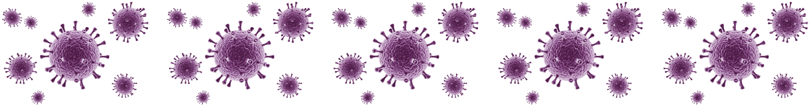
Modèles Linéaires Appliqués / Régression
Régression de Poisson : Inférence

Arthur Charpentier

UQAM

Hiver 2020 - COVID-19 # 8



Régression de Poisson

Données homogènes $\{y_1, \cdots, y_n\}$ de loi de Poisson, $Y_i \sim \mathcal{P}(\lambda)$

$$\mathbb{P}(Y = y) = e^{-\lambda} \frac{\lambda^y}{y!}, \forall y \in \mathbb{N}.$$

Données hétérogènes $\{(y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n)\}$ de loi de Poisson

$$Y_i \sim \mathcal{P}(\lambda_i) \text{ avec } \lambda_i = \exp[\boldsymbol{x}_i^\top \boldsymbol{\beta}].$$

Dans ce modèle, $\mathbb{E}(Y_i|\boldsymbol{x}_i) = \text{Var}(Y_i|\boldsymbol{x}_i) = \lambda_i = \exp[\boldsymbol{x}_i^\top \boldsymbol{\beta}]$.

Modèle dit log-linéaire.

Remarque: on posera parfois $\theta_i = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ et $\mu_i = \lambda_i = \exp[\eta_i]$.

Régression de Poisson

La log-vraisemblance est ici

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

i.e.

$$\log \mathcal{L}(\beta; \boldsymbol{y}) = \sum_{i=1}^{n} y_i \cdot [\boldsymbol{x}_i^\top \beta] - \exp[\boldsymbol{x}_i^\top \beta] - \log(y_i!)$$

Le gradient est ici

$$\nabla \log \mathcal{L}(\beta; \boldsymbol{y}) = \frac{\partial \log \mathcal{L}(\beta; \boldsymbol{y})}{\partial \beta} = \sum_{i=1}^{n} \boldsymbol{x}_i^\top (y_i - \exp[\boldsymbol{x}_i^\top \beta])$$

alors que la matrice Hessienne s'écrit

$$H(\beta) = \frac{\partial^2 \log \mathcal{L}(\beta; \boldsymbol{y})}{\partial \beta \partial \beta^\top} = -\sum_{i=1}^{n} \exp[\boldsymbol{x}_i^\top \beta] \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

Régression de Poisson

La condition du premier ordre est ici

$$\boldsymbol{X}^\top[\boldsymbol{y} - \exp(\boldsymbol{X}\widehat{\beta})] = \boldsymbol{0}, \text{ i.e. } \sum_{i=1}^{n} \boldsymbol{x}_i^\top[\boldsymbol{y} - \exp(\boldsymbol{x}_i^\top\widehat{\beta})] = \boldsymbol{0}$$

ce qui signifie $\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{X}^\top\widehat{\boldsymbol{\lambda}}$ où $\widehat{\boldsymbol{\lambda}} = \exp(\boldsymbol{X}\widehat{\beta})$.

La matrice Hessienne est $H(\beta) = -\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$, $\boldsymbol{\Omega} = -\text{diag}\big(\exp(\boldsymbol{X}\widehat{\beta})\big)$

Cf slides #2, en posant $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{p}(1-\boldsymbol{p}))$,

$$\nabla \log \mathcal{L}(\beta) = \frac{\partial \log \mathcal{L}(\beta)}{\partial \beta} = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{p}) = \boldsymbol{0} \quad (\text{avec } \boldsymbol{p} = \boldsymbol{p}(\beta))$$

$$H(\beta) = \frac{\partial^2 \log \mathcal{L}(\beta)}{\partial \beta \partial \beta^\top} = -\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X} \quad (\text{avec } \boldsymbol{\Omega} = \boldsymbol{\Omega}(\beta))$$

Régression de Poisson

Parfois, les données ne sont pas observées pendant la même durée

Données hétérogènes $\{(y_1, \boldsymbol{x}_1, e_1), \cdots, (y_n, \boldsymbol{x}_n, e_n)\}$ de loi de Poisson (cf #7 processus de Poisson)

$$Y_i \sim \mathcal{P}(e_i\lambda_i) \text{ avec } \tilde{\lambda}_i = e_i\lambda_i = \exp[\boldsymbol{x}_i^\top \beta + \log(e_i)].$$

$\log(e_i)$ est appelée variable offset.

Parfois les données sont regroupées (cf #9 régression binomiale $\mathcal{B}(n,p)$) et on suppose les groupes homogènes

$$Y_i = \sum_{j=1}^{e_i} Y_{i,j} \text{ avec } Y_{i,j} \sim \mathcal{P}(\lambda_i), \text{ i.i.d., i.e. } Y_i \sim \mathcal{P}(e_i\lambda_i).$$

Modèles Log-Linéaires

Ici $Y_i \sim \mathcal{P}(\exp[\boldsymbol{x}_i^\top \beta])$.

On pourra considérer (#12) $Y_i \sim \mathcal{N}(\exp[\boldsymbol{x}_i^\top \beta], \sigma^2)$, mais

$$\underbrace{Y_i \sim \mathcal{N}(\exp[\boldsymbol{x}_i^\top \beta], \sigma^2)}_{\text{modèle GLM sur } y_i} \neq \underbrace{\log Y_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \beta, \sigma^2)}_{\text{modèle LM sur } \log y_i}$$

Dans le premier cas, $\mathbb{E}(Y_i) = \exp[\boldsymbol{x}_i^\top \beta]$

Dans le second cas, $\mathbb{E}(Y_i) = \exp\left[\boldsymbol{x}_i^\top \beta + \dfrac{\sigma^2}{2}\right]$

Cf loi lognormale, si $Y \sim \mathcal{LN}(\mu, \sigma^2)$, $\mathbb{E}(Y) = \exp\left[\mu + \dfrac{\sigma^2}{2}\right]$

Cf inégalité de Jensen, $\mathbb{E}(\log X) \leq \log \mathbb{E}(\log X)$ car log concave.

Cas de données $N_{i,j}$ en 'triangle'...

```
> library(boot)
> data(aids)
> head(aids)
  year quarter delay dud time y
1 1983       3     0   0    1 2
2 1983       3     2   0    1 6
3 1983       3     5   0    1 0
```

Ici `year:quarter` = date (i) et `delay` = développement (j)

```
[33,]   53  175   35   17   13   11    2
[34,]   63  135   24   23   12    1   NA
[35,]   71  161   48   25    5   NA   NA
[36,]   95  178   39    6   NA   NA   NA
[37,]   76  181   16   NA   NA   NA   NA
[38,]   67   66   NA   NA   NA   NA   NA
```

$\rightarrow$ utiliser un facteur ligne $\boldsymbol{L} = (L_1, \cdots, L_I)$
et un facteur colonne $\boldsymbol{C} = (C_1, \cdots, C_J)$

Modèles Log-Linéaires

Ici $N_{i,j} \sim \mathcal{P}(L_i \cdot C_j)$ ou $N_{i,j} \sim \mathcal{P}(\exp(\ell_i + \gamma_j))$

```
> Aids = aids
> Aids$y[Aids$dud==1] = NA
> reg = glm(y~as.factor(delay)+as.factor(time),data=
    Aids,family=poisson)
```

On peut ensuite prévoir, $\widehat{\lambda}_{i,j} = \exp(\widehat{\ell}_i + \widehat{\gamma}_j)$

```
> p = predict(reg,newdata=Aids,type="response")
> matrix(p,length(unique(Aids$time)),byrow = TRUE)
[33,] 53.8 150.9 43.3 26.5 16.0 13.5 10.1  7.7
[34,] 47.6 133.5 38.3 23.4 14.2 11.9  8.9  6.8
[35,] 59.8 167.7 48.1 29.4 17.8 15.0 11.2  8.6
[36,] 67.7 189.9 54.5 33.3 20.2 16.9 12.7  9.7
[37,] 67.5 189.5 54.4 33.2 20.1 16.9 12.7  9.7
[38,] 67.0 188.0 53.9 33.0 19.9 16.8 12.6  9.6
```