Pluie.Demain Prédiction
GLM et Choix de Modèle

ZhuQing ZHONG

08/01/2022

Table des matières

Préparation des données et Import des paquets

Importer les paquets necessaires

```
library(readr)
library(tidyverse)
library(xtable)
library(Hmisc)
library(corrplot)
library(MASS)
library(rJava)
library(glmulti)
library("FactoMineR")
library("factoextra")
```

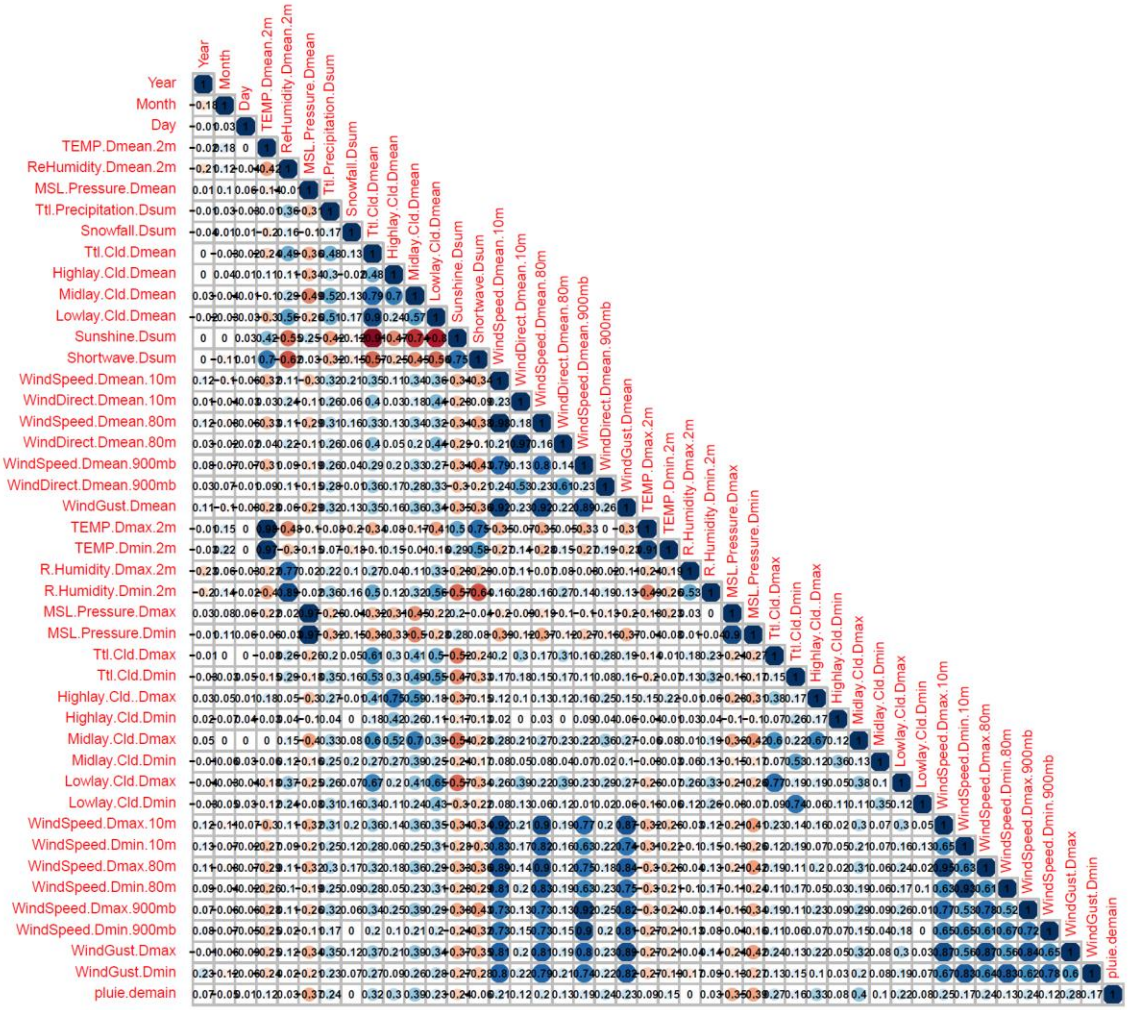
Importer les fichiers “meteo.train” et “meteo.test”

```
meteo.train.initial=read.csv("meteo.train.csv")
meteo.predict.initial=read.csv("meteo.test.csv")
```

Renommer les attributs avec les noms courts pour faciliter l’analyse

```
m=meteo.train.initial
m=m %>%
  rename(
    TEMP.Dmean.2m = Temperature.daily.mean..2.m.above.gnd.,
    ReHumidity.Dmean.2m = Relative.Humidity.daily.mean..2.m.above.gnd.,
    MSL.Pressure.Dmean = Mean.Sea.Level.Pressure.daily.mean..MSL.,
    Ttl.Precipitation.Dsum = Total.Precipitation.daily.sum..sfc.,
    Snowfall.Dsum = Snowfall.amount.raw.daily.sum..sfc.,
    Ttl.Cld.Dmean = Total.Cloud.Cover.daily.mean..sfc.,
    Highlay.Cld.Dmean = High.Cloud.Cover.daily.mean..high.cld.lay.,
    Midlay.Cld.Dmean = Medium.Cloud.Cover.daily.mean..mid.cld.lay.,
    Lowlay.Cld.Dmean = Low.Cloud.Cover.daily.mean..low.cld.lay.,
    Sunshine.Dsum = Sunshine.Duration.daily.sum..sfc.,
    Shortwave.Dsum = Shortwave.Radiation.daily.sum..sfc.,
    WindSpeed.Dmean.10m = Wind.Speed.daily.mean..10.m.above.gnd.,
    WindDirect.Dmean.10m = Wind.Direction.daily.mean..10.m.above.gnd.,
    WindSpeed.Dmean.80m = Wind.Speed.daily.mean..80.m.above.gnd.,
    WindDirect.Dmean.80m = Wind.Direction.daily.mean..80.m.above.gnd.,
    WindSpeed.Dmean.900mb = Wind.Speed.daily.mean..900.mb.,
    WindDirect.Dmean.900mb = Wind.Direction.daily.mean..900.mb.,
    WindGust.Dmean = Wind.Gust.daily.mean..sfc.,
    TEMP.Dmax.2m = Temperature.daily.max..2.m.above.gnd.,
    TEMP.Dmin.2m = Temperature.daily.min..2.m.above.gnd.,
    R.Humidity.Dmax.2m = Relative.Humidity.daily.max..2.m.above.gnd.,
    R.Humidity.Dmin.2m = Relative.Humidity.daily.min..2.m.above.gnd.,
    MSL.Pressure.Dmax = Mean.Sea.Level.Pressure.daily.max..MSL.,
    MSL.Pressure.Dmin = Mean.Sea.Level.Pressure.daily.min..MSL.,
    Ttl.Cld.Dmax = Total.Cloud.Cover.daily.max..sfc.,
    Ttl.Cld.Dmin = Total.Cloud.Cover.daily.min..sfc.,
    Highlay.Cld..Dmax = High.Cloud.Cover.daily.max..high.cld.lay.,
    Highlay.Cld.Dmin = High.Cloud.Cover.daily.min..high.cld.lay.,
    Midlay.Cld.Dmax = Medium.Cloud.Cover.daily.max..mid.cld.lay.,
    Midlay.Cld.Dmin = Medium.Cloud.Cover.daily.min..mid.cld.lay.,
    Lowlay.Cld.Dmax = Low.Cloud.Cover.daily.max..low.cld.lay.,
    Lowlay.Cld.Dmin = Low.Cloud.Cover.daily.min..low.cld.lay.,
    WindSpeed.Dmax.10m = Wind.Speed.daily.max..10.m.above.gnd.,
    WindSpeed.Dmin.10m = Wind.Speed.daily.min..10.m.above.gnd.,
    WindSpeed.Dmax.80m = Wind.Speed.daily.max..80.m.above.gnd.,
    WindSpeed.Dmin.80m = Wind.Speed.daily.min..80.m.above.gnd.,
    WindSpeed.Dmax.900mb = Wind.Speed.daily.max..900.mb.,
```

```
    WindSpeed.Dmin.900mb = Wind.Speed.daily.min..900.mb.,
    WindGust.Dmax = Wind.Gust.daily.max..sfc.,
    WindGust.Dmin = Wind.Gust.daily.min..sfc.
  )

cor.meteo=cor(m[,c(2:4,7:47)],use = 'complete')
corrplot(cor.meteo,type="lower",number.cex=0.4,tl.cex=0.5,addCoef.col = 'black')
```



Choix de modèle via la méthode ‘Pas à Pas’

Avec 43 variables présentes, il est très couteux au niveau de temp à sélectionner les covariables via la méthode exhaustive. Pour cela, j’utilise d’abord la fonction step() pour la choix des modèles prédictives de la manière suivante:

1. Modèle logit via ‘stepwise selection’ en basant sur la critère AIC
2. modèle probit via ‘stepwise selection’ en basant sur la critère AIC
3. modèle logit via ‘stepwise selection’ en basant sur la critère BIC
4. modèle logit via ‘backward selection’ en basant sur la critère AIC

5. modèle probit via 'backward selection' en basant sur la critère AIC
6. modèle logit via 'backward selection' en basant sur la critère BIC

Je m'intéresse à :

• comparer les modèles selectionnés par "stepwise selection" et "backward selection"

• comparer le modèle logit au modèle probit

• comparer les modèles séléctionnés basant sur la critère AIC aux ceux sélectionnés utilisant la critère BIC

Préparer la "full modèle logit" pour la 'step selection'

```
m.glm=m[,c(2:4,7:47)]
glm.logit=glm(pluie.demain~.,data=m.glm,family = binomial)
summary(glm.logit)
```

```
##
## Call:
## glm(formula = pluie.demain ~ ., family = binomial, data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7911  -0.8301   0.2850   0.8297   2.9293
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.646e+01  7.061e+01  -1.083 0.278866
## Year                      6.921e-02  3.486e-02   1.985 0.047125 *
## Month                    -1.860e-02  2.493e-02  -0.746 0.455776
## Day                       1.179e-02  8.160e-03   1.445 0.148438
## TEMP.Dmean.2m             1.830e-01  1.640e-01   1.116 0.264488
## ReHumidity.Dmean.2m       1.986e-02  3.243e-02   0.612 0.540327
## MSL.Pressure.Dmean        5.124e-01  1.394e-01   3.675 0.000237 ***
## Ttl.Precipitation.Dsum    2.586e-02  2.805e-02   0.922 0.356497
## Snowfall.Dsum            -2.853e-01  2.339e-01  -1.220 0.222560
## Ttl.Cld.Dmean             1.247e-02  1.200e-02   1.039 0.298720
## Highlay.Cld.Dmean        -3.253e-03  6.820e-03  -0.477 0.633421
## Midlay.Cld.Dmean          5.590e-03  6.691e-03   0.835 0.403515
## Lowlay.Cld.Dmean         -4.340e-03  8.114e-03  -0.535 0.592768
## Sunshine.Dsum             4.908e-04  8.828e-04   0.556 0.578242
## Shortwave.Dsum            2.938e-05  9.883e-05   0.297 0.766285
## WindSpeed.Dmean.10m      -4.640e-02  9.698e-02  -0.478 0.632361
## WindDirect.Dmean.10m      5.637e-03  5.768e-03   0.977 0.328385
## WindSpeed.Dmean.80m      -9.430e-02  6.947e-02  -1.357 0.174690
## WindDirect.Dmean.80m     -9.489e-03  5.960e-03  -1.592 0.111388
## WindSpeed.Dmean.900mb     1.834e-02  2.594e-02   0.707 0.479457
## WindDirect.Dmean.900mb    5.404e-03  1.451e-03   3.723 0.000197 ***
## WindGust.Dmean            1.782e-02  3.689e-02   0.483 0.629095
## TEMP.Dmax.2m             -1.146e-02  9.593e-02  -0.119 0.904920
## TEMP.Dmin.2m             -1.302e-01  8.631e-02  -1.509 0.131368
## R.Humidity.Dmax.2m        6.722e-05  2.061e-02   0.003 0.997398
## R.Humidity.Dmin.2m       -6.868e-03  1.856e-02  -0.370 0.711278
## MSL.Pressure.Dmax        -2.587e-01  7.502e-02  -3.449 0.000564 ***
## MSL.Pressure.Dmin        -3.206e-01  7.572e-02  -4.234 2.29e-05 ***
## Ttl.Cld.Dmax              3.412e-03  4.864e-03   0.701 0.483090
## Ttl.Cld.Dmin              7.789e-03  6.264e-03   1.243 0.213759
## Highlay.Cld..Dmax         3.423e-03  2.886e-03   1.186 0.235609
## Highlay.Cld.Dmin          6.148e-03  2.093e-02   0.294 0.768986
```

```
## Midlay.Cld.Dmax         6.159e-03  3.164e-03   1.946 0.051598 .
## Midlay.Cld.Dmin        -5.295e-03  9.463e-03  -0.560 0.575746
## Lowlay.Cld.Dmax         2.944e-03  3.397e-03   0.867 0.386126
## Lowlay.Cld.Dmin         1.197e-04  7.017e-03   0.017 0.986395
## WindSpeed.Dmax.10m      5.588e-02  3.448e-02   1.620 0.105153
## WindSpeed.Dmin.10m      1.690e-01  6.415e-02   2.635 0.008415 **
## WindSpeed.Dmax.80m      3.933e-03  2.845e-02   0.138 0.890059
## WindSpeed.Dmin.80m     -5.304e-02  4.219e-02  -1.257 0.208741
## WindSpeed.Dmax.900mb   -1.342e-02  1.213e-02  -1.106 0.268904
## WindSpeed.Dmin.900mb   -4.050e-03  1.911e-02  -0.212 0.832168
## WindGust.Dmax           2.282e-02  1.730e-02   1.319 0.187260
## WindGust.Dmin           5.101e-03  2.800e-02   0.182 0.855427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1232.7  on 1136  degrees of freedom
## AIC: 1320.7
##
## Number of Fisher Scoring iterations: 5
```

Préparer la “full modèle probit” pour la ‘step selection’

```
glm.probit=glm(pluie.demain~.,data=m.glm,family = binomial(link="probit"))
summary(glm.probit)

##
## Call:
## glm(formula = pluie.demain ~ ., family = binomial(link = "probit"),
##     data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8027  -0.8479   0.2559   0.8468   3.1124
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -4.476e+01  4.133e+01  -1.083 0.278856
## Year                    4.022e-02  2.041e-02   1.971 0.048779 *
## Month                  -1.247e-02  1.455e-02  -0.857 0.391363
## Day                     6.537e-03  4.778e-03   1.368 0.171299
## TEMP.Dmean.2m           1.140e-01  9.477e-02   1.203 0.228828
## ReHumidity.Dmean.2m     9.422e-03  1.891e-02   0.498 0.618223
## MSL.Pressure.Dmean      2.875e-01  7.939e-02   3.622 0.000292 ***
## Ttl.Precipitation.Dsum  1.506e-02  1.592e-02   0.946 0.344144
## Snowfall.Dsum          -1.703e-01  1.294e-01  -1.316 0.188215
## Ttl.Cld.Dmean           7.394e-03  7.039e-03   1.050 0.293541
## Highlay.Cld.Dmean      -1.619e-03  3.968e-03  -0.408 0.683202
## Midlay.Cld.Dmean        3.053e-03  3.942e-03   0.774 0.438712
## Lowlay.Cld.Dmean       -2.778e-03  4.770e-03  -0.582 0.560339
## Sunshine.Dsum           3.096e-04  5.192e-04   0.596 0.550919
## Shortwave.Dsum          1.699e-05  5.789e-05   0.294 0.769093
## WindSpeed.Dmean.10m    -3.052e-02  5.685e-02  -0.537 0.591383
## WindDirect.Dmean.10m    3.383e-03  3.373e-03   1.003 0.315880
## WindSpeed.Dmean.80m    -5.637e-02  4.072e-02  -1.384 0.166261
## WindDirect.Dmean.80m   -5.725e-03  3.483e-03  -1.644 0.100231
## WindSpeed.Dmean.900mb   1.229e-02  1.498e-02   0.821 0.411825
```

```
## WindDirect.Dmean.900mb  3.140e-03  8.480e-04   3.702 0.000214 ***
## WindGust.Dmean           1.234e-02  2.144e-02   0.575 0.564971
## TEMP.Dmax.2m            -1.490e-02  5.502e-02  -0.271 0.786507
## TEMP.Dmin.2m            -7.520e-02  5.038e-02  -1.493 0.135526
## R.Humidity.Dmax.2m       9.626e-04  1.205e-02   0.080 0.936330
## R.Humidity.Dmin.2m      -2.122e-03  1.075e-02  -0.197 0.843439
## MSL.Pressure.Dmax       -1.461e-01  4.310e-02  -3.389 0.000701 ***
## MSL.Pressure.Dmin       -1.801e-01  4.286e-02  -4.201 2.66e-05 ***
## Ttl.Cld.Dmax             1.633e-03  2.836e-03   0.576 0.564885
## Ttl.Cld.Dmin             4.055e-03  3.563e-03   1.138 0.255060
## Highlay.Cld..Dmax        2.058e-03  1.712e-03   1.202 0.229290
## Highlay.Cld.Dmin         5.632e-04  1.097e-02   0.051 0.959039
## Midlay.Cld.Dmax          3.859e-03  1.868e-03   2.066 0.038855 *
## Midlay.Cld.Dmin         -1.966e-03  5.359e-03  -0.367 0.713685
## Lowlay.Cld.Dmax          1.812e-03  2.025e-03   0.894 0.371065
## Lowlay.Cld.Dmin          5.102e-04  4.018e-03   0.127 0.898961
## WindSpeed.Dmax.10m       3.560e-02  2.023e-02   1.760 0.078432 .
## WindSpeed.Dmin.10m       9.546e-02  3.745e-02   2.549 0.010805 *
## WindSpeed.Dmax.80m       1.350e-03  1.667e-02   0.081 0.935459
## WindSpeed.Dmin.80m      -2.999e-02  2.484e-02  -1.207 0.227264
## WindSpeed.Dmax.900mb    -8.354e-03  7.030e-03  -1.188 0.234669
## WindSpeed.Dmin.900mb    -4.485e-03  1.109e-02  -0.404 0.685999
## WindGust.Dmax            1.336e-02  1.004e-02   1.330 0.183400
## WindGust.Dmin            4.138e-03  1.634e-02   0.253 0.800060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1236.1  on 1136  degrees of freedom
## AIC: 1324.1
##
## Number of Fisher Scoring iterations: 5
```

Stepwise Selection via step()

Stepwise Selection - 1. modèle logit en basant sur la critère AIC

```
# modèle logit via 'stepwise selection' en basant sur la critère AIC
glm.stepw=step(glm(pluie.demain~1,data=m.glm,family = binomial), pluie.demain ~ Year +
Month + Day + TEMP.Dmean.2m + ReHumidity.Dmean.2m +
                    MSL.Pressure.Dmean + Ttl.Precipitation.Dsum + Snowfall.Dsum +
                    Ttl.Cld.Dmean + Highlay.Cld.Dmean + Midlay.Cld.Dmean + Lowlay.Cld.D
mean +
                    Sunshine.Dsum + Shortwave.Dsum + WindSpeed.Dmean.10m + WindDirect.D
mean.10m +
                    WindSpeed.Dmean.80m + WindDirect.Dmean.80m + WindSpeed.Dmean.900mb
+
                    WindDirect.Dmean.900mb + WindGust.Dmean + TEMP.Dmax.2m +
                    TEMP.Dmin.2m + R.Humidity.Dmax.2m + R.Humidity.Dmin.2m +
                    MSL.Pressure.Dmax + MSL.Pressure.Dmin + Ttl.Cld.Dmax + Ttl.Cld.Dmin
 +
                    Highlay.Cld..Dmax + Highlay.Cld.Dmin + Midlay.Cld.Dmax +
                    Midlay.Cld.Dmin + Lowlay.Cld.Dmax + Lowlay.Cld.Dmin + WindSpeed.Dma
x.10m +
                    WindSpeed.Dmin.10m + WindSpeed.Dmax.80m + WindSpeed.Dmin.80m +
                    WindSpeed.Dmax.900mb + WindSpeed.Dmin.900mb + WindGust.Dmax +
                    WindGust.Dmin,data=m.glm,direction = "both")
```

```
summary(glm.stepw)

##
## Call:
## glm(formula = pluie.demain ~ Midlay.Cld.Dmax + MSL.Pressure.Dmin +
##     TEMP.Dmin.2m + WindGust.Dmax + Ttl.Cld.Dmean + TEMP.Dmax.2m +
##     WindDirect.Dmean.900mb + Year + WindSpeed.Dmean.80m + WindSpeed.Dmin.10m +
##     WindSpeed.Dmax.10m + WindDirect.Dmean.80m + Snowfall.Dsum +
##     Ttl.Cld.Dmin, family = binomial, data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4374  -0.8622   0.3226   0.8524   2.7697
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -61.600275  62.148039  -0.991 0.321595
## Midlay.Cld.Dmax           0.009756   0.002148   4.543 5.55e-06 ***
## MSL.Pressure.Dmin        -0.065489   0.010515  -6.228 4.72e-10 ***
## TEMP.Dmin.2m             -0.055325   0.034996  -1.581 0.113904
## WindGust.Dmax             0.026808   0.010666   2.513 0.011962 *
## Ttl.Cld.Dmean             0.011544   0.003938   2.931 0.003374 **
## TEMP.Dmax.2m              0.101845   0.030895   3.297 0.000979 ***
## WindDirect.Dmean.900mb    0.004732   0.001282   3.691 0.000224 ***
## Year                      0.061821   0.030490   2.028 0.042605 *
## WindSpeed.Dmean.80m      -0.116914   0.029693  -3.937 8.24e-05 ***
## WindSpeed.Dmin.10m        0.100079   0.035327   2.833 0.004613 **
## WindSpeed.Dmax.10m        0.059281   0.022576   2.626 0.008642 **
## WindDirect.Dmean.80m     -0.002899   0.001521  -1.907 0.056557 .
## Snowfall.Dsum            -0.329552   0.201944  -1.632 0.102702
## Ttl.Cld.Dmin              0.005994   0.003858   1.554 0.120262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1262.5  on 1165  degrees of freedom
## AIC: 1292.5
##
## Number of Fisher Scoring iterations: 4
```

Stepwise Selection - 2. modèle probit en basant sur la critère AIC

```
#modèle probit via  'stepwise selection' en basant sur la critère AIC
glm.stepw_probit=step(glm(pluie.demain~1,data=m.glm,family = binomial(link = "probit"))
, pluie.demain ~ Year + Month + Day + TEMP.Dmean.2m + ReHumidity.Dmean.2m +
                 MSL.Pressure.Dmean + Ttl.Precipitation.Dsum + Snowfall.Dsum +
                 Ttl.Cld.Dmean + Highlay.Cld.Dmean + Midlay.Cld.Dmean + Lowlay.Cld.Dmea
n +
                 Sunshine.Dsum + Shortwave.Dsum + WindSpeed.Dmean.10m + WindDirect.Dmea
n.10m +
                 WindSpeed.Dmean.80m + WindDirect.Dmean.80m + WindSpeed.Dmean.900mb +
                 WindDirect.Dmean.900mb + WindGust.Dmean + TEMP.Dmax.2m +
                 TEMP.Dmin.2m + R.Humidity.Dmax.2m + R.Humidity.Dmin.2m +
                 MSL.Pressure.Dmax + MSL.Pressure.Dmin + Ttl.Cld.Dmax + Ttl.Cld.Dmin +
                 Highlay.Cld..Dmax + Highlay.Cld.Dmin + Midlay.Cld.Dmax +
                 Midlay.Cld.Dmin + Lowlay.Cld.Dmax + Lowlay.Cld.Dmin + WindSpeed.Dmax.1
0m +
                 WindSpeed.Dmin.10m + WindSpeed.Dmax.80m + WindSpeed.Dmin.80m +
                 WindSpeed.Dmax.900mb + WindSpeed.Dmin.900mb + WindGust.Dmax +
                 WindGust.Dmin,data=m.glm,direction = "both")
```

```
summary(glm.stepw_probit)

##
## Call:
## glm(formula = pluie.demain ~ Midlay.Cld.Dmax + MSL.Pressure.Dmin +
##     WindDirect.Dmean.900mb + TEMP.Dmax.2m + WindGust.Dmax + Ttl.Cld.Dmean +
##     Year + WindDirect.Dmean.80m + WindSpeed.Dmean.80m + WindSpeed.Dmin.10m +
##     WindSpeed.Dmax.10m + Snowfall.Dsum + Ttl.Cld.Dmin + MSL.Pressure.Dmean +
##     MSL.Pressure.Dmax, family = binomial(link = "probit"), data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5808  -0.8609   0.2600   0.8748   2.9571
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -3.809e+01  3.670e+01  -1.038 0.299405
## Midlay.Cld.Dmax         5.969e-03  1.275e-03   4.681 2.85e-06 ***
## MSL.Pressure.Dmin      -1.622e-01  4.040e-02  -4.016 5.93e-05 ***
## WindDirect.Dmean.900mb  2.555e-03  7.421e-04   3.443 0.000575 ***
## TEMP.Dmax.2m            3.168e-02  6.821e-03   4.644 3.42e-06 ***
## WindGust.Dmax           1.457e-02  6.267e-03   2.325 0.020088 *
## Ttl.Cld.Dmean           5.833e-03  2.193e-03   2.660 0.007816 **
## Year                    3.855e-02  1.804e-02   2.137 0.032610 *
## WindDirect.Dmean.80m   -2.104e-03  8.728e-04  -2.411 0.015915 *
## WindSpeed.Dmean.80m    -6.715e-02  1.742e-02  -3.855 0.000116 ***
## WindSpeed.Dmin.10m      5.672e-02  2.078e-02   2.730 0.006343 **
## WindSpeed.Dmax.10m      3.668e-02  1.318e-02   2.784 0.005376 **
## Snowfall.Dsum          -1.852e-01  1.200e-01  -1.543 0.122754
## Ttl.Cld.Dmin            3.805e-03  2.233e-03   1.704 0.088362 .
## MSL.Pressure.Dmean      2.480e-01  7.453e-02   3.327 0.000878 ***
## MSL.Pressure.Dmax      -1.264e-01  4.059e-02  -3.114 0.001844 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1255.4  on 1164  degrees of freedom
## AIC: 1287.4
##
## Number of Fisher Scoring iterations: 5
```

Stepwise Selection - 3. modèle logit en basant sur la critère BIC

```
# modèle logit via 'stepwise selection' en basant sur la critère BIC
glm.stepw_BIC<- stepAIC(glm(pluie.demain~1,data=m.glm,family = binomial), pluie.demain
~ Year + Month + Day + TEMP.Dmean.2m + ReHumidity.Dmean.2m +
                          MSL.Pressure.Dmean + Ttl.Precipitation.Dsum + Snowfall.Dsum +

                          Ttl.Cld.Dmean + Highlay.Cld.Dmean + Midlay.Cld.Dmean + Lowlay
.Cld.Dmean +
                          Sunshine.Dsum + Shortwave.Dsum + WindSpeed.Dmean.10m + WindDi
rect.Dmean.10m +
                          WindSpeed.Dmean.80m + WindDirect.Dmean.80m + WindSpeed.Dmean.
900mb +
                          WindDirect.Dmean.900mb + WindGust.Dmean + TEMP.Dmax.2m +
                          TEMP.Dmin.2m + R.Humidity.Dmax.2m + R.Humidity.Dmin.2m +
                          MSL.Pressure.Dmax + MSL.Pressure.Dmin + Ttl.Cld.Dmax + Ttl.Cl
d.Dmin +
                          Highlay.Cld..Dmax + Highlay.Cld.Dmin + Midlay.Cld.Dmax +
                          Midlay.Cld.Dmin + Lowlay.Cld.Dmax + Lowlay.Cld.Dmin + WindSpe
```

```
ed.Dmax.10m +
                          WindSpeed.Dmin.10m + WindSpeed.Dmax.80m + WindSpeed.Dmin.80m
+
                          WindSpeed.Dmax.900mb + WindSpeed.Dmin.900mb + WindGust.Dmax +

                          WindGust.Dmin,data=m.glm,direction = "both",k = log(nrow(m)))

summary(glm.stepw_BIC)

##
## Call:
## glm(formula = pluie.demain ~ Midlay.Cld.Dmax + MSL.Pressure.Dmin +
##     WindGust.Dmax + Ttl.Cld.Dmean + TEMP.Dmax.2m, family = binomial,
##     data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2825  -0.8975   0.3940   0.8631   2.6940
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       59.236757  10.376489   5.709 1.14e-08 ***
## Midlay.Cld.Dmax    0.010884   0.002060   5.284 1.27e-07 ***
## MSL.Pressure.Dmin -0.061488   0.010128  -6.071 1.27e-09 ***
## WindGust.Dmax      0.022965   0.005630   4.079 4.52e-05 ***
## Ttl.Cld.Dmean      0.012263   0.002987   4.105 4.04e-05 ***
## TEMP.Dmax.2m       0.064181   0.010523   6.099 1.07e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1300.9  on 1174  degrees of freedom
## AIC: 1312.9
##
## Number of Fisher Scoring iterations: 4
```

Backward Selection via step()

Backward Selection - 4. modèle logit en basant sur la critère AIC

```
#modèle logit via 'backward selection' en basant sur la critère AIC
glm.backw=step(glm.logit,direction = "backward")

summary(glm.backw)

##
## Call:
## glm(formula = pluie.demain ~ Year + TEMP.Dmean.2m + MSL.Pressure.Dmean +
##     Snowfall.Dsum + Midlay.Cld.Dmean + WindSpeed.Dmean.80m +
##     WindDirect.Dmean.80m + WindDirect.Dmean.900mb + TEMP.Dmin.2m +
##     MSL.Pressure.Dmax + MSL.Pressure.Dmin + Ttl.Cld.Dmax + Ttl.Cld.Dmin +
##     Midlay.Cld.Dmax + WindSpeed.Dmax.10m + WindSpeed.Dmin.10m +
##     WindGust.Dmax, family = binomial, data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5754  -0.8336   0.2694   0.8517   2.8827
```

```
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -68.942905  62.791813  -1.098 0.272222
## Year                      0.065852   0.030853   2.134 0.032811 *
## TEMP.Dmean.2m             0.146739   0.049863   2.943 0.003252 **
## MSL.Pressure.Dmean        0.481900   0.131205   3.673 0.000240 ***
## Snowfall.Dsum            -0.316432   0.215283  -1.470 0.141605
## Midlay.Cld.Dmean          0.010810   0.004061   2.662 0.007776 **
## WindSpeed.Dmean.80m      -0.114538   0.029949  -3.824 0.000131 ***
## WindDirect.Dmean.80m     -0.002686   0.001524  -1.763 0.077965 .
## WindDirect.Dmean.900mb    0.004585   0.001289   3.557 0.000375 ***
## TEMP.Dmin.2m             -0.102830   0.054202  -1.897 0.057806 .
## MSL.Pressure.Dmax        -0.242121   0.070640  -3.428 0.000609 ***
## MSL.Pressure.Dmin        -0.306000   0.071569  -4.276 1.91e-05 ***
## Ttl.Cld.Dmax              0.008353   0.003504   2.383 0.017151 *
## Ttl.Cld.Dmin              0.007844   0.003863   2.031 0.042272 *
## Midlay.Cld.Dmax           0.006226   0.002671   2.331 0.019757 *
## WindSpeed.Dmax.10m        0.059655   0.022770   2.620 0.008795 **
## WindSpeed.Dmin.10m        0.111262   0.036110   3.081 0.002062 **
## WindGust.Dmax             0.023669   0.010856   2.180 0.029238 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1246.8  on 1162  degrees of freedom
## AIC: 1282.8
##
## Number of Fisher Scoring iterations: 4
```

Backward Selection - 5. modèle probit en basant sur la critère AIC

```
#modèle probit via  'backward selection' en basant sur la critère AIC
glm.backw_probit=step(glm.probit,direction = "backward")

summary(glm.backw_probit)

##
## Call:
## glm(formula = pluie.demain ~ Year + TEMP.Dmean.2m + MSL.Pressure.Dmean +
##     Snowfall.Dsum + Ttl.Cld.Dmean + WindSpeed.Dmean.80m + WindDirect.Dmean.80m +
##     WindDirect.Dmean.900mb + TEMP.Dmin.2m + MSL.Pressure.Dmax +
##     MSL.Pressure.Dmin + Ttl.Cld.Dmin + Highlay.Cld..Dmax + Midlay.Cld.Dmax +
##     Lowlay.Cld.Dmax + WindSpeed.Dmax.10m + WindSpeed.Dmin.10m +
##     WindGust.Dmax, family = binomial(link = "probit"), data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5704  -0.8471   0.2506   0.8649   3.1323
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -4.180e+01  3.693e+01  -1.132 0.257632
## Year                    4.003e-02  1.813e-02   2.208 0.027236 *
## TEMP.Dmean.2m           9.343e-02  3.177e-02   2.941 0.003269 **
## MSL.Pressure.Dmean      2.592e-01  7.442e-02   3.483 0.000496 ***
## Snowfall.Dsum          -1.703e-01  1.193e-01  -1.428 0.153392
```

```
## Ttl.Cld.Dmean           4.490e-03  2.639e-03   1.701 0.088899 .
## WindSpeed.Dmean.80m    -6.571e-02  1.756e-02  -3.742 0.000183 ***
## WindDirect.Dmean.80m   -2.079e-03  9.050e-04  -2.297 0.021597 *
## WindDirect.Dmean.900mb  2.663e-03  7.544e-04   3.530 0.000416 ***
## TEMP.Dmin.2m           -6.816e-02  3.369e-02  -2.023 0.043038 *
## MSL.Pressure.Dmax      -1.343e-01  4.065e-02  -3.305 0.000950 ***
## MSL.Pressure.Dmin      -1.651e-01  4.037e-02  -4.089 4.32e-05 ***
## Ttl.Cld.Dmin            4.504e-03  2.306e-03   1.953 0.050825 .
## Highlay.Cld..Dmax       2.010e-03  1.294e-03   1.553 0.120337
## Midlay.Cld.Dmax         4.840e-03  1.495e-03   3.237 0.001207 **
## Lowlay.Cld.Dmax         2.424e-03  1.553e-03   1.561 0.118526
## WindSpeed.Dmax.10m      3.495e-02  1.332e-02   2.623 0.008709 **
## WindSpeed.Dmin.10m      6.098e-02  2.099e-02   2.906 0.003665 **
## WindGust.Dmax           1.387e-02  6.308e-03   2.200 0.027836 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1248.7  on 1161  degrees of freedom
## AIC: 1286.7
##
## Number of Fisher Scoring iterations: 5
```

Backward Selection - 6. modèle logit en basant sur la critère BIC

```
#modèle logit via 'backward selection' en basant sur la critère BIC
glm.backw_BIC=step(glm.logit,direction = "backward",k = log(nrow(m)))

summary(glm.backw_BIC)
```

```
##
## Call:
## glm(formula = pluie.demain ~ TEMP.Dmean.2m + MSL.Pressure.Dmean +
##     Midlay.Cld.Dmean + WindSpeed.Dmean.80m + WindDirect.Dmean.900mb +
##     MSL.Pressure.Dmax + MSL.Pressure.Dmin + Midlay.Cld.Dmax +
##     WindSpeed.Dmax.10m + WindSpeed.Dmin.10m, family = binomial,
##     data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4000  -0.8432   0.2488   0.8668   2.6718
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            65.2410474 11.6246630   5.612 2.00e-08 ***
## TEMP.Dmean.2m           0.0524903  0.0116419   4.509 6.52e-06 ***
## MSL.Pressure.Dmean      0.5254040  0.1328791   3.954 7.69e-05 ***
## Midlay.Cld.Dmean        0.0124504  0.0035288   3.528 0.000418 ***
## WindSpeed.Dmean.80m    -0.0978213  0.0283301  -3.453 0.000555 ***
## WindDirect.Dmean.900mb  0.0029399  0.0009984   2.945 0.003232 **
## MSL.Pressure.Dmax      -0.2620604  0.0701288  -3.737 0.000186 ***
## MSL.Pressure.Dmin      -0.3303315  0.0729515  -4.528 5.95e-06 ***
## Midlay.Cld.Dmax         0.0088107  0.0023339   3.775 0.000160 ***
## WindSpeed.Dmax.10m      0.0782404  0.0193211   4.049 5.13e-05 ***
## WindSpeed.Dmin.10m      0.1020170  0.0347159   2.939 0.003297 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1272.3  on 1169  degrees of freedom
## AIC: 1294.3
##
## Number of Fisher Scoring iterations: 4
```

Comparaison de 6 modèles sélectionnés par step()

Je vérifie par les codes ci-dessous :

- AIC de chaque modèle sélectionné par step()
- Modèle Utile ? : la différence entre 'Null Deviance' et 'Residual Deviance' par rapport le nombre des variables sélectionnés
- Modèle suffisant ? : P-value du test khi-deux sur le rapport de vraisemblance entre le modèle sélectionné(Mk) et celui saturé(Msat)

```
# vérifier AIC des 6 modèles selectionés par la méthode de 'pas à pas'
# vérifier Null Deviance et Residual Deviance
print("Modèle 1 logit basant sur AIC 'stepwise selection':")
sprintf(" 'AIC' %f, 'P-value Test Khi2' %f, 'nb variable' %d",AIC(glm.stepw), pchisq(gl
m.stepw$deviance,glm.stepw$df.residual,lower=F),I(glm.stepw$df.null-glm.stepw$df.residu
al))

print("Modèle 2 probit basant sur AIC 'stepwise selection':")
sprintf(" 'AIC' %f, 'P-value Test Khi2' %f, 'nb variable' %d",AIC(glm.stepw_probit), pc
hisq(glm.stepw_probit$deviance,glm.stepw_probit$df.residual,lower=F),I(glm.stepw_probit
$df.null-glm.stepw$df.residual))

print("Modèle 3 logit basant sur BIC 'stepwise selection':")
sprintf(" 'AIC' %f, 'P-value Test Khi2' %f, 'nb variable' %d",AIC(glm.stepw_BIC), pchis
q(glm.stepw_BIC$deviance,glm.stepw_BIC$df.residual,lower=F),I(glm.stepw_BIC$df.null-glm
.stepw_BIC$df.residual))

print("Modèle 4 logit basant sur AIC 'backward selection':")
sprintf(" 'AIC' %f, 'P-value Test Khi2' %f, 'nb variable' %d",AIC(glm.backw), pchisq(gl
m.backw$deviance,glm.backw$df.residual,lower=F),I(glm.backw$df.null-glm.backw$df.residu
al))

print("Modèle 5 probit basant sur AIC 'backward selection':")
sprintf(" 'AIC' %f, 'P-value Test Khi2' %f, 'nb variable' %d",AIC(glm.backw_probit), pc
hisq(glm.backw_probit$deviance,glm.backw_probit$df.residual,lower=F),I(glm.backw_probit
$df.null-glm.backw_probit$df.residual))

print("Modèle 6 logit basant sur BIC 'backward selection':")
sprintf(" 'AIC' %f, 'P-value Test Khi2' %f, 'nb variable' %d",AIC(glm.backw_BIC), pchis
q(glm.backw_BIC$deviance,glm.backw_BIC$df.residual,lower=F),I(glm.backw_BIC$df.null-glm
.backw_BIC$df.residual))

## [1] "Modèle 1 logit basant sur AIC 'stepwise selection':"
## [1] " 'AIC' 1292.494551, 'P-value Test Khi2' 0.023866, 'nb variable' 14"
## [1] "Modèle 2 probit basant sur AIC 'stepwise selection':"
```

```
## [1] " 'AIC' 1287.370763, 'P-value Test Khi2' 0.031439, 'nb variable' 14"
## [1] "Modèle 3 logit basant sur BIC 'stepwise selection':"
## [1] " 'AIC' 1312.926794, 'P-value Test Khi2' 0.005477, 'nb variable' 5"
## [1] "Modèle 4 logit basant sur AIC 'backward selection':"
## [1] " 'AIC' 1282.847288, 'P-value Test Khi2' 0.041577, 'nb variable' 17"
## [1] "Modèle 5 probit basant sur AIC 'backward selection':"
## [1] " 'AIC' 1286.700766, 'P-value Test Khi2' 0.036739, 'nb variable' 18"
## [1] "Modèle 6 logit basant sur BIC 'backward selection':"
## [1] " 'AIC' 1294.337766, 'P-value Test Khi2' 0.018267, 'nb variable' 10"
```

Selection Stepwise VS Selection Backward (step()):

Les 3 modèles sélectionnés par le mode 'backward' paraissent mieux que ceux sélectionnés par le mode 'stepwise'. Parce que leur valeur AIC est plus petit . Et leur p-valeur du test khi-deux est plus grande, ce qui indique le résidu des modèles 'backward sélectionnés' est plus petit. Néanmoins, la 'backward' selection' inclure plus de covariables que 'stepwise selection'.

Modèle Logit VS Modèle Probit(step()):

Par la sélection stepwise, le modèle probit parait mieux que celui de logit. Alors que c'est l'inverse dans le mode de selection 'backward'. Les deux types sont proches, en comparant leurs valeurs AIC et les p-valeur du test khi-deux.

Modèle basant sur AIC VS Modèle basant sur BIC (step()):

Les 4 modèles sélectionnés en basant sur la critère AIC sont mieux que ceux en basant sur la critère BIC, d'un point de vu de valeur AIC et de test khi-deux. Mais AIC a choisi le grand modèle.

Quelles modèles à choisir, et que faire pour pour la suite ?

Tous ces six modèles sont utiles, car la différence entre 'Null Deviance' et 'Residual Deviance' est assez grande par rapport au nombre des covariables sélectionnées. Néanmoins, aucun n'est un modèle suffisant, car leur p-valeur du test khi-deux des six modèles est tous inférieur à 5%. Il devrait manquer les covariables à inclure dans le modèle.

J'observe les points suivants, en comparant les deux modèles logit selectionnés par le mode 'backward' :

- Le modèle basant sur AIC(glm.backw) a 7 covariables en plus que celui basant sur BIC(glm.backw_BIC), et a néanmoins gagné seulement 12 sur la valeur AIC du modèle.
- Le modèle basant sur AIC a sélectionné déja 17 covariables parmi 43 variables en total.
- Certaines covariables sélectionnées semblent poser le problème de colinéarité (eg: le coefficient de MSL.Pressure.Dmean est biaisé par rapport aux MSL.Pressure.Dmax et MSL.Pressure.Dmin)

D'après ces observations, je me pose la question si le modèle basant sur AIC a sélectionné trop des variables par rapport au modèle basant sur BIC. Je suppose également qu'il manque l'impact de l'interaction entre les covariables dans le modèle sélectionné.

- tester la réduction des covariables 'backward selectionnées' dans le modéle basant sur AIC.
- tester l'ajout de l'interaction entre les covariables dans le modèle via 'glmutil()'

Préparation des modèles pour la validation croisée

Modèles avec la réduction des covariables

La différence des covariables présente dans les deux modèles basant sur BIC et AIC :

1. **Year(AIC, ! IN BIC)**
2. Snowfall.Dsum (AIC, ! BIC, P-value>0.5)
3. WindDirect.Dmean.80m (AIC, ! BIC, P-value>0.1)
4. TEMP.Dmin.2m((AIC, ! BIC, P-value>0.1))
5. **Ttl.Cld.Dmax (AIC, ! BIC)**
6. **Ttl.Cld.Dmin(AIC, ! BIC)**
7. **WindGust.Dmax(AIC, ! BIC)**

```
formula(glm.backw)
formula(glm.backw_BIC)
summary(glm.backw)
```

Parmis ces 7 covariables, je supprime ceux dont le coefficient est non-significative (p-value > 0.5).

J'obtiens le modèle **glm2_L1,** en supprimant 5 variables dans le modèle logit backward sélectionné basant sur AIC (glm.backw):

1. Snowfall.Dsum
2. WindDirect.Dmean.80m
3. TEMP.Dmin.2m
4. Ttl.Cld.Dmax
5. Ttl.Cld.Dmin

```
glm1_L1=glm(pluie.demain ~1 + TEMP.Dmean.2m + MSL.Pressure.Dmean + Midlay.Cld.Dmean +
              WindSpeed.Dmean.80m + WindDirect.Dmean.900mb + MSL.Pressure.Dmax +
              MSL.Pressure.Dmin + Midlay.Cld.Dmax + WindSpeed.Dmax.10m +
              WindSpeed.Dmin.10m+ WindGust.Dmax + Year+Ttl.Cld.Dmax+Ttl.Cld.Dmin,data=m.
```

```
glm,family=binomial )

summary(glm1_L1)

##
## Call:
## glm(formula = pluie.demain ~ 1 + TEMP.Dmean.2m + MSL.Pressure.Dmean +
##     Midlay.Cld.Dmean + WindSpeed.Dmean.80m + WindDirect.Dmean.900mb +
##     MSL.Pressure.Dmax + MSL.Pressure.Dmin + Midlay.Cld.Dmax +
##     WindSpeed.Dmax.10m + WindSpeed.Dmin.10m + WindGust.Dmax +
##     Year + Ttl.Cld.Dmax + Ttl.Cld.Dmin, family = binomial, data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4709  -0.8515   0.2813   0.8606   2.7422
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -76.342433  62.166020  -1.228 0.219432
## TEMP.Dmean.2m            0.056072   0.011995   4.674 2.95e-06 ***
## MSL.Pressure.Dmean       0.490713   0.132602   3.701 0.000215 ***
## Midlay.Cld.Dmean         0.009826   0.003959   2.482 0.013066 *
## WindSpeed.Dmean.80m     -0.106257   0.029443  -3.609 0.000308 ***
## WindDirect.Dmean.900mb   0.002561   0.001013   2.530 0.011415 *
## MSL.Pressure.Dmax       -0.246154   0.070136  -3.510 0.000449 ***
## MSL.Pressure.Dmin       -0.311508   0.072903  -4.273 1.93e-05 ***
## Midlay.Cld.Dmax          0.006960   0.002651   2.625 0.008661 **
## WindSpeed.Dmax.10m       0.047286   0.022016   2.148 0.031730 *
## WindSpeed.Dmin.10m       0.104603   0.035646   2.934 0.003341 **
## WindGust.Dmax            0.026219   0.010686   2.454 0.014142 *
## Year                     0.070088   0.030576   2.292 0.021889 *
## Ttl.Cld.Dmax             0.006068   0.003393   1.789 0.073688 .
## Ttl.Cld.Dmin             0.004830   0.003706   1.303 0.192564
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1258.9  on 1165  degrees of freedom
## AIC: 1288.9
##
## Number of Fisher Scoring iterations: 4

# supprimer Ttl.Cld.Dmax, Ttl.Cld.Dmin, car p-value>0.05.
glm1_L1=update(glm1_L1,~.-Ttl.Cld.Dmax -Ttl.Cld.Dmin)
summary(glm1_L1)

##
## Call:
## glm(formula = pluie.demain ~ TEMP.Dmean.2m + MSL.Pressure.Dmean +
##     Midlay.Cld.Dmean + WindSpeed.Dmean.80m + WindDirect.Dmean.900mb +
##     MSL.Pressure.Dmax + MSL.Pressure.Dmin + Midlay.Cld.Dmax +
##     WindSpeed.Dmax.10m + WindSpeed.Dmin.10m + WindGust.Dmax +
##     Year, family = binomial, data = m.glm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4610  -0.8544   0.2654   0.8609   2.5833
##
## Coefficients:
```

```
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -63.371198  61.583613  -1.029 0.303467
## TEMP.Dmean.2m            0.050851   0.011700   4.346 1.38e-05 ***
## MSL.Pressure.Dmean       0.485885   0.132634   3.663 0.000249 ***
## Midlay.Cld.Dmean         0.011764   0.003551   3.312 0.000925 ***
## WindSpeed.Dmean.80m     -0.113120   0.029246  -3.868 0.000110 ***
## WindDirect.Dmean.900mb   0.002713   0.001004   2.703 0.006877 **
## MSL.Pressure.Dmax       -0.248958   0.070148  -3.549 0.000387 ***
## MSL.Pressure.Dmin       -0.303810   0.072782  -4.174 2.99e-05 ***
## Midlay.Cld.Dmax          0.008712   0.002344   3.717 0.000201 ***
## WindSpeed.Dmax.10m       0.052706   0.021890   2.408 0.016052 *
## WindSpeed.Dmin.10m       0.111510   0.035232   3.165 0.001551 **
## WindGust.Dmax            0.025526   0.010664   2.394 0.016675 *
## Year                     0.063840   0.030284   2.108 0.035026 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1264.0  on 1167  degrees of freedom
## AIC: 1290
##
## Number of Fisher Scoring iterations: 4
```

D'après le test ANOVA suivant, la suppression des 5 covariables a fait perdre les informations. Le sous-modèle **glm1_L1** est moins bien que **glm.backw** initialement selectionné. .

Je vais revérifier ce point dans le chapitre <validation croisée> plus tard.

```
anova(glm1_L1,glm.backw,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: pluie.demain ~ TEMP.Dmean.2m + MSL.Pressure.Dmean + Midlay.Cld.Dmean +
##     WindSpeed.Dmean.80m + WindDirect.Dmean.900mb + MSL.Pressure.Dmax +
##     MSL.Pressure.Dmin + Midlay.Cld.Dmax + WindSpeed.Dmax.10m +
##     WindSpeed.Dmin.10m + WindGust.Dmax + Year
## Model 2: pluie.demain ~ Year + TEMP.Dmean.2m + MSL.Pressure.Dmean + Snowfall.Dsum +
##     Midlay.Cld.Dmean + WindSpeed.Dmean.80m + WindDirect.Dmean.80m +
##     WindDirect.Dmean.900mb + TEMP.Dmin.2m + MSL.Pressure.Dmax +
##     MSL.Pressure.Dmin + Ttl.Cld.Dmax + Ttl.Cld.Dmin + Midlay.Cld.Dmax +
##     WindSpeed.Dmax.10m + WindSpeed.Dmin.10m + WindGust.Dmax
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1167     1264.0
## 2      1162     1246.8  5   17.166 0.004196 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modèles avec l'interaction entre les variables

En utilisant glmulti() sur la modèle de base **glm1_L1**, j'obtiens le modèle **glm1_L2**, en prenant en compte de l'intéraction entre les variables. Les codes sont indiqués ci-dessous.

Je ne peux pas produire un telle modèle à partir de **glm.backw** un utilisant glmulti(), car ce modèle backward sélectionné en basant sur AIC contient trop des covariables.

```
#formula(glm1_L1)
glmulti.glm_L2 <- glmulti(pluie.demain ~ TEMP.Dmean.2m + MSL.Pressure.Dmean + Midlay.Cl
d.Dmean + WindSpeed.Dmean.80m + WindDirect.Dmean.900mb + MSL.Pressure.Dmax + MSL.Pressu
re.Dmin + Midlay.Cld.Dmax + WindSpeed.Dmax.10m + WindSpeed.Dmin.10m + WindGust.Dmax + Y
ear, data = m.glm,
                          level = 2,          #test interaction between covariables
                          method = "g",
                          crit = "aic",       # AIC as criteria
                          confsetsize = 3,
                          plotty = F, report = F,
                          fitfunction = "glm",
                          family = binomial)

glmulti.glm_L2@formulas[[1]]
summary(glmulti.glm_L2@objects[[1]])

glm1_L2=glm(formula = pluie.demain ~ 1 + TEMP.Dmean.2m + WindSpeed.Dmin.10m + MSL.Press
ure.Dmean:TEMP.Dmean.2m +
    WindDirect.Dmean.900mb:TEMP.Dmean.2m + WindDirect.Dmean.900mb:MSL.Pressure.Dmean +
    MSL.Pressure.Dmax:MSL.Pressure.Dmean + MSL.Pressure.Dmax:Midlay.Cld.Dmean +
    MSL.Pressure.Dmax:WindSpeed.Dmean.80m + MSL.Pressure.Dmin:MSL.Pressure.Dmean +
    MSL.Pressure.Dmin:WindDirect.Dmean.900mb + Midlay.Cld.Dmax:TEMP.Dmean.2m +
    Midlay.Cld.Dmax:Midlay.Cld.Dmean + Midlay.Cld.Dmax:WindSpeed.Dmean.80m +
    WindSpeed.Dmax.10m:WindDirect.Dmean.900mb + WindSpeed.Dmax.10m:MSL.Pressure.Dmax +
    WindSpeed.Dmax.10m:MSL.Pressure.Dmin + WindSpeed.Dmax.10m:Midlay.Cld.Dmax +
    WindSpeed.Dmin.10m:TEMP.Dmean.2m + WindSpeed.Dmin.10m:MSL.Pressure.Dmean +
    WindSpeed.Dmin.10m:Midlay.Cld.Dmean + WindGust.Dmax:TEMP.Dmean.2m +
    WindGust.Dmax:MSL.Pressure.Dmean + WindGust.Dmax:WindDirect.Dmean.900mb +
    WindGust.Dmax:MSL.Pressure.Dmin + WindGust.Dmax:Midlay.Cld.Dmax +
    Year:TEMP.Dmean.2m + Year:MSL.Pressure.Dmean + Year:MSL.Pressure.Dmax +
    Year:WindSpeed.Dmin.10m, family = "binomial", data = m.glm)

#step(glm1_L2)
summary(glm1_L2)

##
## Call:
## glm(formula = pluie.demain ~ 1 + TEMP.Dmean.2m + WindSpeed.Dmin.10m +
##     MSL.Pressure.Dmean:TEMP.Dmean.2m + WindDirect.Dmean.900mb:TEMP.Dmean.2m +
##     WindDirect.Dmean.900mb:MSL.Pressure.Dmean + MSL.Pressure.Dmax:MSL.Pressure.Dmean
 +
##     MSL.Pressure.Dmax:Midlay.Cld.Dmean + MSL.Pressure.Dmax:WindSpeed.Dmean.80m +
##     MSL.Pressure.Dmin:MSL.Pressure.Dmean + MSL.Pressure.Dmin:WindDirect.Dmean.900mb
+
##     Midlay.Cld.Dmax:TEMP.Dmean.2m + Midlay.Cld.Dmax:Midlay.Cld.Dmean +
##     Midlay.Cld.Dmax:WindSpeed.Dmean.80m + WindSpeed.Dmax.10m:WindDirect.Dmean.900mb
+
##     WindSpeed.Dmax.10m:MSL.Pressure.Dmax + WindSpeed.Dmax.10m:MSL.Pressure.Dmin +
##     WindSpeed.Dmax.10m:Midlay.Cld.Dmax + WindSpeed.Dmin.10m:TEMP.Dmean.2m +
##     WindSpeed.Dmin.10m:MSL.Pressure.Dmean + WindSpeed.Dmin.10m:Midlay.Cld.Dmean +
##     WindGust.Dmax:TEMP.Dmean.2m + WindGust.Dmax:MSL.Pressure.Dmean +
##     WindGust.Dmax:WindDirect.Dmean.900mb + WindGust.Dmax:MSL.Pressure.Dmin +
##     WindGust.Dmax:Midlay.Cld.Dmax + Year:TEMP.Dmean.2m + Year:MSL.Pressure.Dmean +
##     Year:MSL.Pressure.Dmax + Year:WindSpeed.Dmin.10m, family = "binomial",
##     data = m.glm)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.6178  -0.7764   0.1968   0.7788   2.7719
##
## Coefficients:
##                                           Estimate Std. Error z value
## (Intercept)                             -2.506e+02  8.708e+01  -2.878
## TEMP.Dmean.2m                            2.664e+01  1.011e+01   2.637
## WindSpeed.Dmin.10m                       4.655e+01  2.044e+01   2.277
## TEMP.Dmean.2m:MSL.Pressure.Dmean        -1.145e-02  1.952e-03  -5.863
## TEMP.Dmean.2m:WindDirect.Dmean.900mb    -5.915e-04  1.735e-04  -3.408
## MSL.Pressure.Dmean:WindDirect.Dmean.900mb -1.683e-03  7.164e-04  -2.349
## MSL.Pressure.Dmean:MSL.Pressure.Dmax     8.778e-04  2.059e-04   4.262
## MSL.Pressure.Dmax:Midlay.Cld.Dmean       1.690e-04  7.924e-05   2.133
## MSL.Pressure.Dmax:WindSpeed.Dmean.80m   -1.904e-04  6.279e-05  -3.032
## MSL.Pressure.Dmean:MSL.Pressure.Dmin    -1.082e-03  1.911e-04  -5.661
## WindDirect.Dmean.900mb:MSL.Pressure.Dmin  1.694e-03  7.182e-04   2.358
## TEMP.Dmean.2m:Midlay.Cld.Dmax            7.355e-04  2.144e-04   3.431
## Midlay.Cld.Dmean:Midlay.Cld.Dmax        -1.518e-03  7.968e-04  -1.905
## WindSpeed.Dmean.80m:Midlay.Cld.Dmax      1.113e-03  6.747e-04   1.649
## WindDirect.Dmean.900mb:WindSpeed.Dmax.10m  6.093e-04  2.533e-04   2.406
## MSL.Pressure.Dmax:WindSpeed.Dmax.10m     7.706e-03  3.507e-03   2.198
## MSL.Pressure.Dmin:WindSpeed.Dmax.10m    -7.656e-03  3.524e-03  -2.173
## Midlay.Cld.Dmax:WindSpeed.Dmax.10m      -2.095e-03  6.462e-04  -3.242
## TEMP.Dmean.2m:WindSpeed.Dmin.10m        -6.321e-03  4.089e-03  -1.546
## WindSpeed.Dmin.10m:MSL.Pressure.Dmean   -6.141e-03  3.237e-03  -1.897
## WindSpeed.Dmin.10m:Midlay.Cld.Dmean     -1.792e-03  8.515e-04  -2.105
## TEMP.Dmean.2m:WindGust.Dmax             -3.424e-03  1.081e-03  -3.166
## MSL.Pressure.Dmean:WindGust.Dmax        -1.534e-02  4.566e-03  -3.360
## WindDirect.Dmean.900mb:WindGust.Dmax    -2.749e-04  1.809e-04  -1.519
## MSL.Pressure.Dmin:WindGust.Dmax          1.544e-02  4.579e-03   3.371
## Midlay.Cld.Dmax:WindGust.Dmax            8.337e-04  3.515e-04   2.371
## TEMP.Dmean.2m:Year                      -7.337e-03  4.889e-03  -1.501
## MSL.Pressure.Dmean:Year                  9.159e-04  1.434e-04   6.386
## MSL.Pressure.Dmax:Year                  -6.936e-04  1.388e-04  -4.999
## WindSpeed.Dmin.10m:Year                 -1.990e-02  9.945e-03  -2.001
##                                         Pr(>|z|)
## (Intercept)                             0.004007 **
## TEMP.Dmean.2m                           0.008370 **
## WindSpeed.Dmin.10m                      0.022772 *
## TEMP.Dmean.2m:MSL.Pressure.Dmean        4.54e-09 ***
## TEMP.Dmean.2m:WindDirect.Dmean.900mb    0.000654 ***
## MSL.Pressure.Dmean:WindDirect.Dmean.900mb 0.018823 *
## MSL.Pressure.Dmean:MSL.Pressure.Dmax    2.02e-05 ***
## MSL.Pressure.Dmax:Midlay.Cld.Dmean      0.032933 *
## MSL.Pressure.Dmax:WindSpeed.Dmean.80m   0.002432 **
## MSL.Pressure.Dmean:MSL.Pressure.Dmin    1.51e-08 ***
## WindDirect.Dmean.900mb:MSL.Pressure.Dmin  0.018355 *
## TEMP.Dmean.2m:Midlay.Cld.Dmax           0.000602 ***
## Midlay.Cld.Dmean:Midlay.Cld.Dmax        0.056779 .
## WindSpeed.Dmean.80m:Midlay.Cld.Dmax     0.099058 .
## WindDirect.Dmean.900mb:WindSpeed.Dmax.10m 0.016130 *
## MSL.Pressure.Dmax:WindSpeed.Dmax.10m    0.027974 *
## MSL.Pressure.Dmin:WindSpeed.Dmax.10m    0.029807 *
## Midlay.Cld.Dmax:WindSpeed.Dmax.10m      0.001187 **
## TEMP.Dmean.2m:WindSpeed.Dmin.10m        0.122123
## WindSpeed.Dmin.10m:MSL.Pressure.Dmean   0.057787 .
## WindSpeed.Dmin.10m:Midlay.Cld.Dmean     0.035305 *
## TEMP.Dmean.2m:WindGust.Dmax             0.001544 **
## MSL.Pressure.Dmean:WindGust.Dmax        0.000780 ***
## WindDirect.Dmean.900mb:WindGust.Dmax    0.128659
## MSL.Pressure.Dmin:WindGust.Dmax         0.000748 ***
```

```
## Midlay.Cld.Dmax:WindGust.Dmax             0.017719 *
## TEMP.Dmean.2m:Year                        0.133423
## MSL.Pressure.Dmean:Year                   1.71e-10 ***
## MSL.Pressure.Dmax:Year                    5.78e-07 ***
## WindSpeed.Dmin.10m:Year                   0.045349 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1635.4  on 1179  degrees of freedom
## Residual deviance: 1158.8  on 1150  degrees of freedom
## AIC: 1218.8
##
## Number of Fisher Scoring iterations: 5

pchisq(glm1_L2$deviance,glm1_L2$df.residual,lower=F)

## [1] 0.422156
```

Validation Croisée

3 modèles pour la validation croisée

A ce stade, je m'intéresse à tester les 3 modèles suivants par la validation croisée k-fold (k=6):

Modèle	AIC	Test Khi2 (Deviance)	Interaction covariable	Note
glm.backw	1282.84	0.0415	F	stepwise selection basant sur AIC
glm1_L1	1290.01	0.0244	F	réduire 5 variables de glm.backup
glm1_L2	1218.77	0.4221	T, Level2	avec intération entre les covariables, obtenue en utilisant glmulti(),à partir de glm1_L1

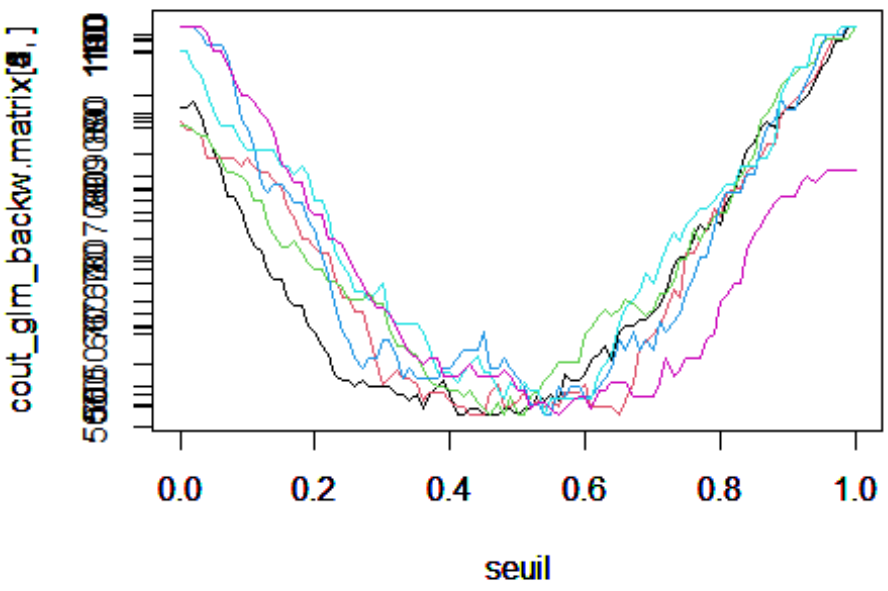
```
AIC(glm.backw)

## [1] 1282.847

AIC(glm1_L1)

## [1] 1290.013

AIC(glm1_L2)

## [1] 1218.774

pchisq(glm.backw$deviance,glm.backw$df.residual,lower=F)

## [1] 0.04157669
```

```
pchisq(glm1_L1$deviance,glm1_L1$df.residual,lower=F)

## [1] 0.02449443

pchisq(glm1_L2$deviance,glm1_L2$df.residual,lower=F)

## [1] 0.422156
```

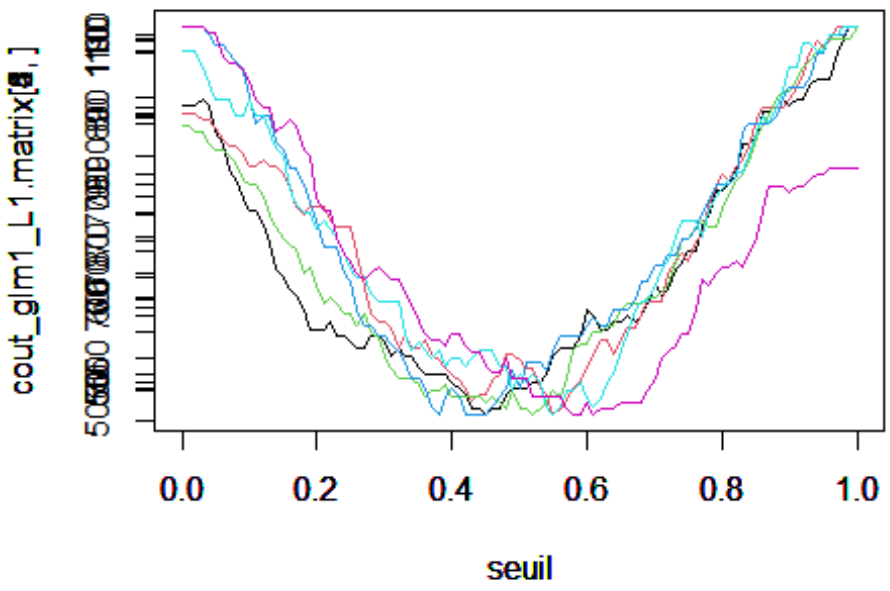
Cherche de seuil optimisé pour la prédiction.

Avant de lancer la validation croisé k-fold, il y a besoin que je cherche le seuil optimisé pour la prédiction. Je suppose que le coût de faire une mauvaise prédiction sur une journée où il pleut, et sur une journée où il ne pleut pas est pareil. Comme chaque échantillon des données (k-ième fold) va donner une courbe de coût différente, je crée une boucle 'for' pour vérifier la courbe de coût sur l'ensemble des données du fichier 'meteo.train.csv'.

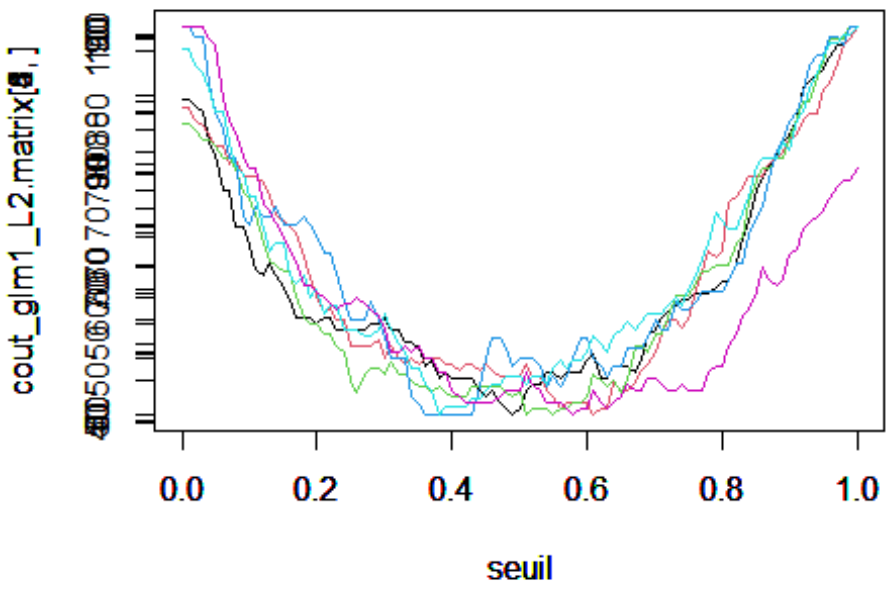
```
k = 6
index = sample(1:k, nrow(m), replace=T)
seuil = seq(0, 1, by=.01)

# chercher le seuil optimisé pour le modèle 'glm.backw'
cout_glm.backw = rep(NA, length(seuil)*k)
for(i in 1:k){

reg=glm(formula = formula(glm.backw),family = binomial,data=m.glm[index != i,])
pred=predict(reg,newdata=m.glm[index==i,],type="response")


  for(j in 1:length(seuil)){

    pred2 = (pred >= seuil[j])
    cout_glm.backw[j+101*(i-1)] = 1 * sum(pred2 & m.glm[index==i,]$pluie.demain==FALSE)
 +
      1* sum(!pred2 & m.glm[index==i,]$pluie.demain==TRUE)

  }

}
cout_glm_backw.matrix=matrix(cout_glm.backw,byrow = T,nrow=k)
plot(seuil,cout_glm_backw.matrix[1,],type = "l")
par(new=T)
plot(seuil,cout_glm_backw.matrix[2,],type = "l",col=2)
par(new=T)
plot(seuil,cout_glm_backw.matrix[3,],type = "l",col=3)
par(new=T)
plot(seuil,cout_glm_backw.matrix[4,],type = "l",col=4)
par(new=T)
plot(seuil,cout_glm_backw.matrix[5,],type = "l",col=5)
par(new=T)
plot(seuil,cout_glm_backw.matrix[6,],type = "l",col=6)
```




```r
# chercher le seuil optimisé pour le modèle 'glm1_L1'
cout_glm1_L1 = rep(NA, length(seuil)*k)
for(i in 1:k){

reg=glm(formula = formula(glm1_L1),family = binomial,data=m.glm[index != i,])
pred=predict(reg,newdata=m.glm[index==i,],type="response")


  for(j in 1:length(seuil)){

    pred2 = (pred >= seuil[j])
    cout_glm1_L1[j+101*(i-1)] = 1 * sum(pred2 & m.glm[index==i,]$pluie.demain==FALSE) +
 1* sum(!pred2 & m.glm[index==i,]$pluie.demain==TRUE)

  }

}
cout_glm1_L1.matrix=matrix(cout_glm1_L1,byrow = T,nrow=k)
plot(seuil,cout_glm1_L1.matrix[1,],type = "l")
par(new=T)
plot(seuil,cout_glm1_L1.matrix[2,],type = "l",col=2)
par(new=T)
plot(seuil,cout_glm1_L1.matrix[3,],type = "l",col=3)
par(new=T)
plot(seuil,cout_glm1_L1.matrix[4,],type = "l",col=4)
par(new=T)
plot(seuil,cout_glm1_L1.matrix[5,],type = "l",col=5)
par(new=T)
plot(seuil,cout_glm1_L1.matrix[6,],type = "l",col=6)
```



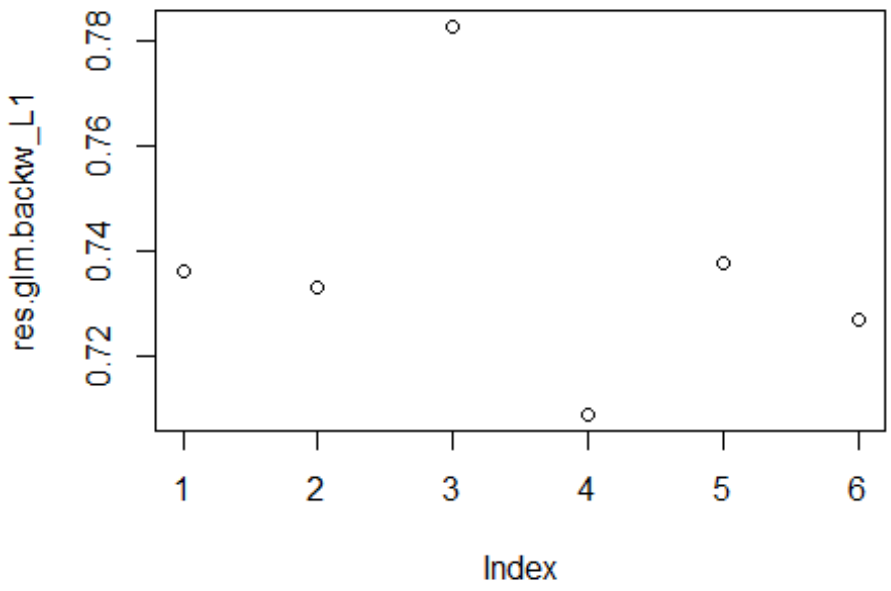

```r
# chercher le seuil optimisé pour le modèle 'glm1_L2'
cout_glm1_L2 = rep(NA, length(seuil)*k)
for(i in 1:k){

reg=glm(formula = formula(glm1_L2),family = binomial,data=m.glm[index != i,])
pred=predict(reg,newdata=m.glm[index==i,],type="response")


  for(j in 1:length(seuil)){

    pred2 = (pred >= seuil[j])
    cout_glm1_L2[j+101*(i-1)] = 1 * sum(pred2 & m.glm[index==i,]$pluie.demain==FALSE) +
 1* sum(!pred2 & m.glm[index==i,]$pluie.demain==TRUE)

  }

}
cout_glm1_L2.matrix=matrix(cout_glm1_L2,byrow = T,nrow=k)
plot(seuil,cout_glm1_L2.matrix[1,],type = "l")
par(new=T)
plot(seuil,cout_glm1_L2.matrix[2,],type = "l",col=2)
par(new=T)
plot(seuil,cout_glm1_L2.matrix[3,],type = "l",col=3)
par(new=T)
plot(seuil,cout_glm1_L2.matrix[4,],type = "l",col=4)
par(new=T)
plot(seuil,cout_glm1_L2.matrix[5,],type = "l",col=5)
par(new=T)
plot(seuil,cout_glm1_L2.matrix[6,],type = "l",col=6)
```



Depuis les graphs de courbe de coût, on voit que le seuil optimisé de prédiction se trouve aux alentours de 0.5 pour tous les trois modèles. Je vais tester les différentes valeurs entre 0.5 et 0.6 dans la validation croisée.

Validation croisée de k-fold modèle - glm.backw

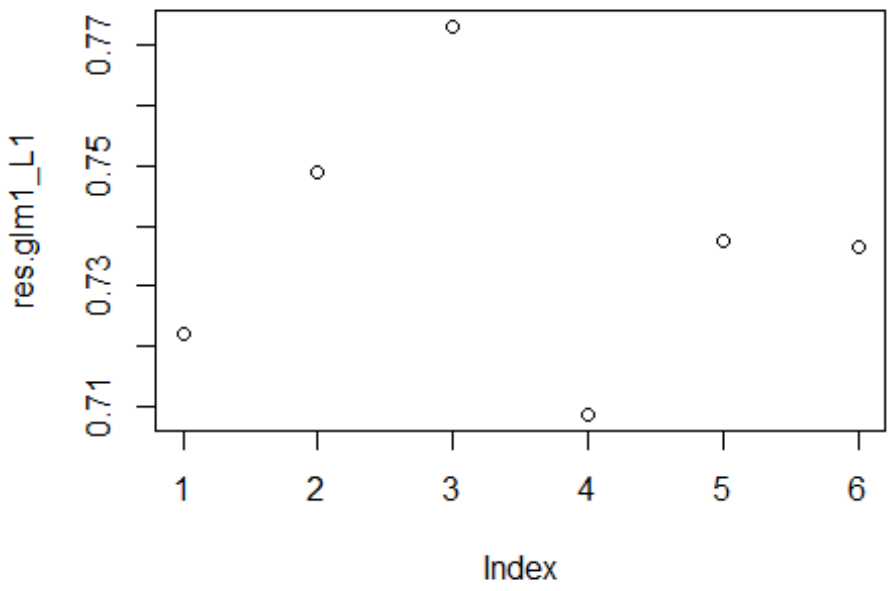
```
res.glm.backw_L1 = rep(NA, k)
for(i in 1:k){
  reg.glm.backw_L1 = glm(
    formula = formula(glm.backw),
    family = binomial,
    data = m[index != i, ]
  )


  pred.glm.backw_L1 = predict(reg.glm.backw_L1, newdata=m[index == i, ],
                              type="response")

  res.glm.backw_L1[i] = mean(m[index==i, "pluie.demain"] == (pred.glm.backw_L1 >.5), na
.rm = T)

}
plot(res.glm.backw_L1)
```




```
mean(res.glm.backw_L1)

## [1] 0.7374591
```

Validation croisée de k-fold modèle - glm1_L1

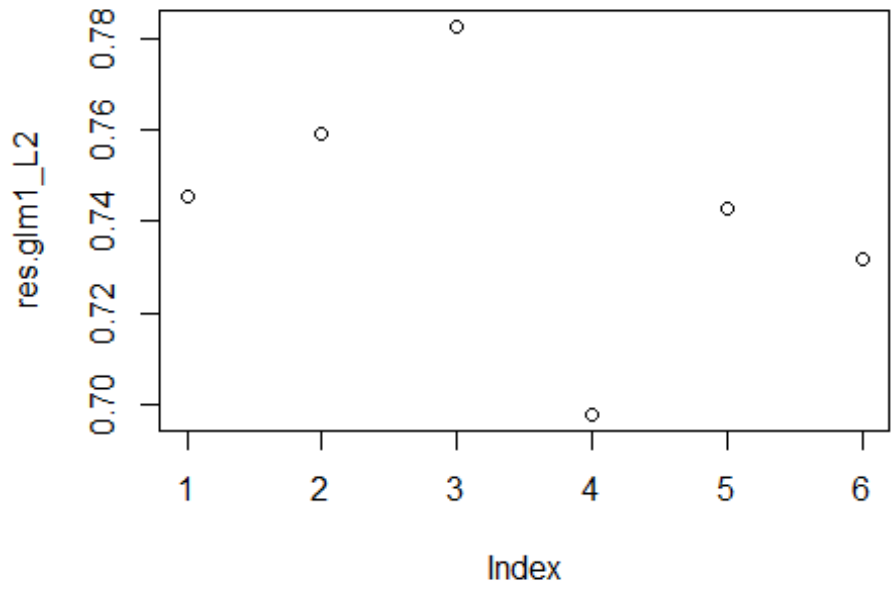
```
res.glm1_L1 = rep(NA, k)
for(i in 1:k){
  reg.glm1_L1 = glm(
    formula = formula(glm1_L1),
    family = binomial,
    data = m[index != i, ]
  )


  pred.glm1_L1 = predict(reg.glm1_L1, newdata=m[index == i, ],type="response")

  res.glm1_L1[i] = mean(m[index==i, "pluie.demain"] == (pred.glm1_L1 >.53), na.rm = T)

}
plot(res.glm1_L1)
```




```
mean(res.glm1_L1)

## [1] 0.7377778
```

Validation croisée de k-fold modèle - glm1_L2

```
res.glm1_L2 = rep(NA, k)
for(i in 1:k){
  reg.glm1_L2 = glm(
    formula = formula(glm1_L2),
    family = binomial,
    data = m[index != i, ]
  )


  pred.glm1_L2 = predict(reg.glm1_L2, newdata=m[index == i, ],type="response")

  res.glm1_L2[i] = mean(m[index==i, "pluie.demain"] == (pred.glm1_L2 >.53), na.rm = T)

}
plot(res.glm1_L2)
```




```
mean(res.glm1_L2)

## [1] 0.7432779
```

Choix de modèle prédictif via la validation croisée

Le performance de prediction entre **glm.backw** et **glm_L1** est très proche dans la validation croisée. Le fait de réduire 5 covariables sur la modèle **glm.backw** semble pas changer la performance de prédiction.

Le meilleur modèle de prédiction selon la validation croisée est glm1_L2. C'est le modèle qui prend en compte de l'interaction entre les covariables.

Je choisi donc glm1_L2 dans le but de prédiction sur le fichier 'meteo.test.csv' . Neanmoins, je remarque que les coefficients sous ce modèle sont difficiles à interprêter, et certaines sont biasés.

Modèle	AIC	Test Khi2 (Deviance)	Seuil Optimisé	Taux de Bonne Prediction
glm.backw	1282.84	0.0415	0.5	0.737
glm1_L1	1290.01	0.0244	0.53	0.737
glm1_L2	1218.77	0.4221	0.53	0.743

Prédiction sur 'meteo.test.csv'

```
meteo.predict=meteo.predict.initial %>%
  rename(
    TEMP.Dmean.2m = Temperature.daily.mean..2.m.above.gnd.,
    ReHumidity.Dmean.2m = Relative.Humidity.daily.mean..2.m.above.gnd.,
    MSL.Pressure.Dmean = Mean.Sea.Level.Pressure.daily.mean..MSL.,
    Ttl.Precipitation.Dsum = Total.Precipitation.daily.sum..sfc.,
    Snowfall.Dsum = Snowfall.amount.raw.daily.sum..sfc.,
    Ttl.Cld.Dmean = Total.Cloud.Cover.daily.mean..sfc.,
    Highlay.Cld.Dmean = High.Cloud.Cover.daily.mean..high.cld.lay.,
    Midlay.Cld.Dmean = Medium.Cloud.Cover.daily.mean..mid.cld.lay.,
    Lowlay.Cld.Dmean = Low.Cloud.Cover.daily.mean..low.cld.lay.,
    Sunshine.Dsum = Sunshine.Duration.daily.sum..sfc.,
    Shortwave.Dsum = Shortwave.Radiation.daily.sum..sfc.,
    WindSpeed.Dmean.10m = Wind.Speed.daily.mean..10.m.above.gnd.,
    WindDirect.Dmean.10m = Wind.Direction.daily.mean..10.m.above.gnd.,
    WindSpeed.Dmean.80m = Wind.Speed.daily.mean..80.m.above.gnd.,
    WindDirect.Dmean.80m = Wind.Direction.daily.mean..80.m.above.gnd.,
    WindSpeed.Dmean.900mb = Wind.Speed.daily.mean..900.mb.,
    WindDirect.Dmean.900mb = Wind.Direction.daily.mean..900.mb.,
    WindGust.Dmean = Wind.Gust.daily.mean..sfc.,
    TEMP.Dmax.2m = Temperature.daily.max..2.m.above.gnd.,
    TEMP.Dmin.2m = Temperature.daily.min..2.m.above.gnd.,
    R.Humidity.Dmax.2m = Relative.Humidity.daily.max..2.m.above.gnd.,
    R.Humidity.Dmin.2m = Relative.Humidity.daily.min..2.m.above.gnd.,
    MSL.Pressure.Dmax = Mean.Sea.Level.Pressure.daily.max..MSL.,
    MSL.Pressure.Dmin = Mean.Sea.Level.Pressure.daily.min..MSL.,
    Ttl.Cld.Dmax = Total.Cloud.Cover.daily.max..sfc.,
    Ttl.Cld.Dmin = Total.Cloud.Cover.daily.min..sfc.,
    Highlay.Cld..Dmax = High.Cloud.Cover.daily.max..high.cld.lay.,
    Highlay.Cld.Dmin = High.Cloud.Cover.daily.min..high.cld.lay.,
    Midlay.Cld.Dmax = Medium.Cloud.Cover.daily.max..mid.cld.lay.,
    Midlay.Cld.Dmin = Medium.Cloud.Cover.daily.min..mid.cld.lay.,
    Lowlay.Cld.Dmax = Low.Cloud.Cover.daily.max..low.cld.lay.,
    Lowlay.Cld.Dmin = Low.Cloud.Cover.daily.min..low.cld.lay.,
    WindSpeed.Dmax.10m = Wind.Speed.daily.max..10.m.above.gnd.,
    WindSpeed.Dmin.10m = Wind.Speed.daily.min..10.m.above.gnd.,
    WindSpeed.Dmax.80m = Wind.Speed.daily.max..80.m.above.gnd.,
    WindSpeed.Dmin.80m = Wind.Speed.daily.min..80.m.above.gnd.,
    WindSpeed.Dmax.900mb = Wind.Speed.daily.max..900.mb.,
    WindSpeed.Dmin.900mb = Wind.Speed.daily.min..900.mb.,
    WindGust.Dmax = Wind.Gust.daily.max..sfc.,
    WindGust.Dmin = Wind.Gust.daily.min..sfc.
  )

meteo.predict$pluie.demain.proba=predict(glm1_L2,newdata = meteo.predict,type = "respon
se")
meteo.predict$pluie.demain.predict=meteo.predict$pluie.demain.proba>=0.53
meteo.predict.initial$pluie.demain.predict=meteo.predict$pluie.demain.predict
write_csv(meteo.predict.initial, "C:\\Users\\ZQFX\\Desktop\\formation Data Science\\Cou
rs Dauphine\\Modeles Lineaires Genarales - R. RYDER-20210919\\Projet RLG\\summit projet
\\meteo.test_tovalidate.csv")
```