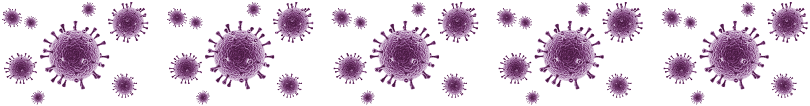
Modèles Linéaires Appliqués / Régression
Régression Logistique: Extensions

Arthur Charpentier

UQAM

Hiver 2020 - COVID-19 # 3



Régression Bernouilli $y = \mathbf{1}_A$

```
> reg1 = glm((Survived==1)~Pclass+Sex+Age+I(Age^2)+I(Age^3)+SibSp,  family=binomial, data=base)
> summary(reg1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.616e+00  6.565e-01   8.554  < 2e-16 ***
Pclass2     -1.360e+00  2.842e-01  -4.786  1.7e-06 ***
Pclass3     -2.557e+00  2.853e-01  -8.962  < 2e-16 ***
Sexmale     -2.658e+00  2.176e-01 -12.216  < 2e-16 ***
Age         -1.905e-01  5.528e-02  -3.446 0.000569 ***
I(Age^2)     4.290e-03  1.854e-03   2.314 0.020669 *
I(Age^3)    -3.520e-05  1.843e-05  -1.910 0.056188 .
SibSp       -5.041e-01  1.317e-01  -3.828 0.000129 ***
> predict(reg1)[1]
-2.592995
> predict(reg1,type="response")[1]
0.06959063
```

Régression Bernoulli $y = \mathbf{1}_{A^C}$

```
> reg0 = glm((Survived==0)~Pclass+Sex+Age+I(Age^2)+I(
    Age^3)+SibSp, family=binomial, data=base)
> summary(reg0)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.616e+00  6.565e-01  -8.554  < 2e-16 ***
Pclass2      1.360e+00  2.842e-01   4.786  1.7e-06 ***
Pclass3      2.557e+00  2.853e-01   8.962  < 2e-16 ***
Sexmale      2.658e+00  2.176e-01  12.216  < 2e-16 ***
Age          1.905e-01  5.528e-02   3.446 0.000569 ***
I(Age^2)    -4.290e-03  1.854e-03  -2.314 0.020669 *
I(Age^3)     3.520e-05  1.843e-05   1.910 0.056188 .
SibSp        5.041e-01  1.317e-01   3.828 0.000129 ***

> predict(reg0)[1]
2.592995
> predict(reg0,type="response")[1]
0.9304094
```

Régression Binomiale

Au lieu de $Y_i \sim \mathcal{B}(p_i)$, $Y_i \sim \mathcal{B}(n_i, p_i)$ où n_i est connnue.

$$\mathbb{E}\left(\frac{Y_i}{n_i}\right) = p_i = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$$

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

On pose $z_i = y_i / n_i$, dont la densité est

$$f(y_i, p_i) = \binom{n_i}{n_i y_i} \exp\left[n_i y_i \log\left(\frac{p}{1-p}\right) + n_i \log(1-p)\right]$$

et on estime $\boldsymbol{\beta}$ par maximum de vraisemblance

Régression Multinomiale

Pour une loi de Bernoulli, $y \in \{0, 1\}$,

$$\mathbb{P}(Y = 1) = \frac{e^{\boldsymbol{x}^\top \beta}}{1 + e^{\boldsymbol{x}^\top \beta}} = \frac{p_1}{p_0 + p_1} \text{ et } \mathbb{P}(Y = 0) = \frac{1}{1 + e^{\boldsymbol{x}^\top}} = \frac{p_0}{p_0 + p_1}$$

Pour une loi multinomiale, $y \in \{A, B, C\}$, $\boldsymbol{y} = (\mathbf{1}_A, \mathbf{1}_B, \mathbf{1}_C)$

$$\mathbb{P}(Y = A) = \frac{p_A}{p_A + p_B + p_C} \propto p_A \text{ i.e. } \mathbb{P}(Y = A) = \frac{e^{\boldsymbol{x}^\top \beta_A}}{e^{\boldsymbol{x}^\top \beta_B} + e^{\boldsymbol{x}^\top \beta_B} + 1}$$

$$\mathbb{P}(Y = B) = \frac{p_B}{p_A + p_B + p_C} \propto p_B \text{ i.e. } \mathbb{P}(Y = B) = \frac{e^{\boldsymbol{x}^\top \beta_B}}{e^{\boldsymbol{x}^\top \beta_A} + e^{\boldsymbol{x}^\top \beta_B} + 1}$$

$$\mathbb{P}(Y = C) = \frac{p_C}{p_A + p_B + p_C} \propto p_C \text{ i.e. } \mathbb{P}(Y = C) = \frac{1}{e^{\boldsymbol{x}^\top \beta_A} + e^{\boldsymbol{x}^\top \beta_B} + 1}$$

Régression Multinomiale

On va essayer de comprendre la classe $y \in \{1, 2, 3\}$ sur les données du Titanic

```
> loc = "http://freakonometrics.free.fr/titanic.RData"
> download.file(loc_fichier, "titanic.RData")
> load("titanic.RData")
> regclass = multinom(Pclass ~ Sex+Age+SibSp, base)
> regclass

Coefficients:
  (Intercept)   Sexmale         Age       SibSp
2    1.416426 0.2662196 -0.04526865 -0.2150871
3    2.420469 1.0330840 -0.07541502 -0.1149161

Residual Deviance: 1347.672
AIC: 1363.672
```

Régression Multinomiale

```
> (b = coefficients(regclass))
  (Intercept)    Sexmale          Age      SibSp
2     1.416426 0.2662196 -0.04526865 -0.2150871
3     2.420469 1.0330840 -0.07541502 -0.1149161
```

Avec ici β_2 et β_3 (la class 1 est la référence)

```
> newbase = data.frame(Sex="female",Age =60,SibSp=0)
> predict(regclass,newdata=newbase,"probs")
         1          2          3
0.71708728 0.19548915 0.08742357
```

Idée : comme la class 1 est la référence,

$$\mathbb{P}(Y=1)\propto 1,\ \mathbb{P}(Y=2)\propto e^{\boldsymbol{x}^\top\beta_2} \text{ et } \mathbb{P}(Y=3)\propto e^{\boldsymbol{x}^\top\beta_3}$$

Régression Multinomiale

$$\mathbb{P}(Y=1) \propto 1,\ \mathbb{P}(Y=2) \propto e^{\boldsymbol{x}^\top \boldsymbol{\beta}_2} \text{ et } \mathbb{P}(Y=3) \propto e^{\boldsymbol{x}^\top \boldsymbol{\beta}_3}$$

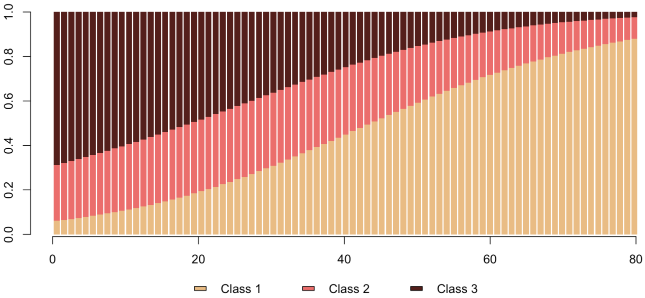
```
> x = c(1,0,60,0)
> b = rbind(rep(0,ncol(b)),b)
> t(exp(b%*%x))
     1         2         3
[1,] 1 0.2726156 0.1219148
```

$$\mathbb{P}(Y=1)\frac{1}{e^{\boldsymbol{x}^\top \boldsymbol{\beta}_2}+e^{\boldsymbol{x}^\top \boldsymbol{\beta}_3}+1},\ \mathbb{P}(Y=2)=\frac{e^{\boldsymbol{x}^\top \boldsymbol{\beta}_A}}{e^{\boldsymbol{x}^\top \boldsymbol{\beta}_B}+e^{\boldsymbol{x}^\top \boldsymbol{\beta}_B}+1},\cdots$$

```
> t(exp(b%*%x))/sum(exp(b%*%x))
             1         2          3
[1,] 0.7170873 0.1954892 0.08742357
```

Régression Multinomiale

```
> x = cbind(1,0,0:80,0)
> p2 = exp(apply((x%*%b[2,]),1,sum))
> p3 = exp(apply((x%*%b[3,]),1,sum))
> pp2 = p2/(1+p2+p3)
> pp3 = p3/(1+p2+p3)
> p = rbind(1-pp2-pp3,pp2,pp3)
> barplot(p)
```



Régression Multinomiale

Considérons une approche alternative : régressions Bernoulli itérées considérons un premier modèle de Bernoulli $y_1 = \mathbf{1}_A$

```
> reg1 = glm((Pclass==1) ~ Sex+Age+SibSp, base, family
    =binomial)
```

considérons un premier modèle de Bernoulli $y_2 = \mathbf{1}_B$, entre les classes B et C

```
> reg2 = glm((Pclass==2) ~ Sex+Age+SibSp, base, family
    =binomial,subset = (Pclass!=1))
```

Idée : $\mathbb{P}(y = B) = \mathbb{P}(y = B|y \neq A) \cdot \mathbb{P}(y \neq A)$

```
> p11 = predict (reg1, newdata=base, type="response")
> p12 = predict (reg2, newdata=base, type="response")
> itp = cbind(p11,(1-p11)*p12,(1-p11)*(1-p12))
```

Régression Multinomiale

On peut comparer les modèles logit itérés (à gauche)
et le modèle multinomial (à droite)

```
> mmp = predict(regclass,newdata=base,"probs")
> head(cbind(itp,mmp))
      1     2     3         1     2     3
1 0.129 0.204 0.668   0.126 0.201 0.673
2 0.459 0.274 0.267   0.462 0.275 0.264
3 0.256 0.328 0.416   0.259 0.330 0.411
4 0.412 0.285 0.303   0.417 0.284 0.299
5 0.227 0.253 0.521   0.229 0.253 0.518
```