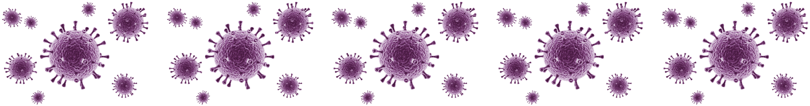
Modèles Linéaires Appliqués / Régression
GLM & Poids et Modèles Tweedie

Arthur Charpentier

UQAM

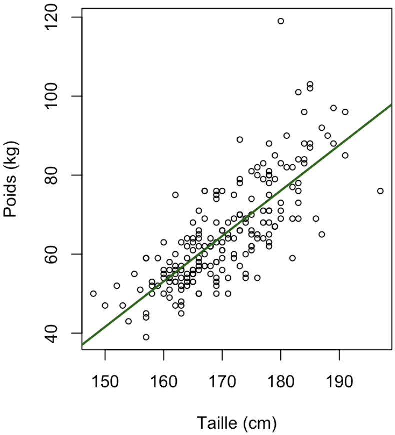
Hiver 2020 - COVID-19 # 16



Les Poids en Régression

Dans la régression classique (Gaussienne) on cherche à résoudre

$$\widehat{\beta} = \text{argmin}\left\{\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top\beta\right)^2\right\}$$



Avec des poids $\boldsymbol{\omega} = (\omega_i)$

$$\widehat{\beta}_{\boldsymbol{\omega}} = \text{argmin}\left\{\sum_{i=1}^{n}\omega_i\left(y_i - \boldsymbol{x}_i^\top\beta\right)^2\right\}$$

Problème des données agrégées,
poids = taille de la population

cf exemple taille et poids des élèves agrégés ici par 'province' ou 'région'

Les Poids en Régression

$$\widehat{\beta}_{\omega} = \text{argmin}\left\{\sum_{i=1}^{n} \omega_i (y_i - \boldsymbol{x}_i^\top \beta)^2\right\}$$

donne la condition du premier ordre, en dérivant par rapport à β_j

$$\sum_{i=1}^{n} \omega_i x_{j,i}(y_i - \boldsymbol{x}_i^\top \beta) = 0,\ \forall j,\ \text{ou } \boldsymbol{X}^\top \boldsymbol{\Omega}(\boldsymbol{y} - \boldsymbol{X}\beta) = \boldsymbol{0}$$

où $\boldsymbol{\Omega} = \text{diag}(\omega_i)$, soit $\widehat{\beta} = (\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{y}$

Dans les GLM, la log-vraisemblance est alors

$$\log \mathcal{L}(\boldsymbol{\theta}, \varphi | \boldsymbol{y}) = \sum_{i=1}^{n} \omega_i \frac{y_i \theta_i - b(\theta_i)}{\varphi} + \text{ terme indépendent de } \theta_i$$

Les Poids en Régression

- Régression de Poisson : poids vs. exposition (offset)

Exposition e_i: $\log \mathcal{L}(\beta; \boldsymbol{y}) = \sum_{i=1}^{n} [y_i \log(e_i \lambda_i) - e_i \lambda_i - \log(y_i!)]$

```
> reg = glm(Y~YEARMARRIAGE+offset(log(EXPOSITION)),
    data=DF,  family=poisson)
```

Poids y_i/e_i: $\log \mathcal{L}^\star(\beta; \boldsymbol{y}) = \sum_{i=1}^{n} e_i [\frac{y_i}{e_i} \log(\lambda_i) - \lambda_i - \log(y_i!)]$

```
> reg = glm(Y/EXPOSITION~YEARMARRIAGE, data=DF,
    weights=EXPOSITION, family=poisson)
```

Les deux modèles sont semblables...

Les Poids en Régression

- Régression Binomiale : poids vs. taille - $Y \sim \mathcal{B}(n, p)$

```
> reg = glm(cbind(Y,n-Y)~X1+X2, data=DF, binomial)
```

Déboublement $(y_i, x_{1,i}, x_{2,i}, n_i) \rightarrow n_i$ lignes
$(1, x_{1,i}, x_{2,i}) \quad \times y_i$ et $(0, x_{1,i}, x_{2,i}) \quad \times n_i - y_i$

```
> reg = glm(y~X1+X2, data=bigDF, family=binomial)
```

Poids modéliser y_i/n_i avec un poids n_i

```
> reg = glm(y/n~X1+X2, data=DF, binomial, weights=n)
```

Les trois modèles sont semblables...

Les Poids en Régression

- Tableau de contigence, y vs x

```
> df = read.table("http://freakonometrics.free.fr/
    baseexo.csv",sep=";",header=TRUE)
> (M=as.matrix(df[,2:ncol(df)]))
M
   14 19 21 23 25 27 29 31 33 35
18 74 13  7  1  0  0  0  0  0  0
23  6  4  2  7  4  0  0  0  0  0
25  2  3  2  2  4  0  0  0  0  0
27  1  1  2  3  3  2  0  0  0  0
29  2  0  1  3  2  0  6  0  0  0
31  2  0  2  1  0  0  1  2  1  0
33  0  0  0  2  0  0  1  1  3  0
35  0  1  0  1  0  0  0  0  2  0
37  0  0  0  0  1  1  0  1  0  1
```

Poids (x, y) et le poids n

```
> W = as.vector(M)
> X = rep(df[,1],ncol(M))
> y = as.numeric(substr(names(df)[-1],2,3))
> Y = rep(y,each=nrow(M))
> cbind(X,Y,W)
   X  Y  W
1 16 14 74
2 23 14  6
3 25 14  2
4 27 14  1
5 29 14  2
> summary(lm(Y ~ X, weights = W))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.35569    2.03022   2.145    0.038 *
X            0.68263    0.09016   7.572 3.04e-09 ***
```

Duplication (x, y) répété n fois

```
> vX = vY = rep(NA,sum(W))
> sumW = c(0,cumsum(W))
> for(i in 1:length(W)){
   if(W[i]>0){
     vX[(1+sumW[i]):sumW[i+1]] = X[i]
     vY[(1+sumW[i]):sumW[i+1]] = Y[i] }}
> base2 = data.frame(x=vX,y=vY)
> tail(base2,10)
     x  y
179 35 33
180 35 33
181 37 35
> summary(lm(y ~ x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.35569    0.95972   4.538 1.04e-05 ***
X2           0.68263    0.04262  16.017  < 2e-16 ***
```

Consider a Tweedie distribution, with variance function power $p \in (1,2)$, mean μ and scale parameter ϕ, then it is a compound Poisson model,

- $N \sim \mathcal{P}(\lambda)$ with $\lambda = \dfrac{\phi\mu^{2-p}}{2-p}$
- $Y_i \sim \mathcal{G}(\alpha,\beta)$ with $\alpha = -\dfrac{p-2}{p-1}$ and $\beta = \dfrac{\phi\mu^{1-p}}{p-1}$

Conversely, consider a compound Poisson model $N \sim \mathcal{P}(\lambda)$ and $Y_i \sim \mathcal{G}(\alpha,\beta)$, then

- variance function power is $p = \dfrac{\alpha+2}{\alpha+1}$
- mean is $\mu = \dfrac{\lambda\alpha}{\beta}$
- scale parameter is $\phi = \dfrac{[\lambda\alpha]^{\frac{\alpha+2}{\alpha+1}-1}\beta^{2-\frac{\alpha+2}{\alpha+1}}}{\alpha+1}$

Modèle Tweedie

In the context of regression

$$N_i \sim \mathcal{P}(\lambda_i) \text{ with } \lambda_i = \exp[\boldsymbol{x}_i^\top \boldsymbol{\beta}_\lambda]$$

$$Y_{j,i} \sim \mathcal{G}(\mu_i, \phi) \text{ with } \mu_i = \exp[\boldsymbol{x}_i^\top \boldsymbol{\beta}_\mu]$$

Then $S_i = Y_{1,i} + \cdots + Y_{N,i}$ has a Tweedie distribution

- variance function power is $p = \dfrac{\phi + 2}{\phi + 1}$
- mean is $\lambda_i \mu_i$
- scale parameter is $\dfrac{\lambda_i^{\frac{1}{\phi+1}-1}}{\mu_i^{\frac{\phi}{\phi+1}}} \left(\dfrac{\phi}{1+\phi} \right)$

There are $1 + 2\dim(\boldsymbol{X})$ degrees of freedom.
Remark Note that the scale parameter should not depend on i.

Modèle Tweedie

A Tweedie regression is

- variance function power is $p \in (1, 2)$
- mean is $\mu_i = \exp[\boldsymbol{x}_i^\mathsf{T} \boldsymbol{\beta}_{\text{Tweedie}}]$
- scale parameter is ϕ

There are $2 + \dim(\boldsymbol{X})$ degrees of freedom.