[1:6]]
}
> round(M,3)
      A     B     C     D     E     F
A    NA 0.886 0.017 0.000 0.000 0.000
B 0.886    NA 0.050 0.000 0.000 0.000
C 0.017 0.050    NA 0.002 0.000 0.001
D 0.000 0.000 0.002    NA 0.005 0.065
E 0.000 0.000 0.000 0.005    NA 0.624
F 0.000 0.000 0.001 0.065 0.624    NA
```

Fréquence d'Accident : x_2

```
> base$zone = factor(base$zone, levels = LETTERS[1:6])
> levels(base$zone) = c("AB","AB","C","D","EF","EF")
> reg2b = glm(nb~zone+offset(log(exposition)),data=base,family=poisson)
> summary(reg2b)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.89048    0.04806 -60.148  < 2e-16 ***
zoneC        0.17829    0.06529   2.731  0.00632 **
zoneD        0.38084    0.06706   5.679 1.36e-08 ***
zoneEF       0.57212    0.06500   8.802  < 2e-16 ***
```

Fréquence d'Accident : $x_1 + x_2$

On considère ici la régression sur deux variables catégorielles,

```
> (E=xtabs(exposition~carburant+zone,data=base))
         zone
carburant        AB        C        D       EF
        D  4307.494 3745.807 2471.297 1994.951
        E  3487.379 3966.076 3150.068 3308.060
> (N=xtabs(nb~carburant+zone,data=base))
         zone
carburant  AB   C   D  EF
        D 261 295 210 232
        E 172 217 247 290
> reg12 = glm(nb~carburant+zone+offset(log(exposition)
   ),data=base,family=poisson)
> yp = predict(reg12,type="response")
> xtabs(yp~base$carburant+base$zone)
              base$zone
base$carburant       AB        C        D       EF
             D 264.4026 279.1828 228.0513 226.3634
             E 168.5974 232.8172 228.9487 295.6366
```

Fréquence d'Accident : $x_1 + x_2$

Méthode des marges, $N_{i,j} \sim \mathcal{P}(E_{i,j} \cdot \lambda_{i,j})$ où $\lambda_{i,j} = A_i \cdot B_j$ où

$$\sum_i E_{i,j} A_i B_j = \sum_i N_{i,j} \text{ et } \sum_j E_{i,j} A_i B_j = \sum_j N_{i,j}$$

soit

$$A_i = \frac{\sum_j N_{i,j}}{\sum_j E_{i,j} B_j} \text{ et } B_j = \frac{\sum_i N_{i,j}}{\sum_i E_{i,j} A_i}$$

```
A=rep(1,length(levels(base$carburant)))
B=rep(1,length(levels(base$zone)))*sum(N)/sum(E)
for(i in 1:1000){
  A=apply(N,1,sum)/apply(t(B*t(E)),1,sum)
  B=apply(N,2,sum)/apply(A*E,2,sum)
}
```

Fréquence d'Accident : $x_1 + x_2$

```
> B
        AB          C          D         EF
0.05447450 0.06614477 0.08189545 0.10069925
> E * A%*%t(B)
         zone
carburant       AB        C        D       EF
        D 264.4026 279.1828 228.0513 226.3634
        E 168.5974 232.8172 228.9487 295.6366
```

$$\sum_i E_{i,j} A_i B_j = \sum_i N_{i,j} \text{ et } \sum_j E_{i,j} A_i B_j = \sum_j N_{i,j}$$

```
> apply(N,2,sum)
 AB   C   D  EF
433 512 457 522
> apply(E * A%*%t(B),2,sum)
 AB   C   D  EF
433 512 457 522
```

Fréquence d'Accident : x_3

Modèle : $Y_i \sim \mathcal{P}(\lambda_i e_i)$ où $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$, x_i est l'âge

```
> reg = glm(nb~ageconducteur+offset(log(exposition)),
+           data=base,family=poisson)
> summary(reg)

Coefficients:
               Estimate Std Error z value Pr(>|z|)
(Intercept)   -2.163061  0.077380  -27.95  < 2e-16 ***
ageconducteur -0.009891  0.001635   -6.05 1.45e-09 ***

(Dispersion parameter taken to be 1)

    Null deviance: 12680  on 49999  degrees of freedom
Residual deviance: 12642  on 49998  degrees of freedom
AIC: 16317

Number of Fisher Scoring iterations: 6
```

Aussi $\widehat{\beta}_0 = -2.1631$ et $\widehat{\beta}_1 = -0.0099$.

Fréquence d'Accident : x_3

Modèle : $Y_i \sim \mathcal{P}(\lambda_i e_i)$ où $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$, x_i est l'âge.

Ici $\widehat{\beta}_0 = -2.1631$ et $\widehat{\beta}_1 = -0.0099$.

$\widehat{\eta}_i = -2.1631 - 0.0099 x_i$ et $\widehat{\lambda}_i = e^{\widehat{\eta}_i} = e^{-2.1631 - 0.0099 x_i}$

On peut aussi utiliser des splines, $\log \lambda_i = \beta_0 + s(x_i)$

```
> library(splines)
> regs = glm(nb~bs(ageconducteur)+
     offset(log(exposition)),data=base,family=poisson)
```

$$s(x) = \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - s_1)_+^3 + \beta_5 (x - s_2)_+^3$$

(splines cubiques)

Fréquence d'Accident : x_3

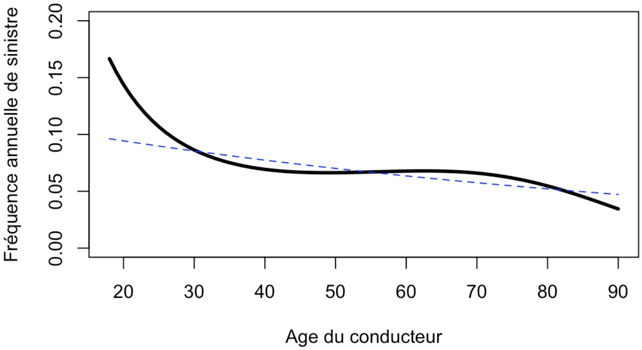
Modèle : $Y_i \sim \mathcal{P}(\lambda_i e_i)$ où $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$, x_i est l'âge

```
a = seq(18,90)
y = predict(regs,newdata=data.frame(ageconducteur=a,
    exposition=1),type="response", se.fit = TRUE)
plot(a,y$fit,type="l")
```

Fréquence d'Accident

Modèle : $Y_i \sim \mathcal{P}(\lambda_i e_i)$ où $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$, x_i est l'âge

```
y0=predict(reg,newdata=data.frame(ageconducteur=a,
    exposition=1),type="response")
lines(a,y0,lty=2)
```



Comparaison avec le modèle linéaire

Fréquence d'Accident : x_3

Modèle : $Y_i \sim \mathcal{P}(\lambda_i e_i)$ où $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$, x_i est l'âge

```
points(a,y$fit,pch=19)
segments(a,y$fit-2*y$se.fit,a,y$fit+2*y$se.fit)
```