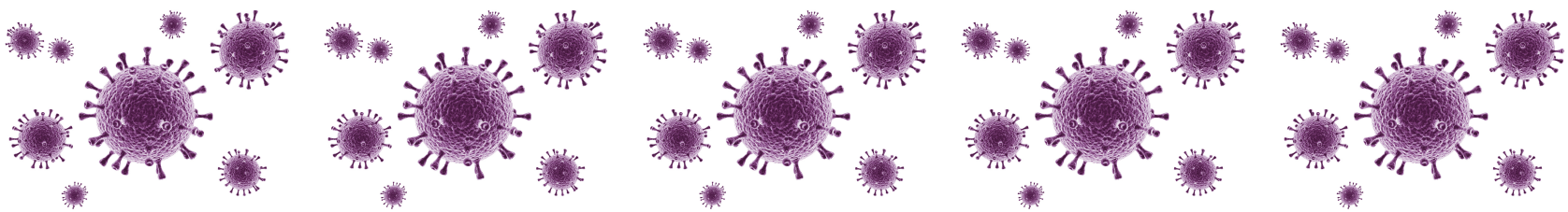
Modèles Linéaires Appliqués / Régression
Sélection de Variables & Stepwise

Arthur Charpentier

UQAM

Hiver 2020 - COVID-19 # 20



Stepwise - Procédure Pas à Pas

Retour sur la régression linéaire

```
> data(swiss)
> str(swiss)
'data.frame': 47 obs. of  6 variables:
 $ Fertility        : num  80.2 83.1 92.5 85.8 76.9 ...
 $ Agriculture      : num  17 45.1 39.7 36.5 43.5 ...
 $ Examination      : int  15 6 5 12 17 9 16 14 12 ...
 $ Education        : int  12 9 5 7 15 7 7 8 7 13 ...
 $ Catholic         : num  9.96 84.84 93.4 33.77 5 ...
 $ Infant           : num  22.2 22.2 20.2 20.3 20.6 ...
```

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_{k,i} x_{k,i} + \varepsilon_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

Test de Student

Idée classique : partir du modèle complet, enlever *la* variable la moins significative, et itérer

```
> summary(lm(Fertility~Examination+Agriculture+InfantM
    +Catholic+Education,data=swiss))

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 66.91518   10.70604   6.250 1.91e-07 ***
Examination -0.25801    0.25388  -1.016  0.31546
Agriculture -0.17211    0.07030  -2.448  0.01873 *
InfantM      1.07705    0.38172   2.822  0.00734 **
Catholic     0.10412    0.03526   2.953  0.00519 **
Education   -0.87094    0.18303  -4.758 2.43e-05 ***
```

(jusqu'à ce que toutes les variables soient significatives)

Stepwise - Procédure Pas à Pas

1. $y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_{k,i} x_{k,i} + \varepsilon_i$ et enlever une variable

```
> drop1(lm(Fertility~Examination+Agriculture+InfantM+
    Catholic+Education,data=swiss), test="F")
Single term deletions

Model:
Fertility ~ Examination + Agriculture + InfantM +
    Catholic + Education
           Df Sum of Sq  RSS    AIC F value    Pr(>F)
<none>                 2105.0 190.69
Examinati  1    53.03 2158.1 189.86  1.0328  0.3154
Agricultu  1   307.72 2412.8 195.10  5.9934  0.0187 *
Infant.M   1   408.75 2513.8 197.03  7.9612  0.0073 **
Catholic   1   447.71 2552.8 197.75  8.7200  0.0051 **
Education  1  1162.56 3267.6 209.36 22.6432 2.4e-05 ***
```

Stepwise - Procédure Pas à Pas

2. Partir de $y_i = \beta_0 + \varepsilon_i$ et rajouter une variable...

```
> add1(lm(Fertility~1, data=swiss), Fertility~
    Agriculture+Infant.Mortality+Catholic+Examination+
    Education , test="F")
Single term additions

Model:
Fertility ~ 1
           Df Sum of Sq  RSS    AIC F value  Pr(>F)
<none>                 7178.0 238.34
Agricultu   1    894.8 6283.1 234.09  6.4089  0.0149 *
Infant.M    1   1245.5 5932.4 231.39  9.4477  0.0035 **
Catholic    1   1543.3 5634.7 228.97 12.3251  0.0010 **
Examinati   1   2994.4 4183.6 214.97 32.2087 9.4e-07 ***
Education   1   3162.7 4015.2 213.04 35.4456 3.6e-07 ***
```

AIC & BIC

Akaike, $\text{AIC} = -2\log\mathcal{L}(\widehat{\beta}) + 2k$

Bayésien/Schwarz, $\text{BIC} = -2\log\mathcal{L}(\widehat{\beta}) + k\log(n)$

```
> n=nrow(swiss)
> reg = lm(Fertility ~ Agriculture + Catholic +
    Education,data=swiss)
> s2 = sum(resid(reg)^2)
> n + n*log(2*pi) + n*log(s2/n) + 2*(1+4)
[1] 331.4126
> -2*logLik(reg)+2*(1+4)
'log Lik.' 331.4126 (df=5)
> AIC(reg, k=2)
[1] 331.4126
> n + n*log(2*pi) + n*log(s2/n) + log(n)*(1+4)
[1] 340.6634
> -2*logLik(reg)+log(n)*(1+4)
'log Lik.' 340.6634 (df=5)
> AIC(reg, k=log(n))
[1] 340.6634
```

AIC & Stepwise

Akaike, $\text{AIC} = -2\log\mathcal{L}(\widehat{\beta}) + 2k$

```
> step(lm(Fertility~1, data=swiss), Fertility~
    Agriculture+Infant.Mortality+Catholic+Examination+
    Education,direction = "forward")
Start:  AIC=238.35
Fertility ~ 1

                Df Sum of Sq    RSS    AIC
+ Education      1    3162.7 4015.2 213.04
+ Examination    1    2994.4 4183.6 214.97
+ Catholic       1    1543.3 5634.7 228.97
+ Infant.M       1    1245.5 5932.4 231.39
+ Agriculture    1     894.8 6283.1 234.09
<none>                       7178.0 238.34
```

On enlève une variable + Education, et on continue

AIC & Stepwise

```
Step:  AIC=213.04
Fertility ~ Education

               Df Sum of Sq    RSS    AIC
+ Catholic      1    961.07 3054.2 202.18
+ Infant.M      1    891.25 3124.0 203.25
+ Examination   1    465.63 3549.6 209.25
<none>                      4015.2 213.04
+ Agriculture   1     61.97 3953.3 214.31

Step:  AIC=202.18
Fertility ~ Education + Catholic

               Df Sum of Sq    RSS    AIC
+ Infant.M      1    631.92 2422.2 193.29
+ Agriculture   1    486.28 2567.9 196.03
<none>                      3054.2 202.18
+ Examination   1      2.46 3051.7 204.15
```

AIC & Stepwise

```
Step:  AIC=193.29
Fertility ~ Education + Catholic + InfantM

               Df Sum of Sq    RSS    AIC
+ Agriculture   1   264.176 2158.1 189.86
<none>                      2422.2 193.29
+ Examination   1     9.486 2412.8 195.10

Step:  AIC=189.86
Fertility ~ Education + Catholic + InfantM + Agricult

               Df Sum of Sq    RSS    AIC
<none>                      2158.1 189.86
+ Examination   1    53.027 2105.0 190.69
```

et le modèle final est

```
lm(formula = Fertility ~ Education + Catholic + Infant
    .Mortality + Agriculture, data = swiss)
```

AIC & Stepwise

Akaike, $\text{AIC} = -2\log\mathcal{L}(\widehat{\beta}) + 2k$

```
> step(lm(Fertility~Agriculture+InfantM+Catholic+
    Examination+Education,data=swiss),direction = "
    backward")
Start:  AIC=190.69
Fertility ~ Agriculture + InfantM + Catholic +
    Examination +
    Education

               Df Sum of Sq    RSS    AIC
 - Examination  1     53.03 2158.1 189.86
<none>                     2105.0 190.69
- Agriculture  1    307.72 2412.8 195.10
- Infant.M     1    408.75 2513.8 197.03
- Catholic     1    447.71 2552.8 197.75
- Education    1   1162.56 3267.6 209.36
```

On enlève une variable - Examination, et on continue

AIC & Stepwise

```
Step:  AIC=189.86
Fertility ~ Agriculture + InfantM + Catholic +
    Education

               Df Sum of Sq    RSS    AIC
<none>                      2158.1 189.86
- Agriculture   1    264.18 2422.2 193.29
- InfantM       1    409.81 2567.9 196.03
- Catholic      1    956.57 3114.6 205.10
- Education     1   2249.97 4408.0 221.43
```

et le modèle final est (encore)

```
lm(formula = Fertility ~ Education + Catholic + Infant
    .Mortality + Agriculture, data = swiss)
```

AIC & BIC

Akaike, $\text{AIC} = -2\log\mathcal{L}(\widehat{\beta}) + 2k$

Bayésien/Schwarz, $\text{BIC} = -2\log\mathcal{L}(\widehat{\beta}) + k\log(n)$

... qui peuvent être aussi utilisés pour les GLM

```
> data("birthwt", package = "MASS")
> str(birthwt)
'data.frame': 189 obs. of  10 variables:
 $ low   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ age   : int  19 33 20 21 18 21 22 17 29 26 ...
 $ lwt   : int  182 155 105 108 107 124 118 103 123 ...
 $ race  : int  2 3 1 1 1 3 1 3 1 1 ...
 $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
 $ ptl   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ht    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ui    : int  1 0 0 1 1 0 0 0 0 0 ...
 $ ftv   : int  0 3 1 2 0 0 1 1 1 0 ...
 $ bwt   : int  2523 2551 2557 2594 2600 2622 2637 ...
```

AIC & BIC

```
> bwt <- with(birthwt, {
+     race <- factor(race, labels = c("white", "black
    ", "other"))
+     ptd <- factor(ptl > 0)
+     ftv <- factor(ftv)
+     levels(ftv)[-(1:2)] <- "2+"
+     data.frame(low = factor(low), age, lwt, race,
    smoke = (smoke > 0), ptd, ht = (ht > 0), ui = (ui
    > 0), ftv) })
> str(bwt)
'data.frame': 189 obs. of  9 variables:
 $ low   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1  ...
 $ age   : int  19 33 20 21 18 21 22 17 29 26 ...
 $ lwt   : int  182 155 105 108 107 124 118 103 123 ...
 $ race  : Factor w/ 3 levels "white","black",..:  ...
 $ smoke: logi  FALSE FALSE TRUE TRUE TRUE FALSE ...
 $ ptd   : Factor w/ 2 levels "FALSE","TRUE": 1 1  ...
 $ ht    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ..
```

AIC & Stepwise

On considère les deux modèles extrêmes

```
> fullmod = glm(low ~ ., data=bwt,family=binomial)
> zeromod = glm(low ~ 1, data=bwt,family=binomial)
```

Puis on peut avancer, pas à pas, à partir du modèle complet

```
> backwards = step(fullmod, direction = "backward")
Start:  AIC=217.48
low ~ age + lwt + race + smoke + ptd + ht + ui + ftv

        Df Deviance    AIC
- ftv    2   196.83 214.83
- age    1   196.42 216.42
<none>       195.48 217.48
- ui     1   197.59 217.59
- smoke  1   198.67 218.67
- race   2   201.23 219.23
- lwt    1   200.95 220.95
- ht     1   202.93 222.93
- ptd    1   203.58 223.58
```

AIC & Stepwise

```
Step:  AIC=214.83
low ~ age + lwt + race + smoke + ptd + ht + ui

       Df Deviance    AIC
- age   1   197.85 213.85
<none>      196.83 214.83
- ui    1   199.15 215.15
- race  2   203.24 217.24
- smoke 1   201.25 217.25
- lwt   1   201.83 217.83
- ptd   1   203.95 219.95
- ht    1   204.01 220.01

Step:  AIC=213.85
low ~ lwt + race + smoke + ptd + ht + ui

       Df Deviance    AIC
<none>      197.85 213.85
```

AIC & Stepwise

On peut aussi avancer, pas à pas, à partir du modèle simple

```
> forwards = step(zeromod, scope=list(lower=formula(
    zeromod),upper=formula(fullmod)), direction="
    forward")
Start:  AIC=236.67
low ~ 1

         Df Deviance    AIC
+ ptd     1   221.90 225.90
+ lwt     1   228.69 232.69
+ ui      1   229.60 233.60
+ smoke   1   229.81 233.81
+ ht      1   230.65 234.65
+ race    2   229.66 235.66
+ age     1   231.91 235.91
<none>        234.67 236.67
+ ftv     2   232.09 238.09
```

AIC & Stepwise

```
Step:  AIC=225.9
low ~ ptd

         Df Deviance    AIC
+ age     1   217.30 223.30
+ lwt     1   217.50 223.50
+ ht      1   217.66 223.66
+ race    2   217.02 225.02
+ ui      1   219.12 225.12
+ smoke   1   219.33 225.33
+ ftv     2   217.88 225.88
<none>        221.90 225.90

Step:  AIC=223.3
low ~ ptd + age

         Df Deviance    AIC
+ ht      1   213.12 221.12
+ lwt     1   214.31 222.31
```

AIC & Stepwise

```
Step:  AIC=221.12
low ~ ptd + age + ht

         Df Deviance    AIC
+ lwt     1   207.43 217.43
+ ui      1   210.13 220.13
+ smoke   1   210.89 220.89
<none>        213.12 221.12
+ race    2   210.06 222.06
+ ftv     2   210.38 222.38

Step:  AIC=217.43
low ~ ptd + age + ht + lwt

         Df Deviance    AIC
+ ui      1   205.15 217.15
+ smoke   1   205.39 217.39
<none>        207.43 217.43
```

AIC & Stepwise

```
Step:  AIC=217.15
low ~ ptd + age + ht + lwt + ui

         Df Deviance    AIC
<none>        205.15 217.15
+ smoke   1   203.24 217.24
+ race    2   201.25 217.25
+ ftv     2   202.41 218.41
```

On a ici deux modèles différents

```
> backwards
Call:  glm(formula = low ~ lwt + race + smoke + ptd +
    ht + ui, family = binomial, data = bwt)

> forwards
Call:  glm(formula = low ~ ptd + age + ht + lwt + ui,
    family = binomial, data = bwt)
```

AIC & Stepwise

On peut aussi utiliser une méthode mixte

```
> bothways = step(zeromod, list(lower=formula(zeromod)
    ,upper=formula(fullmod)), direction="both")
Start:  AIC=236.67
low ~ 1

         Df Deviance    AIC
+ ptd     1   221.90 225.90
+ lwt     1   228.69 232.69
+ ui      1   229.60 233.60
+ smoke   1   229.81 233.81
+ ht      1   230.65 234.65
+ race    2   229.66 235.66
+ age     1   231.91 235.91
<none>        234.67 236.67
+ ftv     2   232.09 238.09
```

AIC & Stepwise

(...)

```
Step:  AIC=217.15
low ~ ptd + age + ht + lwt + ui

         Df Deviance    AIC
<none>        205.15 217.15
+ smoke   1   203.24 217.24
+ race    2   201.25 217.25
- ui      1   207.43 217.43
- age     1   207.51 217.51
+ ftv     2   202.41 218.41
- lwt     1   210.13 220.13
- ht      1   212.70 222.70
- ptd     1   215.48 225.48
```

On retient ici le modèle

```
> bothways
Call:  glm(formula = low ~ ptd + age + ht + lwt + ui,
    family = binomial, data = bwt)
```

AIC & All Models

On peut aussi utiliser `regsubsets`,

```
> library(leaps)
> allreg = regsubsets(low ~ ., data = bwt, nbest = 1,
    nvmax = NULL, force.in = NULL, force.out = NULL,
    method = "exhaustive")
> as.data.frame(summary(allreg)$outmat)
          age lwt rak rar smo ptd htT uiT ft1 f2+
1   ( 1 )                     *
2   ( 1 )                     *   *
3   ( 1 )       *             *   *
4   ( 1 )       *             *   *       *
5   ( 1 )       *   *         *   *       *
6   ( 1 )       *   *         *   *   *   *
7   ( 1 )       *   *   *   * *   *   *
8   ( 1 )       *   *   *   * *   *   *   *
9   ( 1 )   *   *   *   *   * *   *   *   *
10  ( 1 )   *   *   *   *   * *   *   *   *   *
```

AIC & All Models

que l'on peut aussi visualiser dans un graphique

```
> plot(allreg, scale = "bic")
```