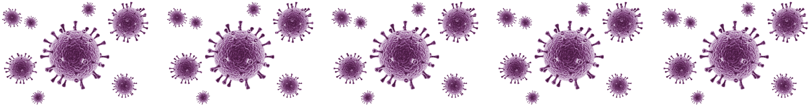
Modèles Linéaires Appliqués / Régression
Régression de Poisson : Interprétations

Arthur Charpentier

UQAM

Hiver 2020 - COVID-19 # 9



Données de Comptage

Base de données, Fair (1978)

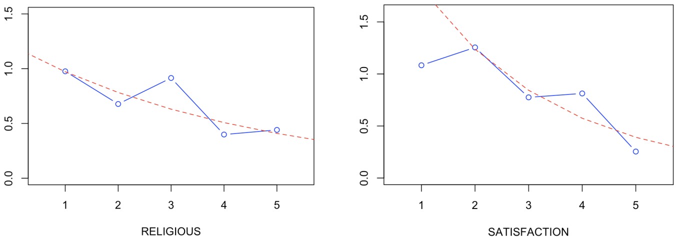
```
> loc = "http://freakonometrics.free.fr/baseaffairs.txt"
> base = read.table(loc,header=TRUE)
> str(base)
'data.frame': 563 obs. of  9 variables:
 $ SEX         : int  1 0 0 1 1 0 0 1 0 1 ...
 $ AGE         : num  37 27 32 57 22 32 22 57 32 ...
 $ YEARMARRIAGE: num  10 4 15 15 0.75 1.5 0.75   ...
 $ CHILDREN    : int  0 0 1 1 0 0 0 1 1 0 ...
 $ RELIGIOUS   : int  3 4 1 5 2 2 2 2 4 4 ...
 $ EDUCATION   : int  18 14 12 18 17 17 12 14 16 ...
 $ OCCUPATION  : int  7 6 1 6 6 5 1 4 1 4 ...
 $ SATISFACTION: int  4 4 4 5 3 5 3 4 2 5 ...
 $ Y           : int  0 0 0 0 0 0 0 0 0 0 ...
```

`OCCUPATION` : échelle d'Hollingshead

Données de Comptage

RELIGIOUS : entre 1 (anti-religieuse) à 5 (très religieuse)
SATISFACTION : de très mécontente (1) à très contente (5)

```
> A = with(base, aggregate(Y,by=list(RELIGIOUS),mean))
> A$x
[1] 0.9761905 0.6776316 0.9152542 0.3989071 0.4411765
> reg = glm(Y~RELIGIOUS,family=poisson)
> predict(reg,type="response",
              newdata=data.frame(RELIGIOUS=1:5))
        1         2         3         4         5
0.9717580 0.7828098 0.6306007 0.5079870 0.4092143
```

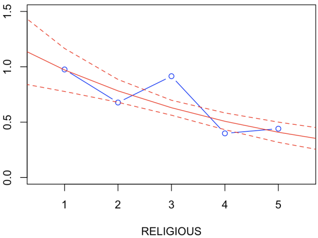


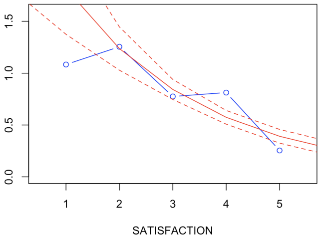
Données de Comptage

RELIGIOUS : entre 1 (anti-religieuse) à 5 (très religieuse)

SATISFACTION : de très mécontente (1) à très contente (5)

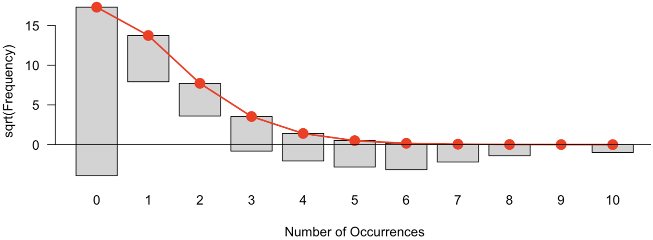
```
> predict(reg,type="response",newdata=data.frame(
   RELIGIOUS=1:5),se.fit=TRUE)
$fit
        1         2         3         4         5 
0.9717580 0.7828098 0.6306007 0.5079870 0.4092143 

$se.fit
        1         2         3         4         5 
0.0971302 0.0515588 0.0337331 0.0378823 0.0456359 
```





Y ne suit pas une loi de Poisson

```
> library(vcd)
> gof = goodfit(base$Y, type = "poisson", method = "ML
   ", par = NULL)
> plot(gof)
```

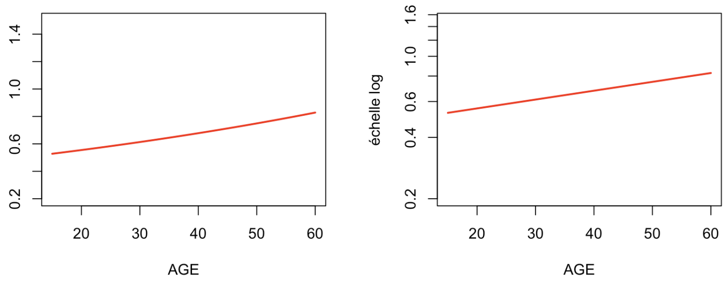


$\rightarrow$ ici on suppose que $Y|\boldsymbol{X} = \boldsymbol{x}$ suit une loi de Poisson

Une Variable Continue x_1

$$\widehat{\lambda}_x = \exp\left[\widehat{\beta}_0 + \widehat{\beta}_1 x\right]$$

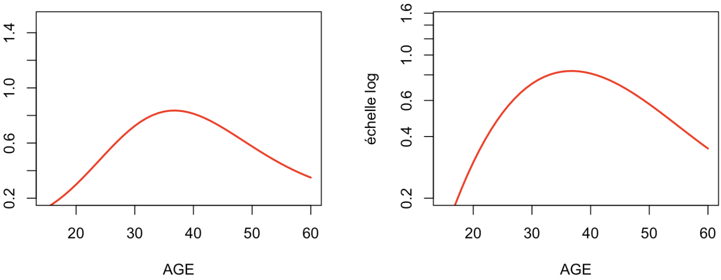
```
> reg = glm(Y~AGE,data=base,family=poisson)
> y = predict(reg,type="response",newdata=data.frame(
    AGE=15:60))
> plot(15:60,y,type="l")
> plot(15:60,y,type="l",log="y")
```



Une Variable Continue x_1

$$\widehat{\lambda}_x = \exp\left[\widehat{\beta}_0 + \widehat{h}(x)\right],\ h(x) = \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - s_1)_+^3 + \cdots$$

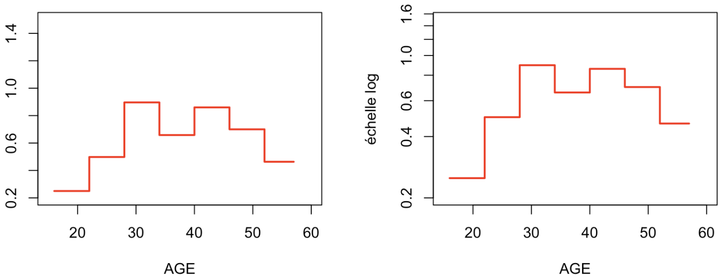
```
> library(splines)
> reg = glm(Y~bs(AGE),data=base,family=poisson)
> nd = data.frame(AGE=15:60)
> y = predict(reg,type="response",newdata=nd)
> plot(15:60,y,type="l")
> plot(15:60,y,type="l",log="y")
```



Une Variable Continue x_1

$$\widehat{\lambda}_x = \exp\left[\widehat{\beta}_0 + \widehat{h}(x)\right],\ h(x) = \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - s_1)_+^3 + \cdots$$

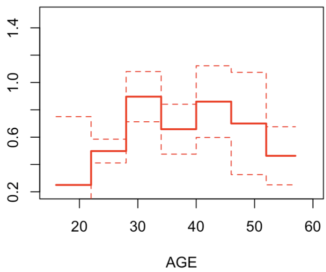
```
> library(splines)
reg = glm(Y~cut(AGE,breaks=seq(15,60,by=6)),data=base,
    family=poisson)
> y = predict(reg,type="response",newdata=nd)
> plot(15:60,y,type="l")
> plot(15:60,y,type="l",log="y")
```

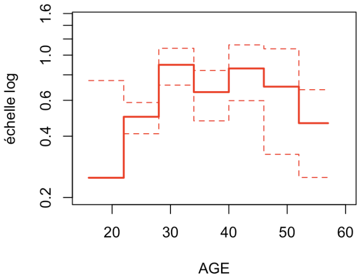


Une Variable Continue x_1

$$\widehat{\lambda}_x = \exp\left[\widehat{\beta}_0 + \widehat{h}(x)\right],\;\; h(x) = \sum_j \beta_j \mathbf{1}(x \in [a_j, a_{j+1}))$$

```
> library(splines)
reg = glm(Y~cut(AGE,breaks=seq(15,60,by=6)),data=base,
    family=poisson)
> y = predict(reg,type="response",newdata=nd, se.fit=
    TRUE)
> plot(15:60,y$fit,type="l")
```

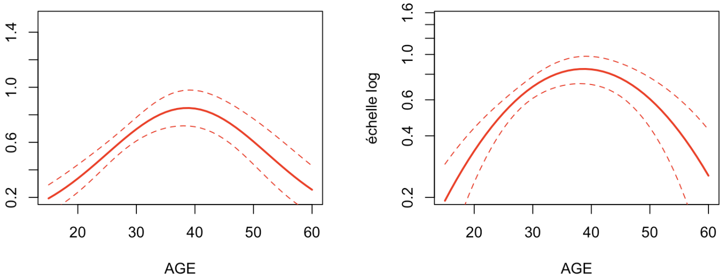




Une Variable Continue x_1

$$\widehat{\lambda}_x = \exp\left[\widehat{\beta}_0 + \widehat{h}(x)\right],\ h(x) = \beta_1 x + \beta_2 x^2$$

```
> library(splines)
reg = glm(Y~poly(AGE,2),data=base,family=poisson)
> y = predict(reg,type="response",newdata=nd)
> plot(15:60,y$fit,type="l")
```



Une Variable Continue x_1

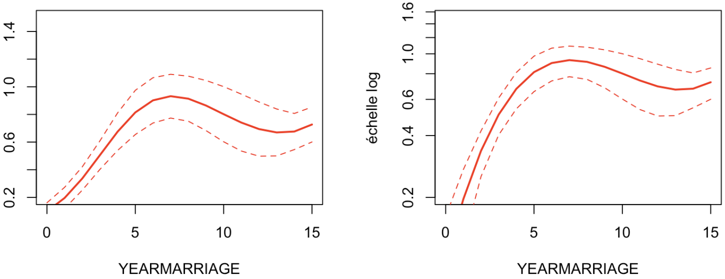
Quel degré pour une transformation polynomiale ?

```
> summary(glm(Y~poly(AGE,3),data=base,family=poisson))

Coefficients:
              Estimate Std. Error z value Pr(>z)
(Intercept)   -0.50257    0.05546  -9.062  < 2e-16 ***
poly(AGE, 3)1  2.63037    1.46107   1.800   0.0718 .
poly(AGE, 3)2 -6.50300    1.43051  -4.546 5.47e-06 ***
poly(AGE, 3)3  1.20600    1.39372   0.865   0.3869

> summary(glm(Y~poly(AGE,2),data=base,family=poisson))

Coefficients:
              Estimate Std. Error z value Pr(>z)
(Intercept)    -0.5006     0.0553  -9.053  < 2e-16 ***
poly(AGE, 2)1   2.3005     1.4207   1.619    0.105
poly(AGE, 2)2  -6.4849     1.4500  -4.472 7.73e-06 ***
```

Une Variable Continue x_2

$$\widehat{\lambda}_x = \exp\left[\widehat{\beta}_0 + \widehat{h}(x)\right], \text{ estimé sur une base de splines}$$

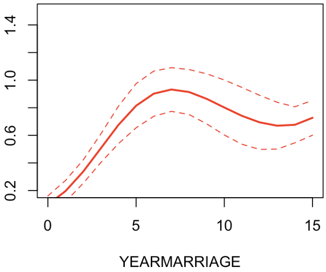
```
> library(splines)
reg = glm(Y~bs(YEARMARRIAGE),data=base,family=poisson)
> y = predict(reg,type="response",newdata=nd)
> plot(0:15,y$fit,type="l")
```

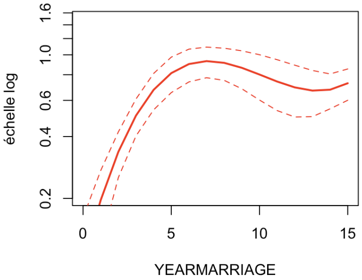


Une Variable Continue x_2

$$\widehat{\lambda}_x = \exp\left[\widehat{\beta}_0 + \widehat{h}(x)\right], \text{ polynôme de degré 3}$$

```
> library(splines)
reg = glm(Y~poly(YEARMARRIAGE,3),data=base,family=
    poisson)
> y = predict(reg,type="response",newdata=nd)
> plot(0:15,y$fit,type="l")
```

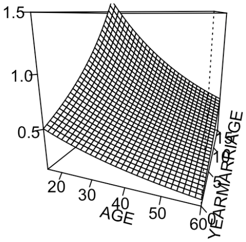


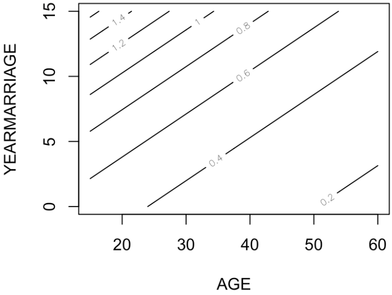


Deux Variables Continues $x_1 + x_2$

$$\widehat{\lambda}_{x_1,x_2} = \exp\left[\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2\right]$$

```
> reg = glm(Y~AGE+YEARMARRIAGE,data=base,family=
    poisson)
```

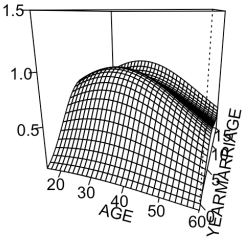


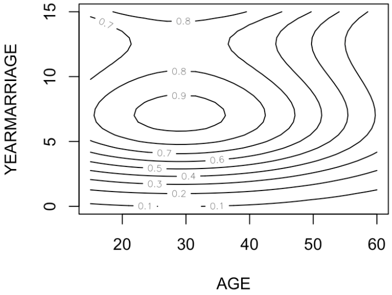


Deux Variables Continues $x_1 + x_2$

$$\widehat{\lambda}_{x_1,x_2} = \exp\left[\widehat{\beta}_0 + \widehat{p}_1(x_1) + \widehat{p}_2(x_2)\right]$$

```
> reg = glm(Y~poly(AGE,2)+poly(YEARMARRIAGE,3),data=
    base,family=poisson)
```





Deux Variables Continues $x_1 + x_2$

Les variables x_1 et x_2 sont (a priori) corrélées

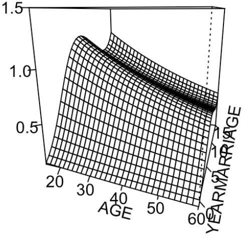
```
> reg = glm(Y~poly(AGE,2)+poly(YEARMARRIAGE,3),data=
    base,family=poisson)
> summary(reg)

Coefficients:
             Estimate Std. Error z value Pr(>z)
(Intercept)  -0.59675    0.06264  -9.527  < 2e-16 ***
poly(AGE,2)1 -3.09579    2.25503  -1.373  0.16980
poly(AGE,2)2 -2.22771    1.60747  -1.386  0.16579
poly(YM,3)1  10.72769    2.48109   4.324 1.53e-05 ***
poly(YM,3)2  -8.44615    1.58266  -5.337 9.47e-08 ***
poly(YM,3)3   4.28586    1.40078   3.060  0.00222 **
```

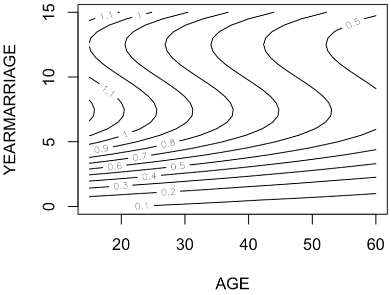
Le polynôme en x_1 n'est plus de degré 2

Deux Variables Continues $x_1 + x_2$

$$\widehat{\lambda}_{x_1,x_2} = \exp\left[\widehat{\beta}_0 + \widehat{p}_1(x_1) + \widehat{p}_2(x_2)\right]$$

```
> reg = glm(Y~AGE+poly(YEARMARRIAGE,3),data=base,
    family=poisson)
```





Deux Variables Continues $x_1 + x_2$

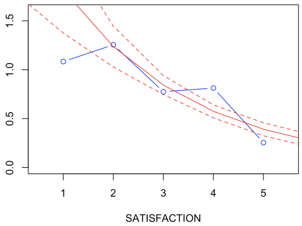
Les variables x_1 et x_2 sont (a priori) corrélées

```
> reg = glm(Y~poly(AGE,2)+poly(YEARMARRIAGE,3),data=
    base,family=poisson)
> summary(reg)

Coefficients:
               Estimate Std. Error z value Pr(>z)
(Intercept)    0.007094   0.305991   0.023  0.98150
AGE           -0.018604   0.009443  -1.970  0.04882 *
poly(YM, 3)1  12.249160   2.264986   5.408 6.37e-08 ***
poly(YM, 3)2  -8.900831   1.549487  -5.744 9.23e-09 ***
poly(YM, 3)3   4.284325   1.404584   3.050  0.00229 **
```

Une Variable Catégorielle x_1

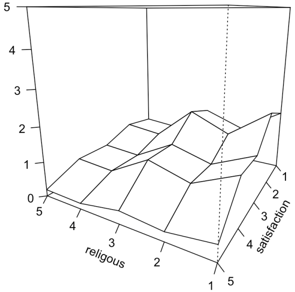
```
> table(base$SATISFACTION,base$Y)

       0   1   2   3   4   5   6   7   8  10
  1    8   0   2   1   0   0   1   0   0   0
  2   33   3   1   6   3   2   2   1   0   0
  3   66   8   4   2   4   2   1   1   1   0
  4  146  10   5   8   2   5   6   3   1   1
  5  198  13   5   2   3   2   1   0   0   0
```

```
> aggregate(base$Y,by=list(
    base$SATISFACTION),mean)$x
[1] 1.083 1.255 0.775 0.813
    0.254
```



Deux Variables Catégorielles $x_1 + x_2$

```
> reg = glm(Y~as.factor(RELIGIOUS)+
    as.factor(SATISFACTION), data=
    base,family=poisson)
```



$x_1 \in \{a_1, \cdots, a_I\}$
$x_2 \in \{b_1, \cdots, b_J\}$
$\rightarrow 1 + (I-1) + (J-1)$ paramètres
$(\beta_0, \beta_{1:2}, \cdots, \beta_{1:I}, \beta_{2:2}, \cdots, \beta_{2:J})$

$\widehat{\lambda}_{1,1} = e^{\widehat{\beta}_0}$ pour (a_1, b_1) (modalités de référence)
$\widehat{\lambda}_{1,j} = e^{\widehat{\beta}_0 + \widehat{\beta}_{2:j}} = e^{\widehat{\beta}_0} \cdot e^{\widehat{\beta}_{2:j}}$ pour (a_1, b_j)
$\widehat{\lambda}_{i,1} = e^{\widehat{\beta}_0 + \widehat{\beta}_{1:i}} = e^{\widehat{\beta}_0} \cdot e^{\widehat{\beta}_{1:i}}$ pour (a_i, b_1)
$\widehat{\lambda}_{i,j} = e^{\widehat{\beta}_0 + \widehat{\beta}_{1:i} + \widehat{\beta}_{2:j}} = e^{\widehat{\beta}_0} \cdot e^{\widehat{\beta}_{1:i}} \cdot e^{\widehat{\beta}_{2:j}}$ pour (a_i, b_j)

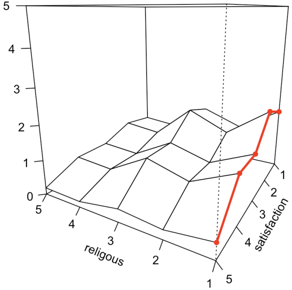
Deux Variables Catégorielles $x_1 + x_2$

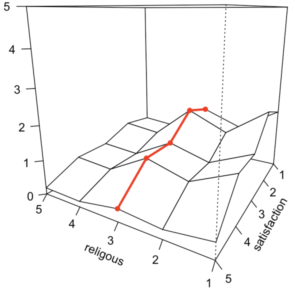
$$\boldsymbol{u} = (e^{\widehat{\beta}_0}, e^{\widehat{\beta}_0+\widehat{\beta}_{2:2}}, \cdots, e^{\widehat{\beta}_0+\widehat{\beta}_{2:5}})$$

prévision pour $x_1 = 1$: $\boldsymbol{u}$

prévision pour $x_1 = i$: $e^{\widehat{\beta}_{1:i}} \cdot \boldsymbol{u}$

```
Coefficients:
                          Estimate
(Intercept)                 0.5875
as.factor(SATISFACTION)2    0.1754
as.factor(SATISFACTION)3   -0.2882
as.factor(SATISFACTION)4   -0.2670
as.factor(SATISFACTION)5   -1.4232

> as.numeric(exp(beta[1])*c(1,exp(
    beta[6:9])))
[1] 1.800 2.145 1.349 1.378 0.434
```





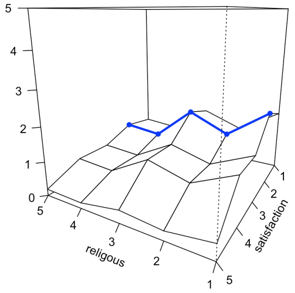
Deux Variables Catégorielles $x_1 + x_2$

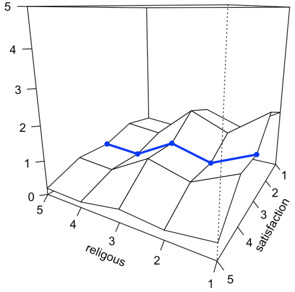
$$\boldsymbol{v} = \left(e^{\widehat{\beta}_0}, e^{\widehat{\beta}_0+\widehat{\beta}_{1:2}}, \cdots, e^{\widehat{\beta}_0+\widehat{\beta}_{1:5}}\right)$$

prévision pour $x_2 = 1$: $\boldsymbol{v}$

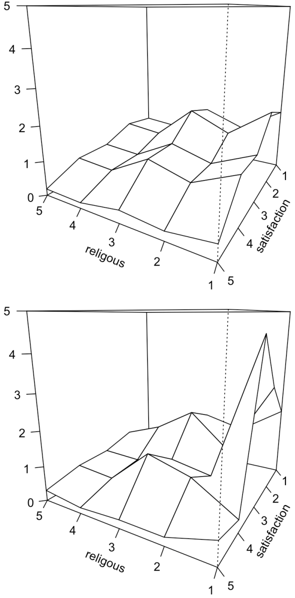
prévision pour $x_2 = j$: $e^{\widehat{\beta}_{2:j}} \cdot \boldsymbol{v}$

```
Coefficients:
                          Estimate
(Intercept)                 0.5875
as.factor(RELIGIOUS)2      -0.5159
as.factor(RELIGIOUS)3      -0.1747
as.factor(RELIGIOUS)4      -0.9792
as.factor(RELIGIOUS)5      -0.8373

> as.numeric(exp(beta[1])*c(1,exp(
    beta[2:5])))
[1] 1.800 1.074 1.511 0.676 0.779
```





Effets Croisés $x_1 * x_2$




```
> reg = glm(Y~as.factor(RELIGIOUS)*
    as.factor(SATISFACTION), data=
    base,family=poisson)
```

25 coefficients (5×5 modalités croisées)