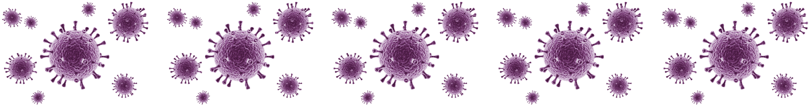
Modèles Linéaires Appliqués / Régression
Modèle de Poisson : Méthode des Marges

Arthur Charpentier

UQAM

Hiver 2020 - COVID-19 # 10



Méthode des Marges

Supposons que l'on prenne en compte ici deux classes de risques. Tableau de contingence et biais minimial, Bailey (1963) et Mildenhall (1999)

On suppose que

$$n_{i,j} \approx e_{i,j}\ L_i C_j,\ \text{i.e. } \boldsymbol{n} \approx \boldsymbol{e} \cdot \boldsymbol{L}\boldsymbol{C}^\top$$

Une idée naturelle pour estimer $\boldsymbol{L} = (L_i)$ et $\boldsymbol{C} = (C_j)$ est d'utiliser une régression de Poisson,

$$N_{i,j} \sim \mathcal{P}(e_{i,j} L_i C_j) \text{ ou } N_{i,j} \sim \mathcal{P}(e_{i,j} \exp[\ell_i + \gamma_j])$$

```
> loc = "http://freakonometrics.free.fr/baseaffairs.
    txt"
> base = read.table(loc,header=TRUE)
```

Méthode des Marges

```
> (E=xtabs((Y>=0)~as.factor(RELIGIOUS)+as.factor(SATISFACTION),data=base))
                      as.factor(SATISFACTION)
as.factor(RELIGIOUS)  1  2  3  4  5
                   1  1  1  4 17 19
                   2  5 19 24 48 56
                   3  2 12 20 38 46
                   4  3 14 30 63 73
                   5  1  5 11 21 30
> (N=xtabs(Y~as.factor(RELIGIOUS)+as.factor(SATISFACTION),data=base))
                      as.factor(SATISFACTION)
as.factor(RELIGIOUS)  1  2  3  4  5
                   1  2  3 18  6 12
                   2  6 17 14 53 13
                   3  3 23 16 55 11
                   4  2 17 15 27 12
                   5  0  4  6 11  9
```

Méthode des Marges

Régression sur données individuelles :

```
> reg12 = glm(Y~as.factor(RELIGIOUS)+as.factor(SATISFACTION),data=base,family=poisson)
> yp = predict(reg12,type="response")
> xtabs(yp~as.factor(RELIGIOUS)+as.factor(SATISFACTION),data=base)
                   as.factor(SATISFACTION)
(RELIGIOUS)      1      2      3      4      5
          1  1.800  2.145  5.396 23.423  8.238
          2  5.371 24.325 19.327 39.482 14.495
          3  3.022 21.610 22.654 43.965 16.748
          4  2.028 11.278 15.200 32.605 11.889
          5  0.779  4.642  6.423 12.525  5.631
```

$$N_{i,j}=\sum_{k=1}^{n}\widehat{\lambda}_k\mathbf{1}(x_{k,1}=i,x_{k,2}=j)$$

Méthode des Marges

Régression sur données individuelles :

```
> A=as.numeric(exp(coefficients(reg12)[1]+c(0,
    coefficients(reg12)[2:5])))
> B=as.numeric(exp(c(0,coefficients(reg12)[6:9])))
> as.numeric(A)
[1] 1.7995049 1.0742896 1.5111070 0.6759457 0.7789961
> as.numeric(B)
[1] 1.0000000 1.1917506 0.7495891 0.7656542 0.2409377
> E* (A%*%t(B))
                    as.factor(SATISFACTION)
(RELIGIOUS)      1      2      3      4      5
          1  1.800  2.145  5.396 23.423  8.238
          2  5.371 24.325 19.327 39.482 14.495
          3  3.022 21.610 22.654 43.965 16.748
          4  2.028 11.278 15.200 32.605 11.889
          5  0.779  4.642  6.423 12.525  5.631
```

Méthode des Marges

Régression sur données groupées : $N_{i,j} \sim \mathcal{P}(e_{i,j} L_i C_j)$,

```
> B=data.frame(N=as.numeric(N),E=as.numeric(E),RELIGIOUS=rep(1:5,5),SATISFACTION=rep(1:5,each=5))
> reg12g = glm(N~as.factor(RELIGIOUS)+as.factor(SATISFACTION)+offset(log(E)),data=B,family=poisson)
> matrix(predict(reg12g,type="response"),5,5)
      [,1]   [,2]   [,3]   [,4]   [,5]
[1,] 1.800  2.145  5.396 23.423  8.238
[2,] 5.371 24.325 19.327 39.482 14.495
[3,] 3.022 21.610 22.654 43.965 16.748
[4,] 2.028 11.278 15.200 32.605 11.889
[5,] 0.779  4.642  6.423 12.525  5.631
```

$$n_{i,j} = e_{i,j} \widehat{L}_i \widehat{C}_j, \text{ i.e. } \boldsymbol{n} = \boldsymbol{e} \cdot \boldsymbol{L}\boldsymbol{C}^\top)$$

Méthode des Marges

Oublions un instant la régression de Poisson,
Cherchons $\boldsymbol{L} = (L_i)$ et de $\boldsymbol{C} = (C_j)$ tels que $\boldsymbol{n} \approx \boldsymbol{e} \cdot \boldsymbol{L}\boldsymbol{C}^\top$
L'estimation de $\boldsymbol{L} = (L_i)$ et de $\boldsymbol{C} = (C_j)$ se fait généralement de trois manières:

- par moindres carrés
- par minimisation d'une distance (e.g. du chi-deux)
- par un principe de balancement (ou méthode des marges)

Méthode des Marges

Il est possible d'utiliser une méthode par moindres carrés (pondérée). On va chercher à minimiser la somme des carrés des erreurs, i.e.

$$D = \sum_{i,j} e_{i,j}(n_{i,j} - L_i C_j)^2$$

La condition du premier ordre donne ici

$$\frac{\partial D}{\partial L_i} = -2\sum_{j} C_j e_{i,j}(n_{i,j} - L_i C_j) = 0$$

$$L_i = \frac{\sum_j C_j e_{i,j} n_{i,j}}{\sum_j e_{i,j} C_j^2} = \frac{\sum_j c_j y_{i,j}}{\sum_j e_{i,j} C_j^2}$$

L'autre condition du premier ordre donne

$$C_j = \frac{\sum_i L_i e_{i,j} n_{i,j}}{\sum_i e_{i,j} L_i^2} = \frac{\sum_i L_i y_{i,j}}{\sum_i e_{i,j} L_i^2}$$

On résoud alors ce petit systeme de maniere itérative (car il n'y a pas de solution analytique simple).

Il est aussi possible d'utiliser une méthode basée sur la distance du chi-deux. On va chercher à minimiser

$$Q = \sum_{i,j} \frac{e_{i,j}(n_{i,j} - L_i C_j)^2}{L_i C_j}$$

Là encore on utilise les conditions du premier ordre, et on obtient

$$L_i = \left(\frac{\sum_j \left(\frac{e_{i,j} y_{i,j}^2}{C_j} \right)}{\sum_j e_{i,j} C_j} \right)^{\frac{1}{2}} \text{ et } L_j = \left(\frac{\sum_i \left(\frac{e_{i,j} y_{i,j}^2}{L_i} \right)}{\sum_i e_{i,j} L_i} \right)^{\frac{1}{2}}$$

où $y_{i,j} = e_{i,j} n_{i,j}$ (que l'on résout itérativement).

Méthode des Marges

Dans la méthode des marges – Bailey (1963) – formellement, on veut

$$\sum_j y_{i,j} = \sum_j e_{i,j} n_{i,j} = \sum_j e_{i,j} L_i C_j,$$

en somment sur la ligne i, pour tout i, ou sur la colonne j,

$$\sum_i y_{i,j} = \sum_i e_{i,j} n_{i,j} = \sum_i e_{i,j} L_i C_j,$$

La première, et la seconde, équation donnent respectivemnet

$$L_i = \frac{\sum_j y_{i,j}}{\sum_j e_{i,j} C_j} \text{ et } C_j = \frac{\sum_i y_{i,j}}{\sum_i e_{i,j} L_i}.$$

Cette solution... correspond à la régression de Poisson.

Méthode des Marges

La première, et la seconde, équation donnent respectivement

$$L_i = \frac{\sum_j y_{i,j}}{\sum_j e_{i,j} C_j} \text{ et } C_j = \frac{\sum_i y_{i,j}}{\sum_i e_{i,j} L_i}.$$

```
> SATISFACTION=as.factor(base$SATISFACTION)
> RELIGIOUS=as.factor(base$RELIGIOUS)
> A=rep(1,length(levels(SATISFACTION)))
> B=rep(1,length(levels(RELIGIOUS)))*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) }
> E * A%*%t(B)
                        as.factor(SATISFACTION)
(RELIGIOUS)          1       2       3      4       5
            1    1.800   2.145   5.396 23.423   8.238
            2    5.371  24.325  19.327 39.482  14.495
            3    3.022  21.610  22.654 43.965  16.748
            4    2.028  11.278  15.200 32.605  11.889
            5    0.779   4.642   6.423 12.525   5.631
```

Méthode des Marges

Cette solution vérifie

$$\sum_j y_{i,j} = \sum_j e_{i,j} L_i C_j, \text{ et } \sum_i y_{i,j} = \sum_i e_{i,j} L_i C_j,$$

```
> N0 = E * A%*%t(B)
> apply(N0,1,sum)
  1   2   3   4   5
 41 103 108  73  30
> apply(N,1,sum)
  1   2   3   4   5
 41 103 108  73  30
> apply(N0,2,sum)
  1   2   3   4   5
 13  64  69 152  57
> apply(N,2,sum)
  1   2   3   4   5
 13  64  69 152  57
```