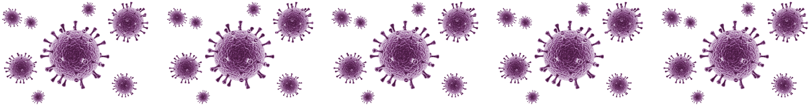
Modèles Linéaires Appliqués / Régression
Régression Logistique : x continue

Arthur Charpentier

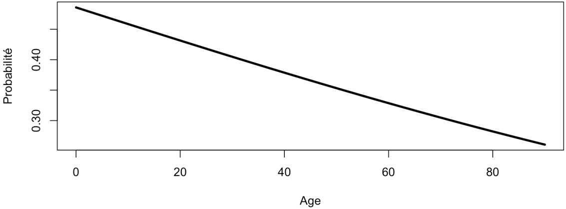
UQAM

Hiver 2020 - COVID-19 # 5



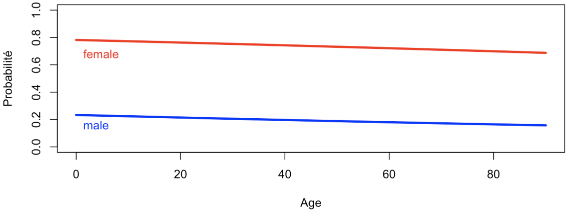
Régression sur x_1

```
> loc = "http://freakonometrics.free.fr/titanic.RData"
> download.file(loc, "titanic.RData")
> load("titanic.RData")
> base = base[!is.na(base$Age),1:7]
> reg = glm(Survived~Age,family=binomial,data=base)
> summary(reg)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.05672    0.17358  -0.327   0.7438
Age         -0.01096    0.00533  -2.057   0.0397 *
```



Régression sur $x_1 + \mathbf{1}_2$

```
> reg = glm(Survived~Age+Sex,family=binomial)

Coefficients:
             Estimate Std. Error z value Pr(>z)
(Intercept)  1.277273   0.230169   5.549 2.87e-08 ***
Age         -0.005426   0.006310  -0.860     0.39
Sexmale     -2.465920   0.185384 -13.302  < 2e-16 ***
```



Régression sur $x_1 + \mathbf{1}_2$

Prédiction : $\widehat{p}_{x,s} = \dfrac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 \mathbf{1}_M}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 \mathbf{1}_M}}$ où $s \in \{F, M\}$.

Pas de relation simple entre $\widehat{p}_{x,M}$ et $\widehat{p}_{x,F}$

ou $\text{cote}_{x,s} = \dfrac{\widehat{p}_{x,s}}{1 - \widehat{p}_{x,s}} = e^{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 \mathbf{1}_M}$, aussi,

$$\text{cote}_{x,M} = \text{cote}_{x,F} \cdot e^{\widehat{\beta}_2}$$

Régression sur x_1

Comme dans les ‘modèles linéaires’, on peut chercher une transformation non-linéaire de x_1

- régression polynomiale

```
> reg3=glm(Survived~poly(Age,3), family=binomial)
> summary(reg3)

Coefficients:
              Estimate Std. Error z value Pr(>z)
(Intercept)   -0.39327    0.07739  -5.082 3.74e-07 ***
poly(Age, 3)1 -4.81433    2.15139  -2.238  0.02524 *
poly(Age, 3)2  2.40500    2.29333   1.049  0.29432
poly(Age, 3)3 -6.83820    2.42311  -2.822  0.00477 **
```

$\text{logit}(p) = P(x_1)$ avec P polynôme de degré 3 (significatif).
avec `poly(,3)` écrit dans une base de polynômes orthogonaux

Régression sur x_1

écriture dans la base usuelle, $\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$

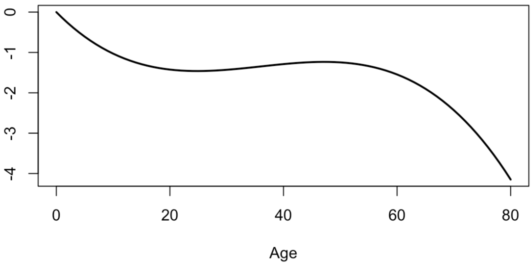
```
> reg3b=glm(Survived~Age+I(Age^2)+I(Age^3), family=
    binomial,data=base[!is.na(base$Age),])
> summary(reg3b)

Coefficients:
              Estimate Std. Error z value Pr(>z)
(Intercept)  9.258e-01  3.612e-01   2.563 0.010364 *
Age         -1.428e-01  4.078e-02  -3.501 0.000463 ***
I(Age^2)     4.392e-03  1.418e-03   3.096 0.001960 **
I(Age^3)    -4.069e-05  1.442e-05  -2.822 0.00477 **
```

Remarque La p-value du test $H_0 : \beta_3 = 0$ est identique...
Les deux modèles sont équivalents

Régression sur x_1

On peut visualiser $x_1 \mapsto \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x^2 + \widehat{\beta}_3 x_1^3$

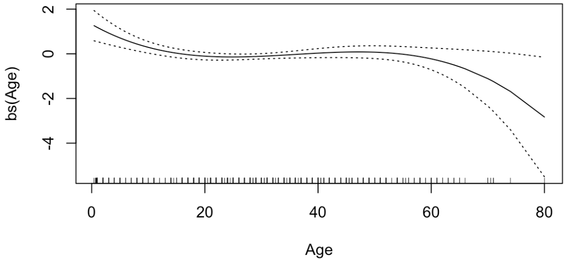
```
> beta = coefficients(reg3b)
> xage = 0:80
> yage = beta[2]*xage+beta[3]*xage^2+beta[4]*xage^3
> plot(xage,yage)
```



Régression sur x_1

- régression spline

```
> library(splines)
> res = glm(Survived~bs(Age),family=binomial)
> library(gam)
> regam=gam(Survived~bs(Age),family=binomial)
> plot(reg, se=TRUE)
```



Cf cas linéaire : splines cubiques $\cdots + (x_1 - s_1)_+^3 + +(x_1 - s_2)_+^3$