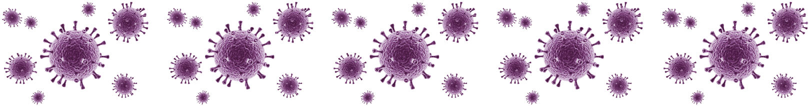
Modèles Linéaires Appliqués / Régression
GLM & Tests

Arthur Charpentier

UQAM

Hiver 2020 - COVID-19 # 18



Tests & Intervalles de Confiance

La déviance était définie par

$$D(\boldsymbol{y},\widehat{\boldsymbol{\mu}}) = 2\varphi\left[\log\mathcal{L}(\boldsymbol{y}) - \log\mathcal{L}(\widehat{\boldsymbol{\mu}})\right]$$

$$D^\star(\boldsymbol{y},\widehat{\boldsymbol{\mu}}) = 2\left[\log\mathcal{L}(\boldsymbol{y}) - \log\mathcal{L}(\widehat{\boldsymbol{\mu}})\right] = \frac{D}{\varphi}$$

Si on veut tester $H_0 : \boldsymbol{\beta}_I = \boldsymbol{0}$ pour un sous ensemble de variables, $I \subset \{1, 2, \cdots, k\}$, on fait un test de rapport de vraisemblance

$$LRT = D^\star(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\mu}}_I) = D^\star(\boldsymbol{y},\widehat{\boldsymbol{\mu}}_I) - D^\star(\boldsymbol{y},\widehat{\boldsymbol{\mu}})$$

Sous H_0, $LRT \overset{\mathcal{L}}{\to} \chi^2(k - \dim(I))$

Convergences Asymptotiques

On a vu que

Proposition sous les hypothèses mentionnées auparavant

- l'estimateur du maximum de vraisemblance $\widehat{\beta}$ existe et est unique,
- $\widehat{\beta} \overset{p.s.}{\rightarrow} \beta$ (fortement consistant)
- lorsque $n \rightarrow \infty$, $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \overset{\mathcal{L}}{\rightarrow} \mathcal{N}(\mathbf{0}, \varphi\boldsymbol{\Sigma})$

où $\boldsymbol{\Sigma} = (\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}$ et $\boldsymbol{W} = \text{diag}((V(\mu_i)g'(\mu_i)^2)^{-1})$.

Proposition sous les hypothèses mentionnées auparavant

- lorsque $n \rightarrow \infty$, $\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta} \overset{\mathcal{L}}{\rightarrow} \mathcal{N}(\mathbf{0}, \varphi\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top)$
- lorsque $n \rightarrow \infty$, $\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} \overset{\mathcal{L}}{\rightarrow} \mathcal{N}(\mathbf{0}, \varphi\boldsymbol{\Delta}^{-2}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top)$

Convergences Asymptotiques

- lorsque $n \to \infty$, $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \varphi \mathbf{\Sigma})$

```
> reg = glm(Survived~Sex+Pclass+Age,family="binomial")
> confint(reg)
                  2.5 %      97.5 %
(Intercept)  3.01450041  4.58890189
Sexmale     -2.93866348 -2.12456642
Pclass2     -1.86335813 -0.77181573
Pclass3     -3.14672951 -2.04176314
Age         -0.05229015 -0.02223298
```

- lorsque $n \to \infty$, $\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \varphi \mathbf{\Delta}^{-2} \boldsymbol{X} \mathbf{\Sigma} \boldsymbol{X}^\top)$

```
> predict(reg,newdata=newbase,se.fit = TRUE,type="
   response")
$fit
        1         2
0.9588402 0.1124358
$se.fit
         1          2
0.01217853 0.01834374
```

Convergences Asymptotiques

En réalité, on utilise ici la vraisemblance profilée

```
> confint(reg)
Waiting for profiling to be done...
              2.5 %        97.5 %
Age    -0.05229015 -0.02223298
```

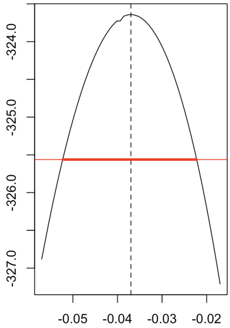
$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{k-1}, \beta_k)$. Pour β_k, on définit

$$\log \mathcal{L}_p(\beta_k) = \log \mathcal{L}((\boldsymbol{\beta}^\star_{-k}, \beta_k))$$

où $\boldsymbol{\beta}^\star_{-k} = \text{argmax}\{\log \mathcal{L}((\boldsymbol{\beta}_{-k}, \beta_k))\}$
L'intervalle de confiance est en :

$$\max\{\log \mathcal{L}_p(\beta_k)\} - \frac{1}{2} q_{\chi^2}(.95)$$

cf test du rapport de vraisemblance



Convergences Asymptotiques

Soit $\boldsymbol{R}$ une matrice $q \times p$ (avec $q \leq p$), alors

$$\frac{1}{\varphi}(\boldsymbol{R}(\widehat{\beta}-\beta))^\top\left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1}(\boldsymbol{R}(\widehat{\beta}-\beta))\xrightarrow{\mathcal{L}}\chi^2(n-p)$$

Si on veut tester $H_0 : \boldsymbol{R}\beta = \boldsymbol{r}_0$, contre $H_1 : \boldsymbol{R}\beta \neq \boldsymbol{r}_0$, on utilise

$$Q = \frac{1}{\varphi}(\boldsymbol{R}\widehat{\beta}-\boldsymbol{r}_0)^\top\left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1}(\boldsymbol{R}\widehat{\beta}-\boldsymbol{r}_0)$$

et sous H_0, $Q\xrightarrow{\mathcal{L}}\chi^2(n-p)$

Vraisemblance Profilée

```
> p0 = rep(0,4)
> profL = function(b){
  Lp = function(p){
    eta = X%*%c(p,b)
    mu = exp(eta)/(1+exp(eta))
    -sum(dbinom(reg$y,size = 1,prob=mu,
   log=TRUE))}
  -optim(par=p0,fn=Lp)$value }
> m=profL(coefficients(reg)[5])
> minf=m-qchisq(.95,df=1)/2
> abline(h=minf,col="red")
> (b1=uniroot(function(z) Vectorize(
   profL)(z)-minf,c(-.06,-.04))$root)
[1] -0.05227511
> (b2=uniroot(function(z) Vectorize(
   profL)(z)-minf,c(-.04,-.02))$root)
[1] -0.02225003
```