CC Modèles logistiques et analyse discriminante

Paul Le Breton et Timothée Templier

16/04/2024

Régression Logistique sur les observations journalières météorolgiques à Melbourne

Mise en place des données :

Importation des packages nécessaires à l'étude :

```
library(boot)
library(glmnet)
library(generalhoslem)
library(car)
library(questionr)
```

Importation du jeu de données :

```
data <- read.table('weatherMelbourne.txt')
```

Transformation de la variable à expliquer en variable binaire 1 - 0:

```
data$RainTomorrow <- ifelse(data$RainTomorrow == "Yes", 1, 0)

data$RainToday <- ifelse(data$RainToday == "Yes", 1, 0)
```

Transformation des variables qualitatives et de la variable à expliquer en factor :

```
data$WindDir9am <- as.factor(data$WindDir9am)
data$WindDir3pm <- as.factor(data$WindDir3pm)
data$RainToday <- as.factor(data$RainToday)
data$Cloud3pm <- as.factor(data$Cloud3pm)
data$Cloud9am <- as.factor(data$Cloud9am)
data$RainTomorrow <- as.factor(data$RainTomorrow)
```

Introduction

Ce document se penche sur l'analyse de la météo de Melbourne, on dispose de 19 variables explicatives sur la meteo à différents moments de la journée,telles que la direction et la vitesse du vent, la température et l'humidité. L'objectif principal est de déterminer quelles variables sont les plus significatives pour prédire s'il va pleuvoir le lendemain.

Pour ce faire, plusieurs modèles seront étudiés, et le choix du meilleur modèle se fera en se basant sur les critères enseignés au cours de cette année universitaire.

Construction des modèles

Construction du premier modèle qui explique RainTomorrow avec l'ensemble des autres variables du jeu de données:

```
m1 <- glm(RainTomorrow~., family=binomial, data=data)
summary(m1)
```

```
##
## Call:
## glm(formula = RainTomorrow ~ ., family = binomial, data = data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    78.855821  12.377121   6.371 1.88e-10 ***
## MinTemp        -0.034716   0.047885  -0.725 0.468452
## MaxTemp        -0.114379   0.062808  -1.821 0.068593 .
## Rainfall        0.043358   0.013795   3.143 0.001672 **
## Evaporation     0.016213   0.032852   0.493 0.621663
## Sunshine       -0.106407   0.036976  -2.878 0.004005 **
## WindGustSpeed   0.032259   0.008322   3.877 0.000106 ***
## WindDir9amENE   0.153630   1.273210   0.121 0.903958
## WindDir9amESE -14.807744 419.693578  -0.035 0.971855
## WindDir9amN     0.038246   0.996612   0.038 0.969388
## WindDir9amNE    0.962986   1.079297   0.892 0.372267
## WindDir9amNNE  -0.219772   1.025463  -0.214 0.830302
## WindDir9amNNW  -0.372782   1.027192  -0.363 0.716669
## WindDir9amNW   -0.366352   1.048371  -0.349 0.726752
## WindDir9amS    -0.731210   1.075248  -0.680 0.496480
## WindDir9amSE   -0.519607   1.152619  -0.451 0.652130
## WindDir9amSSE  -1.441151   1.166289  -1.236 0.216580
## WindDir9amSSW  -0.530478   1.064533  -0.498 0.618259
## WindDir9amSW    0.142465   1.029942   0.138 0.889985
## WindDir9amW     0.108826   1.019280   0.107 0.914973
## WindDir9amWNW   0.170217   1.034949   0.164 0.869362
## WindDir9amWSW  -0.069958   1.020410  -0.069 0.945341
## WindDir3pmENE   1.140313   1.846868   0.617 0.536951
## WindDir3pmESE   0.279948   1.737227   0.161 0.871978
## WindDir3pmN     1.144317   1.404785   0.815 0.415310
## WindDir3pmNE    2.128791   1.513426   1.407 0.159545
## WindDir3pmNNE   1.433231   1.436228   0.998 0.318321
## WindDir3pmNNW   1.931493   1.415597   1.364 0.172430
## WindDir3pmNW    1.973200   1.423178   1.386 0.165602
## WindDir3pmS     1.097020   1.413161   0.776 0.437579
## WindDir3pmSE    1.054318   1.486212   0.709 0.478077
## WindDir3pmSSE   0.919840   1.423723   0.646 0.518227
## WindDir3pmSSW   1.308689   1.411594   0.927 0.353875
## WindDir3pmSW    0.745588   1.426665   0.523 0.601247
## WindDir3pmW     1.918430   1.424238   1.347 0.177985
## WindDir3pmWNW   2.049525   1.441016   1.422 0.154946
## WindDir3pmWSW   1.571657   1.421979   1.105 0.269047
## WindSpeed9am   -0.011263   0.009974  -1.129 0.258794
## WindSpeed3pm   -0.009392   0.010789  -0.871 0.384015
## Humidity9am    -0.003994   0.009269  -0.431 0.666558
```

```
## Humidity3pm     0.058543   0.009312   6.287 3.24e-10 ***
## Pressure9am     0.089062   0.041382   2.152 0.031380 *
## Pressure3pm    -0.173073   0.041411  -4.179 2.92e-05 ***
## Cloud9am1       0.242526   0.673646   0.360 0.718832
## Cloud9am2       0.504493   0.701432   0.719 0.471998
## Cloud9am3      -0.118394   0.688161  -0.172 0.863403
## Cloud9am4       0.168446   0.724628   0.232 0.816182
## Cloud9am5      -0.157289   0.695742  -0.226 0.821144
## Cloud9am6       0.132955   0.673541   0.197 0.843517
## Cloud9am7      -0.075124   0.657286  -0.114 0.909005
## Cloud9am8       0.125872   0.697081   0.181 0.856706
## Cloud3pm1      -0.834281   0.859868  -0.970 0.331925
## Cloud3pm2       0.101805   0.845003   0.120 0.904104
## Cloud3pm3      -0.845626   0.913956  -0.925 0.354843
## Cloud3pm4      -0.051111   0.839799  -0.061 0.951470
## Cloud3pm5      -0.201017   0.821086  -0.245 0.806597
## Cloud3pm6      -0.081511   0.807069  -0.101 0.919553
## Cloud3pm7       0.079154   0.801652   0.099 0.921346
## Cloud3pm8       0.326332   0.852383   0.383 0.701833
## Temp9am         0.085676   0.067268   1.274 0.202788
## Temp3pm         0.088066   0.068083   1.294 0.195832
## RainToday1      0.366468   0.189078   1.938 0.052600 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2126.9  on 1897  degrees of freedom
## Residual deviance: 1413.7  on 1836  degrees of freedom
## AIC: 1537.7
##
## Number of Fisher Scoring iterations: 14
```

Pour ce modèle, la contrainte d'identifiabilité est telle que le coefficient associé à la 1ère modalité de chaque variable qualitative explicative est nul.

Pour le premier modèle, on peut voir que seulement 6 variables semblent significatives au niveau 5%. Ces variables sont : Rainfall, Sunshine, WindGustSpeed, Humidity3pm, Pressure9am, Pressure3pm.

Détections des meilleurs modèles selon les crière d'AIC et BIC On peut trouver de nouveaux modèles grâce au package bestglm qui permet de sélectionner le meilleur groupe de variables explicatives selon le critère AIC ou BIC.

On a rencontré un problème : il y a trop de variables explicatives dans le modèle.

On va donc passer par une méthode pas à pas forward pour sélectionner les meilleurs variables explicatives.

```
m0<-glm(RainTomorrow~1,data=data,family=binomial)
stepAIC<-step(m0,scope=formula(m1),direction="forward")
```

```
stepBIC<-step(m0,scope=formula(m1),direction="forward",k=log(nrow(data)))
```

On réecrit donc les modèles avec les variables explicatives fournis selon ces méthodes pas à pas forward :

```r
modAIC <- glm(RainTomorrow~WindSpeed3pm+WindSpeed9am+Temp9am+Temp3pm+Humidity9am+MaxTemp+MinTemp+WindDi

modBIC <- glm(RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+WindSpeed9am+Temp9am+Evaporation+
```

```r
summary(modAIC)
```

```
##
## Call:
## glm(formula = RainTomorrow ~ WindSpeed3pm + WindSpeed9am + Temp9am +
##     Temp3pm + Humidity9am + MaxTemp + MinTemp + WindDir9am +
##     Cloud3pm + Cloud9am, family = binomial, data = data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.868502   1.475130  -3.300 0.000966 ***
## WindSpeed3pm    0.016101   0.007610   2.116 0.034354 *
## WindSpeed9am    0.020778   0.008035   2.586 0.009708 **
## Temp9am         0.273689   0.054332   5.037 4.72e-07 ***
## Temp3pm        -0.174723   0.045001  -3.883 0.000103 ***
## Humidity9am     0.032102   0.006592   4.870 1.12e-06 ***
## MaxTemp        -0.053772   0.052241  -1.029 0.303334
## MinTemp        -0.005536   0.041921  -0.132 0.894939
## WindDir9amENE   0.125018   1.183966   0.106 0.915906
## WindDir9amESE -14.599571 444.485134  -0.033 0.973797
## WindDir9amN     0.148324   0.932932   0.159 0.873679
## WindDir9amNE    0.752834   1.012585   0.743 0.457193
## WindDir9amNNE   0.100436   0.959095   0.105 0.916599
## WindDir9amNNW   0.117713   0.955839   0.123 0.901987
## WindDir9amNW   -0.133934   0.971943  -0.138 0.890398
## WindDir9amS    -1.389845   0.997260  -1.394 0.163419
## WindDir9amSE   -1.152691   1.060805  -1.087 0.277205
## WindDir9amSSE  -1.957186   1.064608  -1.838 0.066002 .
## WindDir9amSSW  -1.282680   0.986222  -1.301 0.193395
## WindDir9amSW   -0.737069   0.956015  -0.771 0.440718
## WindDir9amW    -0.180115   0.945937  -0.190 0.848989
## WindDir9amWNW   0.153453   0.965703   0.159 0.873746
## WindDir9amWSW  -0.545635   0.950953  -0.574 0.566119
## Cloud3pm1      -0.424231   0.850828  -0.499 0.618055
## Cloud3pm2       0.287899   0.834781   0.345 0.730185
## Cloud3pm3      -0.605724   0.900687  -0.673 0.501257
## Cloud3pm4       0.439162   0.825101   0.532 0.594551
## Cloud3pm5       0.272169   0.804431   0.338 0.735109
## Cloud3pm6       0.573457   0.788327   0.727 0.466959
## Cloud3pm7       1.092983   0.776351   1.408 0.159177
## Cloud3pm8       1.946755   0.811088   2.400 0.016387 *
## Cloud9am1       0.535843   0.661354   0.810 0.417813
## Cloud9am2       0.717125   0.690183   1.039 0.298788
## Cloud9am3       0.395003   0.674929   0.585 0.558378
## Cloud9am4       0.684899   0.709170   0.966 0.334157
## Cloud9am5       0.279568   0.680237   0.411 0.681083
## Cloud9am6       0.743189   0.656614   1.132 0.257698
## Cloud9am7       0.581968   0.641451   0.907 0.364265
## Cloud9am8       1.096666   0.670424   1.636 0.101886
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2126.9  on 1897  degrees of freedom
## Residual deviance: 1645.6  on 1859  degrees of freedom
## AIC: 1723.6
##
## Number of Fisher Scoring iterations: 14
```

Pour le meilleur modèle selon le critère AIC, on obtient 6 variables explicatives significatives au niveau 5%. Ces variables sont : WindSpeed3pm, WindSpeed9am, Temp9am, Temp3pm, Humidity9am et Cloud3pm8.

```
summary(modBIC)
```

```
##
## Call:
## glm(formula = RainTomorrow ~ Pressure9am + WindSpeed3pm + RainToday +
##     MinTemp + WindSpeed9am + Temp9am + Evaporation + MaxTemp +
##     Humidity9am + Temp3pm + Cloud3pm + Cloud9am + WindDir3pm +
##     WindDir9am, family = binomial, data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    77.830926  11.207750   6.944 3.80e-12 ***
## Pressure9am    -0.080205   0.010674  -7.514 5.72e-14 ***
## WindSpeed3pm    0.005076   0.008647   0.587 0.557211
## RainToday1      0.397327   0.155536   2.555 0.010632 *
## MinTemp        -0.047936   0.045805  -1.047 0.295317
## WindSpeed9am    0.007261   0.008702   0.834 0.404065
## Temp9am         0.258852   0.058640   4.414 1.01e-05 ***
## Evaporation    -0.012636   0.030831  -0.410 0.681925
## MaxTemp        -0.039585   0.056472  -0.701 0.483326
## Humidity9am     0.028247   0.007601   3.716 0.000202 ***
## Temp3pm        -0.166147   0.050057  -3.319 0.000903 ***
## Cloud3pm1      -0.758513   0.849163  -0.893 0.371724
## Cloud3pm2      -0.042188   0.834124  -0.051 0.959662
## Cloud3pm3      -1.119222   0.918082  -1.219 0.222811
## Cloud3pm4      -0.127987   0.828771  -0.154 0.877271
## Cloud3pm5      -0.314789   0.809053  -0.389 0.697214
## Cloud3pm6      -0.023789   0.789628  -0.030 0.975966
## Cloud3pm7       0.478166   0.777366   0.615 0.538481
## Cloud3pm8       1.282119   0.814367   1.574 0.115401
## Cloud9am1       0.255008   0.671787   0.380 0.704245
## Cloud9am2       0.579144   0.699398   0.828 0.407636
## Cloud9am3       0.062365   0.687106   0.091 0.927680
## Cloud9am4       0.406845   0.719909   0.565 0.571982
## Cloud9am5       0.028609   0.691667   0.041 0.967007
## Cloud9am6       0.484760   0.666867   0.727 0.467274
## Cloud9am7       0.435322   0.647659   0.672 0.501490
## Cloud9am8       0.831343   0.680175   1.222 0.221613
## WindDir3pmENE   1.259372   1.556516   0.809 0.418460
```

```
## WindDir3pmESE  -0.130292   1.501257  -0.087 0.930840
## WindDir3pmN     0.956691   1.199206   0.798 0.425004
## WindDir3pmNE    1.689791   1.321860   1.278 0.201129
## WindDir3pmNNE   1.228381   1.229621   0.999 0.317799
## WindDir3pmNNW   1.556890   1.209616   1.287 0.198061
## WindDir3pmNW    1.350832   1.218556   1.109 0.267623
## WindDir3pmS     0.405253   1.201236   0.337 0.735843
## WindDir3pmSE    0.424256   1.282628   0.331 0.740817
## WindDir3pmSSE   0.248493   1.215614   0.204 0.838027
## WindDir3pmSSW   0.646634   1.202273   0.538 0.590685
## WindDir3pmSW    0.094470   1.220176   0.077 0.938287
## WindDir3pmW     1.101789   1.215062   0.907 0.364525
## WindDir3pmWNW   1.332928   1.231513   1.082 0.279097
## WindDir3pmWSW   0.714658   1.211184   0.590 0.555157
## WindDir9amENE  -0.025045   1.202331  -0.021 0.983381
## WindDir9amESE -14.694655 436.456552  -0.034 0.973142
## WindDir9amN    -0.218654   0.939024  -0.233 0.815876
## WindDir9amNE    0.654416   1.026481   0.638 0.523778
## WindDir9amNNE  -0.289049   0.967433  -0.299 0.765108
## WindDir9amNNW  -0.772546   0.970152  -0.796 0.425850
## WindDir9amNW   -0.662895   0.987002  -0.672 0.501823
## WindDir9amS    -1.221617   1.016106  -1.202 0.229265
## WindDir9amSE   -0.789712   1.086636  -0.727 0.467379
## WindDir9amSSE  -1.740729   1.091147  -1.595 0.110641
## WindDir9amSSW  -1.219033   1.005906  -1.212 0.225560
## WindDir9amSW   -0.476684   0.971182  -0.491 0.623548
## WindDir9amW    -0.463973   0.961624  -0.482 0.629459
## WindDir9amWNW  -0.310582   0.979128  -0.317 0.751090
## WindDir9amWSW  -0.533294   0.962792  -0.554 0.579645
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2126.9  on 1897  degrees of freedom
## Residual deviance: 1515.1  on 1841  degrees of freedom
## AIC: 1629.1
##
## Number of Fisher Scoring iterations: 14
```

Pour le meilleur modèle selon le critère du BIC, on obtient 5 variables explicatives significatives au niveau 5%. Ces variables sont : Pressure9am, RainToday1, Temp9am, Humidity9am, Temp3pm.

Tests de nouveaux modèles avec intéractions On va tester des intéractions entre variables explicatives dans ces 2 modèles.

Pour modAIC :
Pour retenir le meilleur modèle, on va comparer l'AIC des modèles tests (avec intéraction) avec celui du meilleur modèle selon le critère de l'AIC.

```
# Intéraction entre WindSpeed9am et Temp9am
modtest1 <- glm(RainTomorrow~WindSpeed3pm+WindSpeed9am:Temp9am+Temp3pm+Humidity9am+MaxTemp+MinTemp+Windl

# Intéraction entre WindSpeed3pm et Temp3pm
```

```
modtest2 <- glm(RainTomorrow~WindSpeed3pm:Temp3pm+WindSpeed9am+Temp9am+Humidity9am+MaxTemp+MinTemp+Windl

# Intéraction entre WindSpeed9am et Humidity9am
modtest3 <- glm(RainTomorrow~WindSpeed3pm+Temp3pm+WindSpeed9am:Humidity9am+Temp9am+MaxTemp+MinTemp+Windl

# Intéraction entre Temp9am et Humidity9am
modtest4 <- glm(RainTomorrow~WindSpeed3pm+Temp3pm+WindSpeed9am+Temp9am:Humidity9am+MaxTemp+MinTemp+Windl

# Intéraction entre WindSpeed9am et Humidity9am et entre WindSpeed9am et Humidity9am
modtest5 <- glm(RainTomorrow~WindSpeed3pm:Temp3pm+WindSpeed9am:Humidity9am+Temp9am+MaxTemp+MinTemp+Windl
```

```
c(AIC(modAIC),AIC(modtest1))
```

```
## [1] 1723.556 1746.546
```

```
c(AIC(modAIC),AIC(modtest2))
```

```
## [1] 1723.556 1739.594
```

```
c(AIC(modAIC),AIC(modtest3))
```

```
## [1] 1723.556 1732.103
```

```
c(AIC(modAIC),AIC(modtest4))
```

```
## [1] 1723.556 1733.350
```

```
c(AIC(modAIC),AIC(modtest5))
```

```
## [1] 1723.556 1744.305
```

Au regard du critère de l'AIC, aucun de ces modèles avec intéraction n'est préférable à celui construit à l'aide de la méthode pas à pas.
On va donc retenir le modèle modAIC.

Pour modBIC :

Ici, même chose : on va comparer le BIC de différents modèles construits avec une intéraction avec le modèle construit selon la méthode pas à pas forward avec critère BIC.

```
# Intéraction entre Temp3pm et Cloud3pm
modBIC_bis1<-glm(formula = RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+Evaporation+MaxTemp+

# Intéraction entre WindDir3pm et WindSpeed3pm
modBIC_bis2<-glm(formula = RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+Evaporation+MaxTemp+

# Intéraction entre MaxTemp et RainToday
modBIC_bis3<-glm(formula = RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+Evaporation+MaxTemp+

# Intéraction entre Humidity9am et MaxTemp
```

```r
modBIC_bis4<-glm(formula = RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+Evaporation+MaxTemp+

# Intéraction entre Pressure9am et RainToday
modBIC_bis5<-glm(formula = RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+Evaporation+MaxTemp+

# Intéraction entre Cloud9am et RainToday
modBIC_bis6<-glm(formula = RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+Evaporation+MaxTemp+

# Intéraction entre Cloud3pm et RainToday
modBIC_bis7<-glm(formula = RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+Evaporation+MaxTemp+

# Intéraction entre Humidity9am et RainToday
modBIC_bis<-glm(formula = RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+Evaporation+MaxTemp+Hu
```

```r
c(BIC(modBIC,modBIC_bis1))
```

```
## $df
## [1] 57 63
##
## $BIC
## [1] 1945.397 1989.573
```

```r
c(BIC(modBIC,modBIC_bis2))
```

```
## $df
## [1] 57 70
##
## $BIC
## [1] 1945.397 2044.707
```

```r
c(BIC(modBIC,modBIC_bis3))
```

```
## $df
## [1] 57 56
##
## $BIC
## [1] 1945.397 1950.957
```

```r
c(BIC(modBIC,modBIC_bis4))
```

```
## $df
## [1] 57 56
##
## $BIC
## [1] 1945.397 1955.261
```

```r
c(BIC(modBIC,modBIC_bis5))
```

```
## $df
## [1] 57 56
```

```
##
## $BIC
## [1] 1945.397 1956.775
```

```
c(BIC(modBIC,modBIC_bis6))
```

```
## $df
## [1] 57 63
##
## $BIC
## [1] 1945.397 2005.544
```

```
c(BIC(modBIC,modBIC_bis7))
```

```
## $df
## [1] 57 63
##
## $BIC
## [1] 1945.397 2003.444
```

```
c(BIC(modBIC,modBIC_bis))
```

```
## $df
## [1] 57 56
##
## $BIC
## [1] 1945.397 1944.867
```

On peut voir que le dernier modèle construit (modBIC_bis), celui ayant comme intéractivité Humidity9am selon RainToday, minimise le critère du BIC face au modèle modBIC. C'est donc ce modèle que l'on va retenir pour la suite.

```
summary(modBIC_bis)
```

```
##
## Call:
## glm(formula = RainTomorrow ~ Pressure9am + WindSpeed3pm + RainToday +
##     MinTemp + Evaporation + MaxTemp + Humidity9am + Temp3pm +
##     Cloud3pm + Cloud9am + WindDir3pm + WindDir9am + Humidity9am:RainToday,
##     family = binomial, data = data)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             87.448496  11.080794   7.892 2.98e-15 ***
## Pressure9am             -0.087635   0.010590  -8.275  < 2e-16 ***
## WindSpeed3pm             0.005161   0.007590   0.680 0.496541
## RainToday1              -2.633354   0.858153  -3.069 0.002150 **
## MinTemp                  0.102165   0.030895   3.307 0.000944 ***
## Evaporation             -0.027122   0.030952  -0.876 0.380892
## MaxTemp                  0.049526   0.051420   0.963 0.335466
## Humidity9am             -0.001446   0.007320  -0.198 0.843364
```

```
## Temp3pm                -0.169340   0.049995  -3.387 0.000706 ***
## Cloud3pm1              -0.700142   0.854887  -0.819 0.412794
## Cloud3pm2               0.026114   0.841910   0.031 0.975255
## Cloud3pm3              -1.008844   0.916344  -1.101 0.270921
## Cloud3pm4              -0.024056   0.835207  -0.029 0.977022
## Cloud3pm5              -0.184324   0.814623  -0.226 0.820992
## Cloud3pm6               0.158741   0.795034   0.200 0.841742
## Cloud3pm7               0.688549   0.782584   0.880 0.378946
## Cloud3pm8               1.570884   0.818152   1.920 0.054853 .
## Cloud9am1               0.443577   0.671873   0.660 0.509119
## Cloud9am2               0.785893   0.697747   1.126 0.260026
## Cloud9am3               0.221479   0.686544   0.323 0.746998
## Cloud9am4               0.556820   0.719689   0.774 0.439111
## Cloud9am5               0.225230   0.691143   0.326 0.744515
## Cloud9am6               0.641696   0.666820   0.962 0.335888
## Cloud9am7               0.605741   0.647916   0.935 0.349836
## Cloud9am8               0.909847   0.682583   1.333 0.182549
## WindDir3pmENE           1.102749   1.471351   0.749 0.453567
## WindDir3pmESE          -0.445228   1.447790  -0.308 0.758446
## WindDir3pmN             0.882060   1.102697   0.800 0.423762
## WindDir3pmNE            1.607296   1.225157   1.312 0.189550
## WindDir3pmNNE           1.113448   1.132823   0.983 0.325658
## WindDir3pmNNW           1.364889   1.112742   1.227 0.219973
## WindDir3pmNW            1.164688   1.123883   1.036 0.300059
## WindDir3pmS             0.387336   1.103655   0.351 0.725620
## WindDir3pmSE            0.353647   1.194073   0.296 0.767101
## WindDir3pmSSE           0.154189   1.120522   0.138 0.890553
## WindDir3pmSSW           0.607054   1.105815   0.549 0.583030
## WindDir3pmSW            0.093944   1.125123   0.083 0.933456
## WindDir3pmW             1.044292   1.119791   0.933 0.351038
## WindDir3pmWNW           1.194259   1.137742   1.050 0.293868
## WindDir3pmWSW           0.591009   1.115920   0.530 0.596378
## WindDir9amENE           0.068110   1.151845   0.059 0.952848
## WindDir9amESE         -14.522351 436.158824  -0.033 0.973439
## WindDir9amN            -0.061068   0.882049  -0.069 0.944803
## WindDir9amNE            0.734406   0.979223   0.750 0.453261
## WindDir9amNNE          -0.224864   0.916891  -0.245 0.806266
## WindDir9amNNW          -0.641212   0.920559  -0.697 0.486087
## WindDir9amNW           -0.503119   0.937545  -0.537 0.591520
## WindDir9amS            -0.978747   0.969139  -1.010 0.312537
## WindDir9amSE           -0.746491   1.039568  -0.718 0.472709
## WindDir9amSSE          -1.609367   1.042751  -1.543 0.122737
## WindDir9amSSW          -1.022394   0.952926  -1.073 0.283316
## WindDir9amSW           -0.361460   0.918472  -0.394 0.693917
## WindDir9amW            -0.323862   0.910060  -0.356 0.721938
## WindDir9amWNW          -0.209681   0.929550  -0.226 0.821534
## WindDir9amWSW          -0.407961   0.910666  -0.448 0.654167
## RainToday1:Humidity9am  0.042038   0.011601   3.624 0.000290 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2126.9  on 1897  degrees of freedom
```

```
## Residual deviance: 1522.1  on 1842  degrees of freedom
## AIC: 1634.1
##
## Number of Fisher Scoring iterations: 14
```

Pour le meilleur modèle selon le critère du BIC, on obtient 5 variables explicatives significatives au niveau 5%. Ces variables sont : Pressure9am, RainToday1, MinTemp,Temp3pm,RainToday1:Humidity9am

Comparaisons des modèles modAIC et modBIC_bis Comparaisons avec erreurs VC K-fold:

```
cv.glm(data,modAIC,K=10)$delta[1]
```

```
## [1] 0.14642
```

```
cv.glm(data,modBIC_bis,K=10)$delta[1]
```

```
## [1] 0.1393087
```

Au regard du critère des erreurs VC K-fold, on peut dire que le modèle le plus pertinent est modBIC_bis. Cependant, la différence étant très faible, on ne peut pas retenir un seul modèle. On va donc garder les 2. Ces erreurs sont assez négligeable étant donné qu'elles se situent aux alentours de 0,14 par rapport aux valeurs de Y qui prend 0 ou 1.

On peut aussi comparer ces 2 modèles en regardant la valeur de leur AIC et de leur BIC :

```
c(AIC(modAIC), AIC(modBIC_bis))
```

```
## [1] 1723.556 1634.148
```

```
c(BIC(modAIC), BIC(modBIC_bis))
```

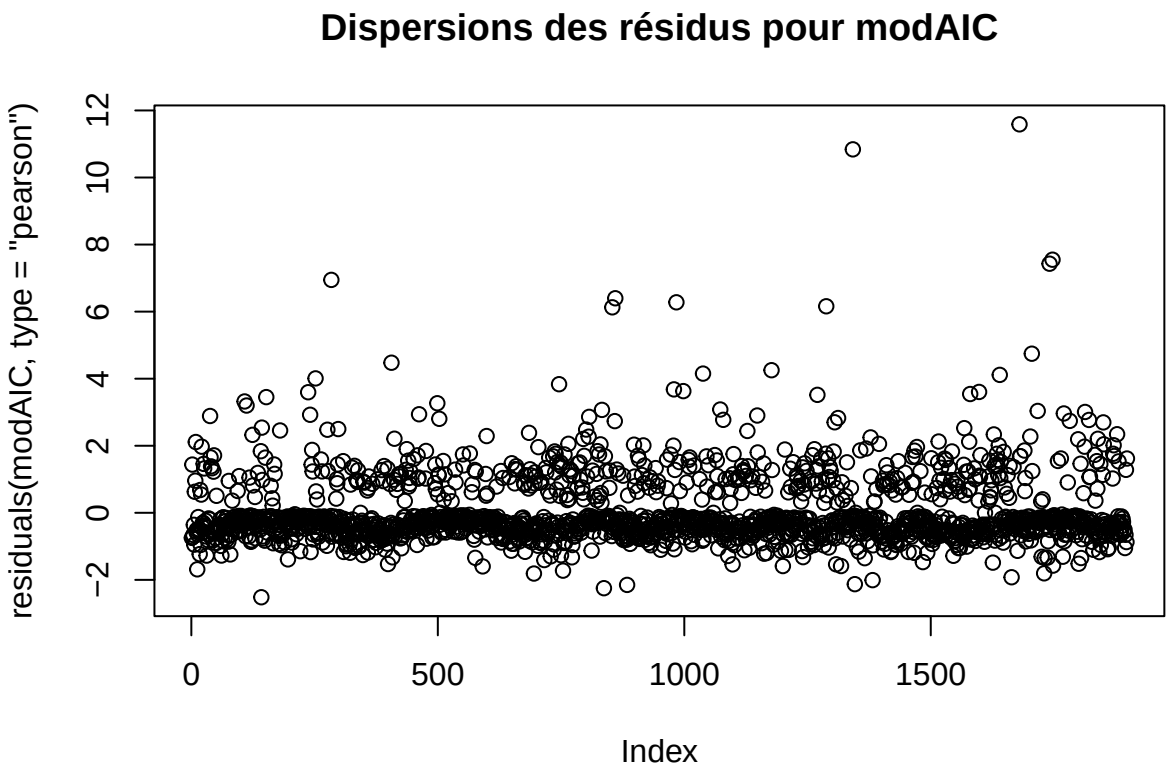
```
## [1] 1939.949 1944.867
```

Au regard du critère de l'AIC, c'est étonnement le modèle modBIC_bis qui semble le plus pertinent pour expliquer la variable RainTomorrow. Au contraire, avec le critère du BIC, c'est le modèle modAIC qui a la plus petite valeur, et que l'on peut donc retenir comme modèle le plus pertinent.

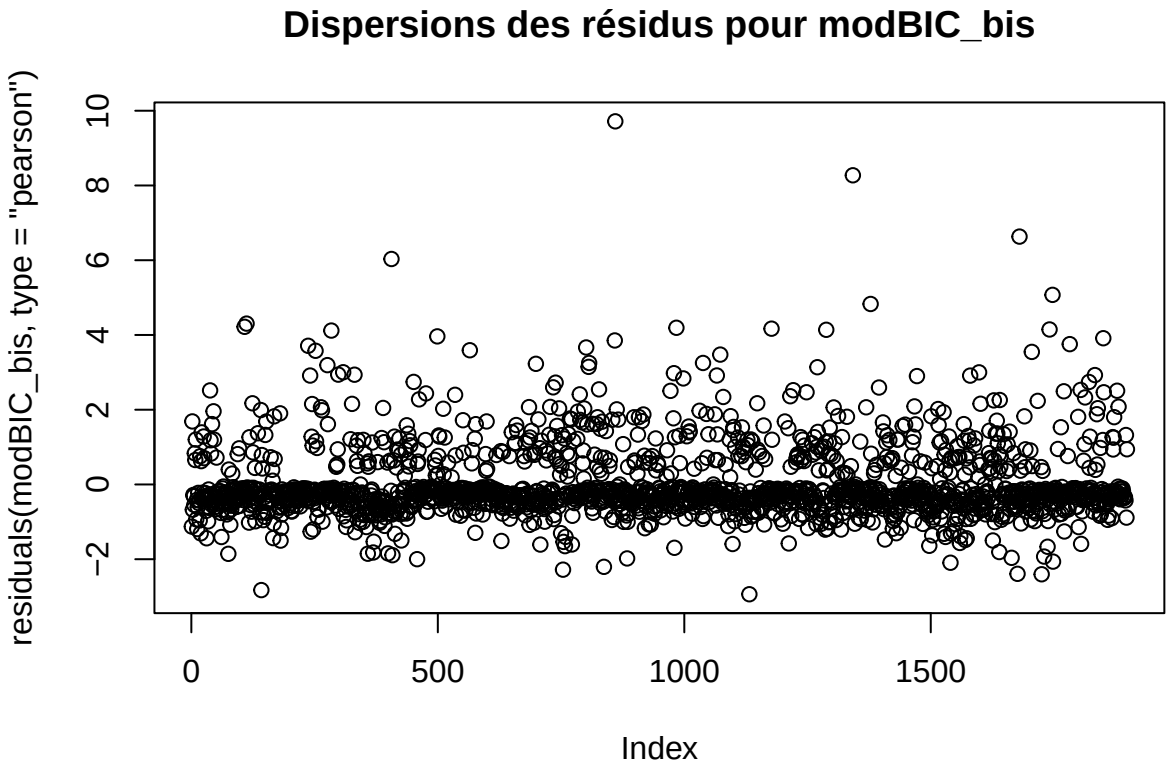
On va donc garder ces 2 modèles car on ne peut pas faire de choix au regard de ces 2 critères.

Analyse des résidus des deux modèles :

```
plot(residuals(modAIC,type="pearson"),main = "Dispersions des résidus pour modAIC")
```




```
plot(residuals(modBIC_bis,type="pearson"),main = "Dispersions des résidus pour modBIC_bis")
```



Les redidus pour les deux modèles sont correctes.

Test d'Hosmer-Lemeshow : -> H0: modèle adapté VS H1: modèle pas adapté

```
logitgof(data$ RainTomorrow,fitted(modAIC))
```

```
##
##  Hosmer and Lemeshow test (binary model)
##
## data:  data$RainTomorrow, fitted(modAIC)
## X-squared = 4.9199, df = 8, p-value = 0.7661
```

Au seuil 5% on ne rejette pas le fait que modAIC soit un bon modèle.

```
logitgof(data$ RainTomorrow,fitted(modBIC_bis))
```

```
##
##  Hosmer and Lemeshow test (binary model)
##
## data:  data$RainTomorrow, fitted(modBIC_bis)
## X-squared = 8.999, df = 8, p-value = 0.3424
```

Au seuil 5% on ne rejette pas le fait que modBIC_bis soit un bon modèle.

```r
id<-sample(c(1:1898),size = 1265)
ech_ap<-data[id,]
ech_test<-data[-id,]
```

```r
modBIC_bis_ap<-glm(formula = RainTomorrow~Pressure9am+WindSpeed3pm+RainToday+MinTemp+Evaporation+MaxTemp

modAIC_ap <- glm(RainTomorrow~WindSpeed3pm+WindSpeed9am+Temp9am+Temp3pm+Humidity9am+MaxTemp+MinTemp+Wind
```

```r
predBIC<-round(predict(modBIC_bis_ap,newdata=ech_test,type="response"),0)
mean(predBIC != ech_test$RainTomorrow)
```

Comparaison des prédictions des deux modèles

```
## [1] 0.2132701
```

```r
predAIC<-round(predict(modAIC_ap,newdata=ech_test,type="response"),0)
mean(predAIC != ech_test$RainTomorrow)
```

```
## [1] 0.2195893
```

On peut voir que les deux modèles présentent des taux d'erreurs de prédiction similaires.Il n'est pas possible de conclure qu'un modèle prédit mieux que l'autre.

Vérifications des modèles finals Test multiple sur les variables des modèles finals :

Nous allons maintenant regarder les valeurs critiques pour les tests de nullité des coefficients associés à chaque variable explicative, par un test de rapport de vraissemblance, pour les 2 modèles retenus.

```r
Anova(modAIC, type=3, test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: RainTomorrow
##               LR Chisq Df Pr(>Chisq)
## WindSpeed3pm     4.479  1   0.034307 *
## WindSpeed9am     6.725  1   0.009509 **
## Temp9am         25.935  1  3.531e-07 ***
## Temp3pm         15.335  1  9.003e-05 ***
## Humidity9am     24.329  1  8.119e-07 ***
## MaxTemp          1.062  1   0.302842
## MinTemp          0.017  1   0.894954
## WindDir9am      67.541 15  1.220e-08 ***
## Cloud3pm        60.897  8  3.107e-10 ***
## Cloud9am         9.884  8   0.273268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avec un test multiple on retrouve bien les mêmes sorties qu'avec un summary.
Dans le modèle modAIC, qui a été construit à l'aide de la méthode pas à pas, on retrouve bien les variables significatives au niveau 5% suivantes : WindSpeed3pm, WindSpeed9am, Temp9am, Temp3pm, Humidity9am, WindDir9am, et Cloud3pm. Comme le montre ces résultats, le modèle prédit bien la variable RainTomorrow.

```
Anova(modBIC_bis, type=3, test.statistic = "LR")
```

```
## Analysis of Deviance Table (Type III tests)
##
## Response: RainTomorrow
##                         LR Chisq Df Pr(>Chisq)
## Pressure9am              74.102  1  < 2.2e-16 ***
## WindSpeed3pm              0.462  1  0.4966803
## RainToday                 9.825  1  0.0017217 **
## MinTemp                  11.094  1  0.0008661 ***
## Evaporation               0.772  1  0.3796166
## MaxTemp                   0.929  1  0.3352108
## Humidity9am               0.039  1  0.8433611
## Temp3pm                  11.658  1  0.0006393 ***
## Cloud3pm                 54.711  8  5.024e-09 ***
## Cloud9am                  7.170  8  0.5183972
## WindDir3pm               23.426 15  0.0755069 .
## WindDir9am               24.375 15  0.0589974 .
## RainToday:Humidity9am    13.654  1  0.0002197 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lorsque nous effectuons un test multiple, nous obtenons des résultats similaires pour les variables continues. Cependant, pour les variables qualitatives, nous observons que Cloud3pm devient significative, alors qu'aucune modalité n'était significative dans le test simple. Ainsi, nous avons identifié 6 variables significatives avec un seuil de 5 % : Pressure9am, RainToday, MinTemp, Temp3pm, Cloud3pm et RainToday.

Conclusion

En conclusion, cette analyse des données météorologiques de Melbourne a été réalisée en utilisant des modèles linéaires généralisés. L'objectif principal était de déterminer quelles variables étaient les plus significatives pour prédire s'il allait pleuvoir le lendemain.

2 modèles ont été construits et évalués en se basant sur la méthode de recherche de meilleurs modèles 'pas à pas forward', à l'aide des critères AIC et BIC. Pour le modèle retenu selon le critère AIC, six variables explicatives se sont avérées significatives : WindSpeed3pm, WindSpeed9am, Temp9am, Temp3pm, Humidity9am et Cloud3pm8. Quant au modèle retenu selon le critère BIC, cinq variables explicatives se sont avérées significatives : Pressure9am, RainToday, Temp9am, Humidity9am et Temp3pm.

Des tests d'interactions entre les variables explicatives ont également été effectués, et un nouveau modèle, meilleur que celui construit à l'aide du critère du BIC à été gardé. Ses variables signficatives sont les suivantes : RainToday, MinTemp, Temp3pm, Cloud3pm et RainToday:Humidity9am.

Enfin, on décide donc de retenir ces 2 modèles: modAIC et modBIC_bis du fait qu'ils minimisent chacun le critère BIC ou AIC. Ces 2 modèles permettent de prédire de manière optimale s'il va pleuvoir le lendemain à Melbourne, en se basant sur les données météorologiques disponibles.

Cette démarche démontre l'importance de choisir judicieusement les variables explicatives et de considérer les interactions potentielles pour obtenir un modèle prédictif robuste.