Lab 3 - Survival Analysis

Philippe Real

02 février, 2020

Contents

1 Exercice N°2 - Comparaison des approches analyse de survie et classification

On souhaite prévoir la probabilité de rechute ("recurrent") à 24 mois. Pour cela, on comparerez les méthodes de l'analyse de survie (modèles de Cox, survival random forests, ...) aux méthodes de classification. Les mesures de performances (notamment l'AUC) se feront sur un sous-échantillon de test formé de 20 à 30% des données (attention à bien stratifier !).

1.1 Import des données wpc

```
## Observations: 198
## Variables: 35
## $ id         <dbl> 119513, 8423, 842517, 843483, 843584, 843786, 84435...
## $ recur      <chr> "N", "N", "N", "N", "R", "R", "N", "R", "N", "N", "...
## $ time       <dbl> 31, 61, 116, 123, 27, 77, 60, 77, 119, 76, 123, 125...
## $ V1         <dbl> 18.02, 17.99, 21.37, 11.42, 20.29, 12.75, 18.98, 13...
## $ V2         <dbl> 27.60, 10.38, 17.44, 20.38, 14.34, 15.29, 19.61, 20...
## $ V3         <dbl> 117.50, 122.80, 137.50, 77.58, 135.10, 84.60, 124.4...
## $ V4         <dbl> 1013.0, 1001.0, 1373.0, 386.1, 1297.0, 502.7, 1112....
## $ V5         <dbl> 0.09489, 0.11840, 0.08836, 0.14250, 0.10030, 0.1189...
## $ V6         <dbl> 0.10360, 0.27760, 0.11890, 0.28390, 0.13280, 0.1569...
## $ V7         <dbl> 0.10860, 0.30010, 0.12550, 0.24140, 0.19800, 0.1664...
## $ V8         <dbl> 0.07055, 0.14710, 0.08180, 0.10520, 0.10430, 0.0766...
## $ V9         <dbl> 0.1865, 0.2419, 0.2333, 0.2597, 0.1809, 0.1995, 0.1...
## $ V10        <dbl> 0.06333, 0.07871, 0.06010, 0.09744, 0.05883, 0.0716...
## $ V11        <dbl> 0.6249, 1.0950, 0.5854, 0.4956, 0.7572, 0.3877, 0.5...
## $ V12        <dbl> 1.8900, 0.9053, 0.6105, 1.1560, 0.7813, 0.7402, 0.8...
## $ V13        <dbl> 3.972, 8.589, 3.928, 3.445, 5.438, 2.999, 3.592, 3....
## $ V14        <dbl> 71.55, 153.40, 82.15, 27.23, 94.44, 30.85, 61.21, 5...
## $ V15        <dbl> 0.004433, 0.006399, 0.006167, 0.009110, 0.011490, 0...
## $ V16        <dbl> 0.014210, 0.049040, 0.034490, 0.074580, 0.024610, 0...
## $ V17        <dbl> 0.03233, 0.05373, 0.03300, 0.05661, 0.05688, 0.0456...
## $ V18        <dbl> 0.009854, 0.015870, 0.018050, 0.018670, 0.018850, 0...
## $ V19        <dbl> 0.01694, 0.03003, 0.03094, 0.05963, 0.01756, 0.0177...
## $ V20        <dbl> 0.003495, 0.006193, 0.005039, 0.009208, 0.005115, 0...
## $ V21        <dbl> 21.63, 25.38, 24.90, 14.91, 22.54, 15.51, 23.39, 17...
## $ V22        <dbl> 37.08, 17.33, 20.98, 26.50, 16.67, 20.37, 25.45, 28...
## $ V23        <dbl> 139.70, 184.60, 159.10, 98.87, 152.20, 107.30, 152....
## $ V24        <dbl> 1436.0, 2019.0, 1949.0, 567.7, 1575.0, 733.2, 1593....
## $ V25        <dbl> 0.1195, 0.1622, 0.1188, 0.2098, 0.1374, 0.1706, 0.1...
## $ V26        <dbl> 0.1926, 0.6656, 0.3449, 0.8663, 0.2050, 0.4196, 0.3...
## $ V27        <dbl> 0.3140, 0.7119, 0.3414, 0.6869, 0.4000, 0.5999, 0.2...
## $ V28        <dbl> 0.11700, 0.26540, 0.20320, 0.25750, 0.16250, 0.1709...
## $ V29        <dbl> 0.2677, 0.4601, 0.4334, 0.6638, 0.2364, 0.3485, 0.2...
## $ V30        <dbl> 0.08113, 0.11890, 0.09067, 0.17300, 0.07678, 0.1179...
## $ tumor_size <dbl> 5.0, 3.0, 2.5, 2.0, 3.5, 2.5, 1.5, 4.0, 2.0, 6.0, 2...
## $ lymph      <chr> "5", "2", "0", "0", "0", "0", "?", "10", "1", "20",...
```

1.2 Label pour la tâche de classification.

A partir de la variable `recur` (rechute) variable binaire.

```
wpbc = wpbc %>% mutate(id = factor(id)) %>% mutate( recur = recode_factor(recur , 'N' = FALSE, 'R' = TRU
wpbc$time <- as.numeric(wpbc$time )
wpbc<-filter(wpbc,wpbc$lymph!="?")
wpbc$lymph <- as.numeric(wpbc$lymph )
```

```
data.cox<-wpbc
data.cox$recur<-as.numeric(data.cox$recur)
data.cox$recur<-data.cox$recur-1
head(data.cox)
```

```
## # A tibble: 6 x 35
##   id    recur  time    V1    V2    V3    V4     V5    V6    V7     V8    V9
##   <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1 1195~     0    31  18.0  27.6 118.  1013  0.0949 0.104 0.109 0.0706 0.186
## 2 8423      0    61  18.0  10.4 123.  1001  0.118  0.278 0.300 0.147  0.242
## 3 8425~     0   116  21.4  17.4 138.  1373  0.0884 0.119 0.126 0.0818 0.233
## 4 8434~     0   123  11.4  20.4  77.6  386. 0.142  0.284 0.241 0.105  0.260
## 5 8435~     1    27  20.3  14.3 135.  1297  0.100  0.133 0.198 0.104  0.181
## 6 8437~     1    77  12.8  15.3  84.6  503. 0.119  0.157 0.166 0.0767 0.200
## # ... with 23 more variables: V10 <dbl>, V11 <dbl>, V12 <dbl>, V13 <dbl>,
## #   V14 <dbl>, V15 <dbl>, V16 <dbl>, V17 <dbl>, V18 <dbl>, V19 <dbl>,
## #   V20 <dbl>, V21 <dbl>, V22 <dbl>, V23 <dbl>, V24 <dbl>, V25 <dbl>,
## #   V26 <dbl>, V27 <dbl>, V28 <dbl>, V29 <dbl>, V30 <dbl>,
## #   tumor_size <dbl>, lymph <dbl>
```

1.3 Création des jeux de données train et test

En fixant la racine du générateur aléatoire (fonction R set.seed), on crée un jeu de données de train et un de test. On utilise la fonction *stratified* du package *fifer*.

On doit retouver le même type de distribution en particulier pour les variables importantes dans nos différents jeux de données. En annexe on présente les summary des différents échantillons.

```
set.seed(123)
dataTrain <-stratified(data.cox,c("recur","lymph"),size=0.7)
dataTest <- anti_join(data.cox,dataTrain)
```

- fréquences absolues des classes - éch. d'apprentissage

```
##
##   0   1
## 105  34
```

- fréquences relatives des classes dans l'éch. d'apprentissage

```
##
##         0         1
## 0.7553957 0.2446043
```

- distribution des classes dans l'éch. test

```
##
##         0         1
## 0.7818182 0.2181818
```

1.4 Méthodes d'analyse de survie

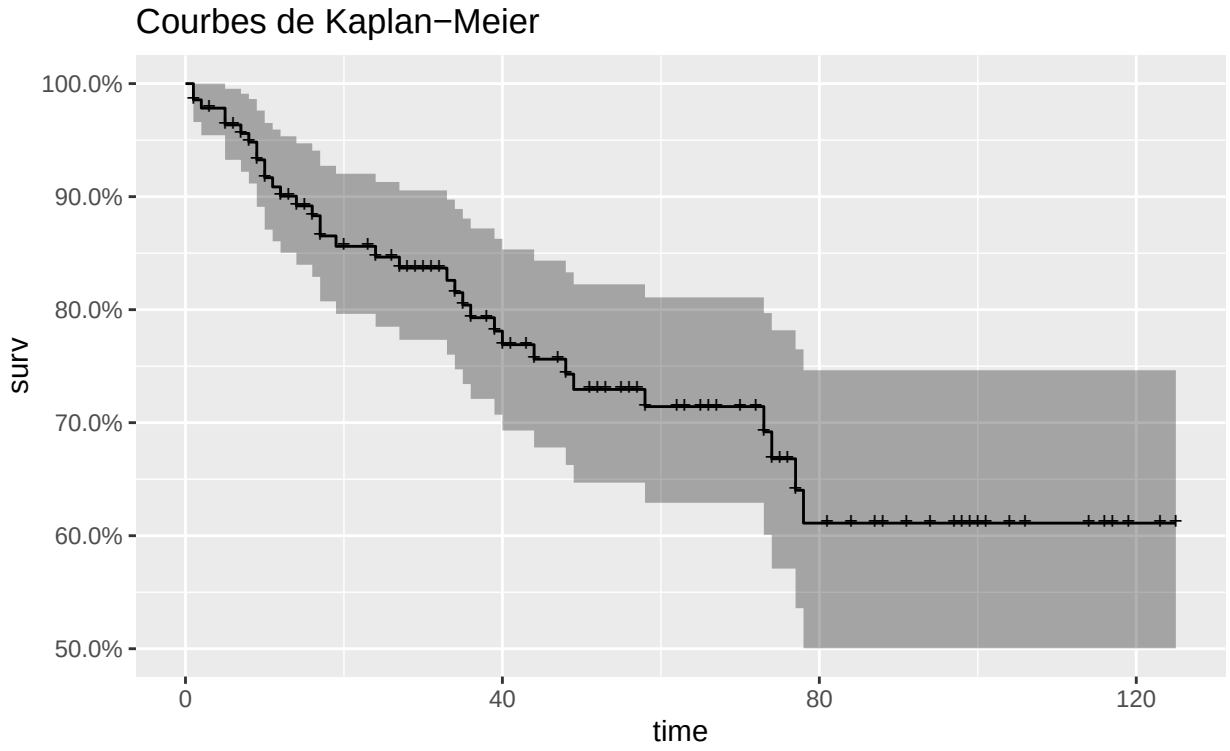
Avant de construire un mopdèle de Cox on commence par regarder la courbe de survie de Kplan-Meyer.

1.4.1 Courbe de survie de Kaplan-Meyer

```
km.fit <- survfit(Surv(time, recur) ~ 1, data=dataTrain)
```

Prévision à 24 mois avec l'estimateur de Kapplan-Meyer

```
## Call: survfit(formula = Surv(time, recur) ~ 1, data = dataTrain)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    24     90      19    0.847  0.0326        0.785        0.913
```



1.4.2 Modèle de Cox complet

```
## Call:
## coxph(formula = f.cox, data = dataTrain)
##
##   n= 139, number of events= 34
##
##                   coef   exp(coef)    se(coef)      z Pr(>|z|)
## tumor_size   8.681e-02   1.091e+00   1.476e-01  0.588  0.55647
## lymph        1.233e-01   1.131e+00   4.425e-02  2.785  0.00535 **
## V1          -8.358e+00   2.346e-04   4.209e+00 -1.986  0.04706 *
## V2          -3.637e-02   9.643e-01   2.013e-01 -0.181  0.85663
## V3           1.252e+00   3.497e+00   6.242e-01  2.006  0.04490 *
## V4           1.249e-03   1.001e+00   9.141e-03  0.137  0.89131
## V5           7.797e+01   7.245e+33   6.820e+01  1.143  0.25295
## V6          -3.160e+01   1.891e-14   3.823e+01 -0.827  0.40848
## V7          -1.811e+01   1.360e-08   2.171e+01 -0.834  0.40411
## V8          -2.228e+01   2.116e-10   3.777e+01 -0.590  0.55532
## V9           6.914e+00   1.006e+03   2.031e+01  0.340  0.73355
## V10         -2.504e+02  1.707e-109   1.704e+02 -1.470  0.14155
## V11          1.740e+01   3.599e+07   1.037e+01  1.677  0.09349 .
## V12         -2.647e+00   7.086e-02   1.681e+00 -1.575  0.11531
## V13          4.292e-01   1.536e+00   1.312e+00  0.327  0.74366
## V14         -1.354e-01   8.734e-01   5.217e-02 -2.595  0.00947 **
## V15         -1.754e+01   2.412e-08   2.946e+02 -0.060  0.95253
## V16          9.820e+01   4.453e+42   7.549e+01  1.301  0.19332
## V17         -9.243e+01   7.247e-41   6.993e+01 -1.322  0.18629
## V18         -1.482e+02   4.249e-65   1.855e+02 -0.799  0.42439
## V19          1.320e+02   2.092e+57   8.735e+01  1.511  0.13080
## V20          4.472e+02  1.687e+194   5.624e+02  0.795  0.42646
## V21          3.121e-01   1.366e+00   1.146e+00  0.272  0.78536
## V22          6.867e-02   1.071e+00   1.898e-01  0.362  0.71746
## V23         -8.384e-03   9.917e-01   1.179e-01 -0.071  0.94330
## V24          1.184e-03   1.001e+00   5.785e-03  0.205  0.83790
## V25          6.159e+01   5.616e+26   3.777e+01  1.631  0.10292
## V26         -1.313e+01   1.992e-06   9.386e+00 -1.398  0.16198
## V27          9.222e+00   1.012e+04   7.542e+00  1.223  0.22138
## V28          1.376e+01   9.504e+05   1.946e+01  0.707  0.47933
## V29         -7.114e+00   8.134e-04   1.303e+01 -0.546  0.58498
## V30         -1.250e+01   3.712e-06   5.911e+01 -0.212  0.83246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef)  lower .95  upper .95
## tumor_size  1.091e+00  9.169e-01  8.167e-01  1.457e+00
## lymph       1.131e+00  8.840e-01  1.037e+00  1.234e+00
## V1          2.346e-04  4.262e+03  6.136e-08  8.971e-01
## V2          9.643e-01  1.037e+00  6.499e-01  1.431e+00
## V3          3.497e+00  2.860e-01  1.029e+00  1.188e+01
## V4          1.001e+00  9.988e-01  9.835e-01  1.019e+00
## V5          7.245e+33  1.380e-34  6.441e-25  8.149e+91
## V6          1.891e-14  5.287e+13  5.452e-47  6.561e+18
## V7          1.360e-08  7.352e+07  4.504e-27  4.107e+10
## V8          2.116e-10  4.726e+09  1.500e-42  2.984e+22
```

```
## V9          1.006e+03  9.940e-04  5.184e-15  1.952e+20
## V10        1.707e-109 5.858e+108 1.644e-254  1.773e+36
## V11         3.599e+07  2.779e-08  5.326e-02  2.432e+16
## V12         7.086e-02  1.411e+01  2.628e-03  1.911e+00
## V13         1.536e+00  6.510e-01  1.173e-01  2.012e+01
## V14         8.734e-01  1.145e+00  7.885e-01  9.674e-01
## V15         2.412e-08  4.146e+07 3.928e-259 1.481e+243
## V16         4.453e+42  2.246e-43  2.450e-22 8.093e+106
## V17         7.247e-41  1.380e+40 2.155e-100  2.437e+19
## V18         4.249e-65  2.354e+64 4.893e-223  3.689e+93
## V19         2.092e+57  4.780e-58  9.248e-18 4.733e+131
## V20        1.687e+194 5.928e-195 3.527e-285        Inf
## V21         1.366e+00  7.319e-01  1.446e-01  1.291e+01
## V22         1.071e+00  9.336e-01  7.384e-01  1.554e+00
## V23         9.917e-01  1.008e+00  7.871e-01  1.249e+00
## V24         1.001e+00  9.988e-01  9.899e-01  1.013e+00
## V25         5.616e+26  1.781e-27  4.003e-06  7.878e+58
## V26         1.992e-06  5.019e+05  2.041e-14  1.945e+02
## V27         1.012e+04  9.881e-05  3.853e-03  2.658e+10
## V28         9.504e+05  1.052e-06  2.598e-11  3.477e+22
## V29         8.134e-04  1.229e+03  6.636e-15  9.970e+07
## V30         3.712e-06  2.694e+05  1.814e-56  7.596e+44
##
## Concordance= 0.81  (se = 0.04 )
## Likelihood ratio test= 50.4  on 32 df,   p=0.02
## Wald test            = 29.56  on 32 df,   p=0.6
## Score (logrank) test = 48.02  on 32 df,   p=0.03
```

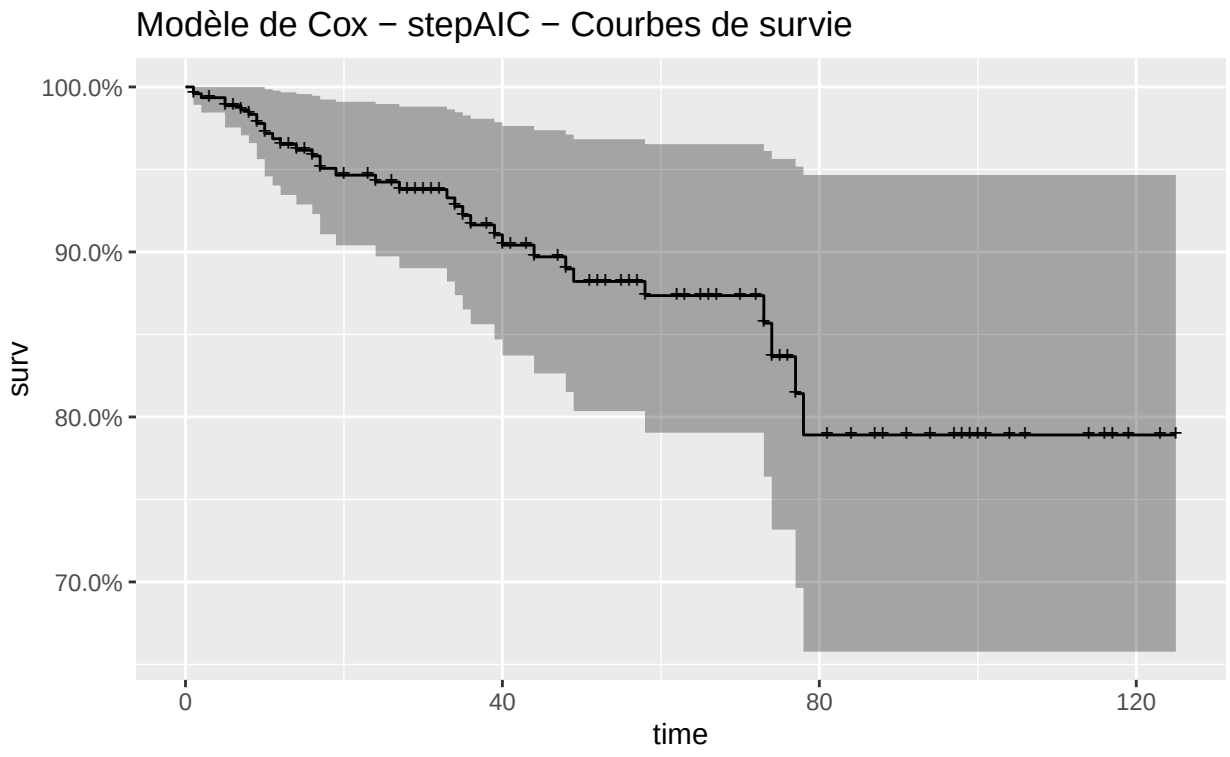
1.4.3 Modèle de Cox et sélection de variables par AIC

On utilise la fonction de R stepAIC pour faire le choix de variables.

```
## Call:
## coxph(formula = Surv(time, recur) ~ lymph + V1 + V3 + V6 + V11 +
##     V12 + V14 + V16 + V17 + V19 + V24 + V25 + V26 + V27, data = dataTrain)
##
##   n= 139, number of events= 34
##
##              coef  exp(coef)   se(coef)      z Pr(>|z|)
## lymph  1.242e-01  1.132e+00  3.420e-02  3.633 0.000280 ***
## V1    -5.770e+00  3.120e-03  2.585e+00 -2.232 0.025620 *
## V3     8.786e-01  2.408e+00  3.867e-01  2.272 0.023079 *
## V6    -5.886e+01  2.749e-26  1.991e+01 -2.956 0.003113 **
## V11    1.679e+01  1.952e+07  4.733e+00  3.547 0.000390 ***
## V12   -1.939e+00  1.438e-01  7.856e-01 -2.469 0.013552 *
## V14   -1.209e-01  8.861e-01  3.639e-02 -3.322 0.000893 ***
## V16    1.448e+02  7.949e+62  4.255e+01  3.404 0.000664 ***
## V17   -9.846e+01  1.729e-43  4.055e+01 -2.428 0.015186 *
## V19    8.341e+01  1.679e+36  3.218e+01  2.592 0.009537 **
## V24    3.574e-03  1.004e+00  1.647e-03  2.170 0.029977 *
## V25    5.906e+01  4.469e+25  1.875e+01  3.150 0.001630 **
## V26   -1.098e+01  1.699e-05  5.167e+00 -2.126 0.033539 *
## V27    6.222e+00  5.037e+02  4.278e+00  1.454 0.145815
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## lymph 1.132e+00  8.832e-01 1.059e+00 1.211e+00
## V1    3.120e-03  3.205e+02 1.966e-05 4.951e-01
## V3    2.408e+00  4.154e-01 1.128e+00 5.137e+00
## V6    2.749e-26  3.637e+25 3.112e-43 2.429e-09
## V11   1.952e+07  5.124e-08 1.828e+03 2.083e+11
## V12   1.438e-01  6.955e+00 3.083e-02 6.704e-01
## V14   8.861e-01  1.128e+00 8.252e-01 9.516e-01
## V16   7.949e+62  1.258e-63 4.825e+26 1.309e+99
## V17   1.729e-43  5.785e+42 5.215e-78 5.730e-09
## V19   1.679e+36  5.956e-37 6.840e+08 4.122e+63
## V24   1.004e+00  9.964e-01 1.000e+00 1.007e+00
## V25   4.469e+25  2.238e-26 4.925e+09 4.055e+41
## V26   1.699e-05  5.885e+04 6.794e-10 4.250e-01
## V27   5.037e+02  1.985e-03 1.150e-01 2.206e+06
##
## Concordance= 0.787  (se = 0.042 )
## Likelihood ratio test= 41.8  on 14 df,   p=1e-04
## Wald test            = 30.91  on 14 df,   p=0.006
## Score (logrank) test = 37.18  on 14 df,   p=7e-04
```

La sélection de variable, semble avoir bien amélioré le modèle. Les variables retenues semblent, plutôt très significatives et les tests meilleurs.



1.4.4 Forêts-aléaoires de survie

```
r.forest
```

```
## Ranger result
##
## Call:
##  ranger(f.cox, data = data.cox, mtry = 7, importance = "permutation",      splitrule = "extratrees",
##
## Type:                             Survival
## Number of trees:                  500
## Sample size:                      194
## Number of independent variables:  32
## Mtry:                             7
## Target node size:                 3
## Variable importance mode:         permutation
## Splitrule:                        extratrees
## Number of unique death times:     94
## Number of random splits:          1
## OOB prediction error (1-C):       0.3876442
```

1.5 Méthode de classification - Modèle de regression logistique

Comme méthode de classification, on va considérer un modèle de regression logistique.

1.5.1 Ajout du critère de décision

Pour utiliser un modèle logit on ajoute une variable de décision binaire, qui correspond à une rechute entre 0 et 24 mois. Cette nouvelle variable est aussi ajoutée au jeu de test.

```
data.logit.Train=data.logit.Train %>% mutate( rechute24 = data.logit.Train$time<25 & data.logit.Train$r
data.logit.Test=data.logit.Test %>% mutate( rechute24 = data.logit.Test$time<25 & data.logit.Test$recur
```

1.5.2 Modèle logit complet ou saturé

On construit tout d'abord le modèle complet. Puis à partir de ce modèle complet et par minimisation du critère AIC on obtiendra le modèle logit final.

```
##
## Call:
## glm(formula = f.glm, family = binomial, data = data.logit.Train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.53940  -0.39476  -0.14071  -0.02504   2.54415
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.27406   22.30290   0.147   0.8833
## tumor_size    0.23377    0.24872   0.940   0.3473
```

```
## lymph          0.13437    0.08007   1.678   0.0933 .
## V1            -8.52550    7.63594  -1.116   0.2642
## V2             0.30751    0.30020   1.024   0.3057
## V3             0.87499    1.08569   0.806   0.4203
## V4             0.01379    0.01617   0.853   0.3936
## V5             1.60539  136.87303   0.012   0.9906
## V6            23.30146   63.27183   0.368   0.7127
## V7           -36.52348   46.99144  -0.777   0.4370
## V8           -12.77134   80.26313  -0.159   0.8736
## V9           -50.33918   37.47788  -1.343   0.1792
## V10         -283.03360  309.43355  -0.915   0.3604
## V11           13.50575   20.39624   0.662   0.5079
## V12           -0.88357    2.73271  -0.323   0.7464
## V13           -1.07510    2.67304  -0.402   0.6875
## V14           -0.05979    0.08408  -0.711   0.4770
## V15          -88.33125  528.92031  -0.167   0.8674
## V16          175.65356  161.96252   1.085   0.2781
## V17          -69.78819  121.27399  -0.575   0.5650
## V18         -383.23631  300.81390  -1.274   0.2027
## V19          229.74863  152.57007   1.506   0.1321
## V20          -68.32664 1153.69259  -0.059   0.9528
## V21            2.19842    2.27617   0.966   0.3341
## V22           -0.31559    0.29876  -1.056   0.2908
## V23            0.08471    0.21138   0.401   0.6886
## V24           -0.01099    0.01178  -0.933   0.3509
## V25          107.00024   66.23903   1.615   0.1062
## V26          -31.44715   19.69413  -1.597   0.1103
## V27           11.96414   13.05235   0.917   0.3593
## V28           47.98071   33.48490   1.433   0.1519
## V29           -4.03644   22.44539  -0.180   0.8573
## V30           20.37737  102.55008   0.199   0.8425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.897  on 138  degrees of freedom
## Residual deviance:  62.177  on 106  degrees of freedom
## AIC: 128.18
##
## Number of Fisher Scoring iterations: 8
```

La variable lymph semble particulièrement significative.

1.5.3 Modèle logit final

On va sélectionner le modèle logit final en choissant le modèle qui minimise le critère AIC. Pour cela on utilisera la fonction R: *step*

```
##
## Call:
## glm(formula = rechute24 ~ lymph + V1 + V9 + V18 + V19 + V21 +
##     V22 + V24 + V28, family = binomial, data = data.logit.Train)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4942  -0.4435  -0.2425  -0.1102   2.6732
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.238e+00  7.215e+00  -0.172   0.8637
## lymph        8.577e-02  5.079e-02   1.689   0.0913 .
## V1          -7.378e-01  3.134e-01  -2.354   0.0186 *
## V9          -6.465e+01  2.270e+01  -2.848   0.0044 **
## V18         -2.215e+02  1.058e+02  -2.094   0.0363 *
## V19          1.297e+02  5.405e+01   2.399   0.0164 *
## V21          1.418e+00  6.828e-01   2.077   0.0378 *
## V22         -1.086e-01  5.796e-02  -1.874   0.0609 .
## V24         -5.418e-03  3.851e-03  -1.407   0.1595
## V28          2.955e+01  1.323e+01   2.234   0.0255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.897  on 138  degrees of freedom
## Residual deviance:  76.917  on 129  degrees of freedom
## AIC: 96.917
##
## Number of Fisher Scoring iterations: 6
```

1.5.4 Comparaison des deux modèles *logit* par un test anova

```
## Analysis of Deviance Table
##
## Model 1: rechute24 ~ 1 + tumor_size + lymph + V1 + V2 + V3 + V4 + V5 +
##     V6 + V7 + V8 + V9 + V10 + V11 + V12 + V13 + V14 + V15 + V16 +
##     V17 + V18 + V19 + V20 + V21 + V22 + V23 + V24 + V25 + V26 +
##     V27 + V28 + V29 + V30
## Model 2: rechute24 ~ lymph + V1 + V9 + V18 + V19 + V21 + V22 + V24 + V28
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1       106     62.177
## 2       129     76.917 -23   -14.74   0.9037
```

Le test accepte la nullité des paramètres du logit complet qui ne sont pas dans le logit obtenu avec la fonction de choix de modèles step. On prévilégiera le modèle step: $m_logit.BwdFwd$.

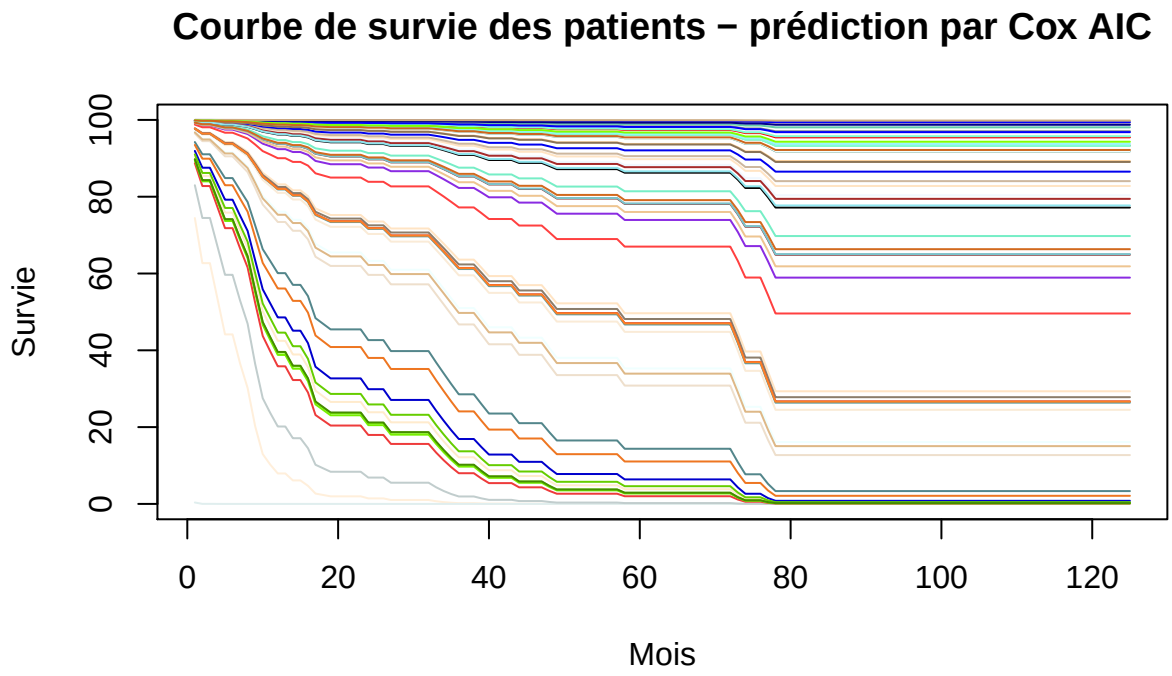
1.6 Prédire dans les 2 modèles les probabilités de rechute à 24 mois.

1.6.1 Modèles de *Cox* prédiction de rechute à 24 mois.

On va tracer la courbe de survie pour les différents patients du jeu de test pour chacun des 2 modèles obtenu. On calculera aussi, simultanément la probabilité de survie à 24 mois pour chacun des patients en utilisant la fonction $survfit$.

- Courbe de survie des différents patients à partir du modèle de Cox obtenu par stepAIC

```
DataPredict<-dataTest
survival.aic.df<-survival(dataTest,coxph.m.AIC,24,"AIC")
```




```
summary(survival.aic.df)
```

```
##       id                 time         n.risk      n.event
##  Length:55          Min.   :24   Min.   :90   Min.   :19
##  Class :character   1st Qu.:24   1st Qu.:90   1st Qu.:19
##  Mode  :character   Median :24   Median :90   Median :19
##                     Mean   :24   Mean   :90   Mean   :19
##                     3rd Qu.:24   3rd Qu.:90   3rd Qu.:19
##                     Max.   :24   Max.   :90   Max.   :19
##     survival         relapse             std.err          lowerCI_95
##  Min.   :0.0000   Min.   :0.0004245   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.6275   1st Qu.:0.0172813   1st Qu.:0.01352   1st Qu.:0.3039
##  Median :0.9137   Median :0.0862998   Median :0.04963   Median :0.8214
##  Mean   :0.7505   Mean   :0.2494948   Mean   :0.09757   Mean   :0.6330
##  3rd Qu.:0.9827   3rd Qu.:0.3724813   3rd Qu.:0.16328   3rd Qu.:0.9548
##  Max.   :0.9996   Max.   :1.0000000   Max.   :0.41318   Max.   :0.9977
##    uperCI_95
##  Min.   :0.9820
##  1st Qu.:1.0000
##  Median :1.0000
##  Mean   :0.9997
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

- Probabilité de rechute à 24 mois pour les modèles de Cox

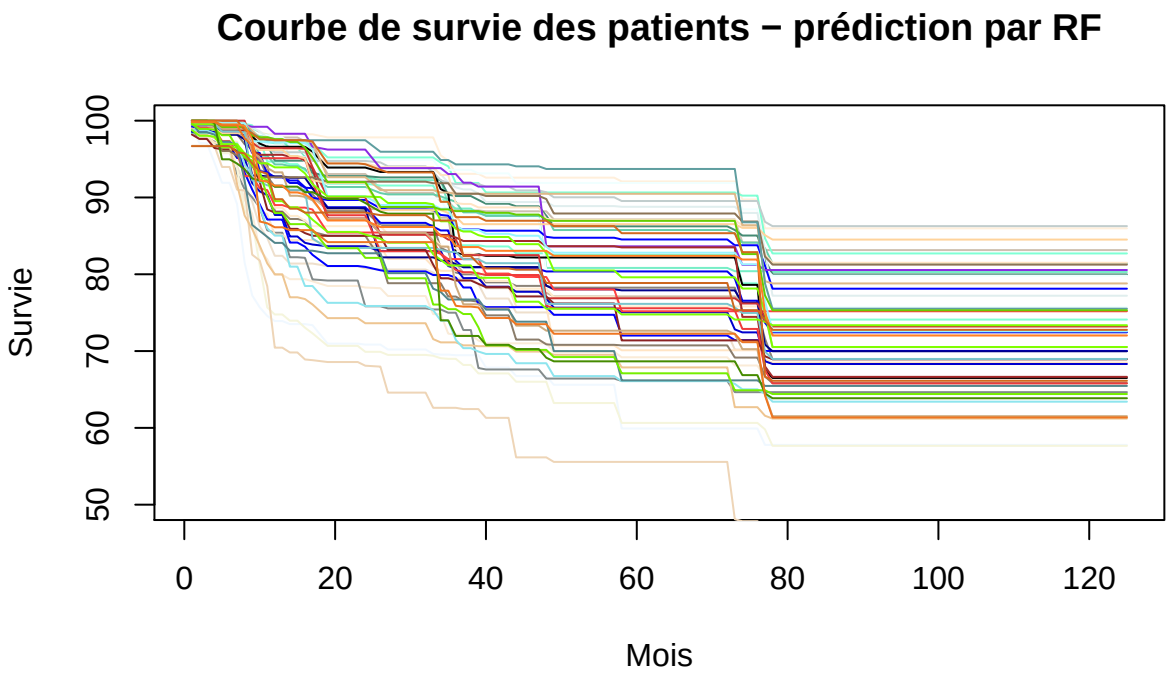
La probabilité de rechute (moyenne) à 24 mois peut aussi être obtenue globalement en utilisant : $survfit$ sur tout l'échantillon.

```
cox.m.comp.fit <-survfit(coxph.m.comp)
cox.m.aic.fit <-survfit(coxph.m.AIC)
```

Modèle	Proba-rechute %	Proba-survie %
Cox - complet	4.919	95.081
Cox - stepAIC	6.218	93.782

1.6.2 Survival Random-Forest prédiction de rechute à 24 mois.

- Courbe de survie des différents patients à partir du modèle Random-Forest



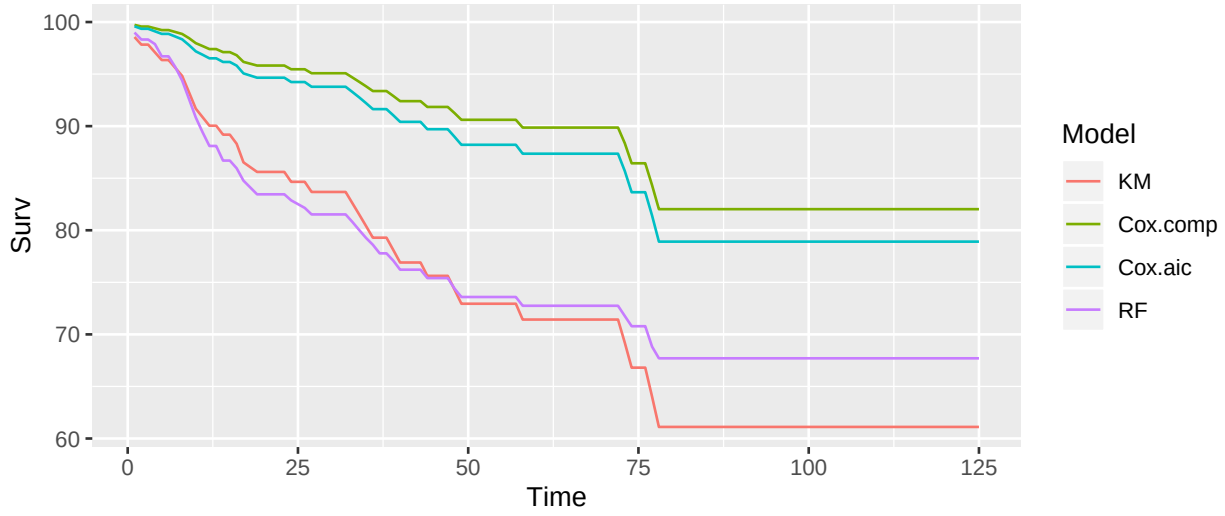
On donne aussi la probabilité de rechute à 24 mois pour la random forest:

```
proba.mean<- round(100*(1-mean(sapply(1:dim(dataTrain)[1], function(n) r.forest$survival[n,][21]))),2)
surv.mean <-round(100-proba.mean,2)
```

Modèle	Proba-rechute %	Proba-survie %
Random forest	16.88	83.12

1.6.3 Comparaison des courbes de survie moyennes des différents modèles

On trace les courbes de survie des modèles étudiés : Kapplan Meyer, Cox et RF. On remarque que les 2 modèles de Cox sont très proches, largement au dessus de KM et RF.



1.6.4 Modèles *logit* prédiction de rechute à 24 mois.

On utilise la fonction predict de R.

```
prev_m_logit.cmp <- predict(m_logit.cmp, newdata = data.logit.Test, type = "response" )
prev_m_logit.aic <- predict(m_logit.BwdFwd,newdata = data.logit.Test, type = "response" )
```

On stock dans un tableau les probabilités de rechute/survie pour chaque individus obtenues à partir des différents modèles

```
##   r.forest coxph.cmp coxph.aic logit.cmp logit.aic
## 1    0.194     0.007     0.012     0.046     0.806
## 2    0.292     0.004     0.053     0.000     0.027
## 3    0.217     0.019     0.297     0.416     0.487
```

- Probabilité de rechute pour le logit - complet

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## 0.0000004 0.0023800 0.0396087 0.2183830 0.3458826 0.9991659
```

- Probabilité de rechute pour le logit - aic

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## 0.0003221 0.0216097 0.0822378 0.2143556 0.3491117 0.8671045
```

Modèle	Proba-rechute %	Proba-survie %
logit-complet	21.84	78.16
lgit-stepAIC	21.44	78.56

1.7 Comparaison des modèles en termes de précision (accuracy) et d'AUC.

1.7.1 Pourcentage de réussite des modèles par rapport à l'observation

- Estimation au seuil de 0.5

On confronte les probabilités obtenues aux seuil de 0.5. Dés que la prévision dépasse 50% on prédit qu'il y a rechute.

```
pred_0.5 <- apply(pred_proba >=0.5, 2, factor)#,labels=c("no","yes"))
result<-cbind(pred_0.5,data.logit.Test$rechute24)
```

- tableau des % de réussite de prédiction pour les différents modèles

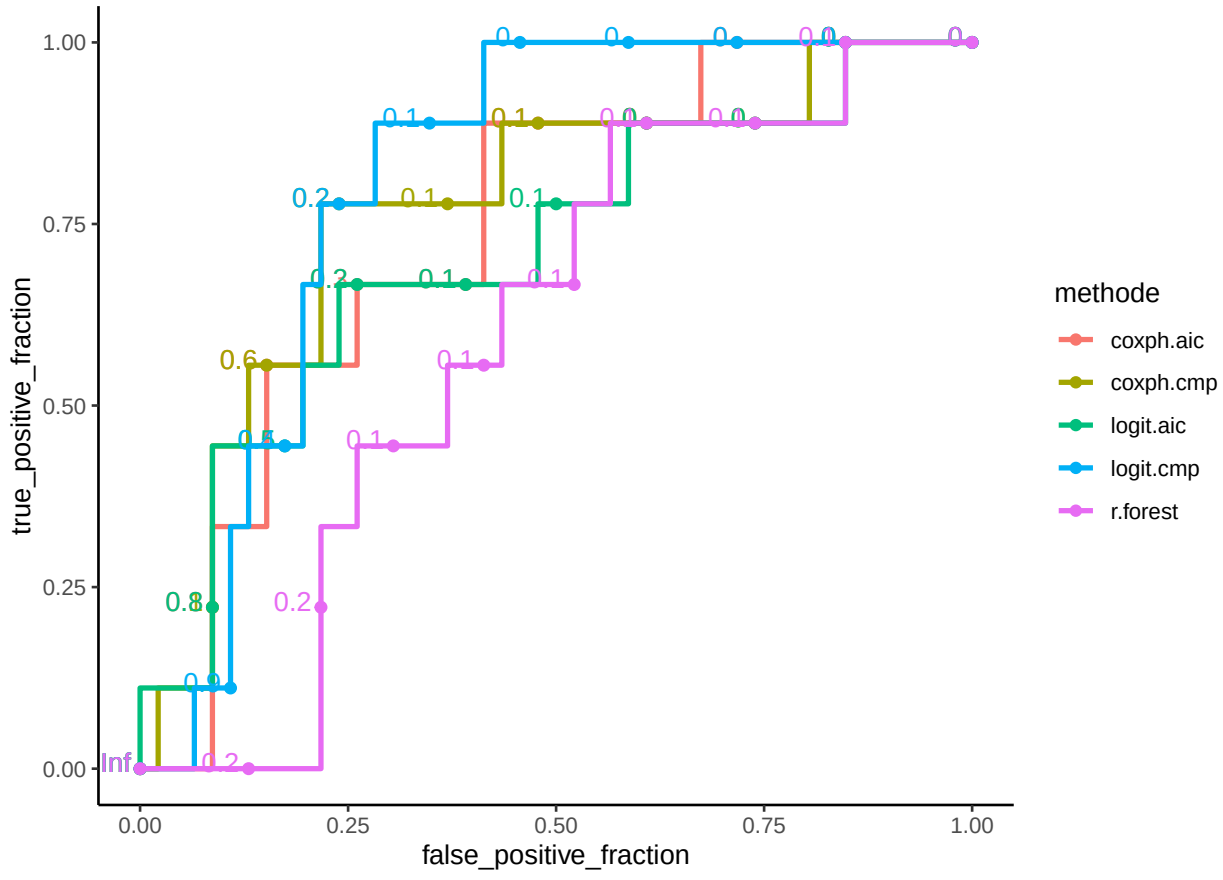
```
##   r.forest coxph.cmp coxph.aic logit.cmp logit.aic Observe
## 1  83.6364        80        80   78.1818   83.6364     100
```

- tableau des % d'erreur de prédiction pour les différents modèles

```
##   r.forest coxph.cmp coxph.aic logit.cmp logit.aic Observe
## 1  16.3636        20        20   21.8182   16.3636       0
```

1.7.2 AUC des différents modèles

1.7.3 Courbes ROC et AUC des différents modèles



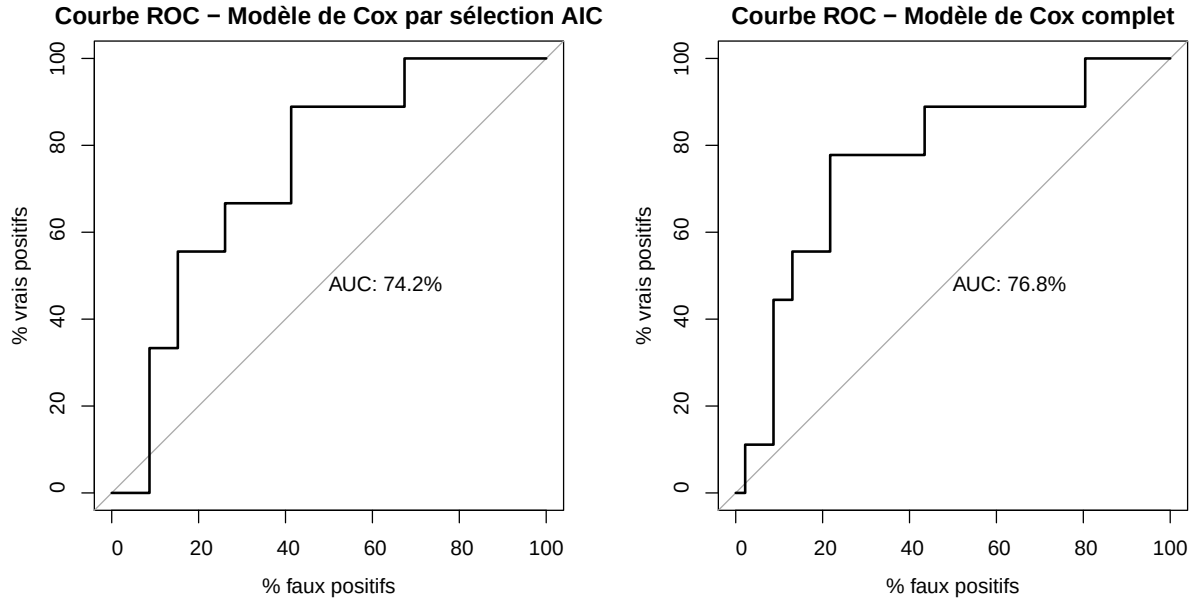
On trace les courbes ROC des différents modèles sur un même graphique. Le meilleur estimateur aura une aire sous la courbe le plus proche possible de 1. La courbe idéale serait perpendiculaire aux abscisse du point origine jusqu'au point (0.1) puis prallèle jusqu'au point (1,1). Si bien que l'aire sous cette courbe serait égale à 1.

```
df_roc.all %>% group_by(methode) %>% summarize(AUC=auc(obs,score))
```

```
## # A tibble: 5 x 2
##   methode     AUC
##   <chr>     <dbl>
## 1 coxph.aic 0.742
## 2 coxph.cmp 0.768
## 3 logit.aic 0.710
## 4 logit.cmp 0.809
## 5 r.forest  0.594
```

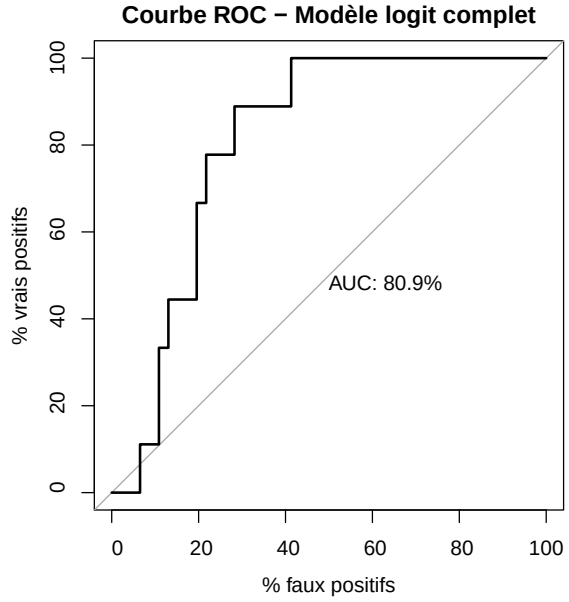
on confirme ces résultats en utilisant une autre méthode, la méthode roc du package R *pROC*

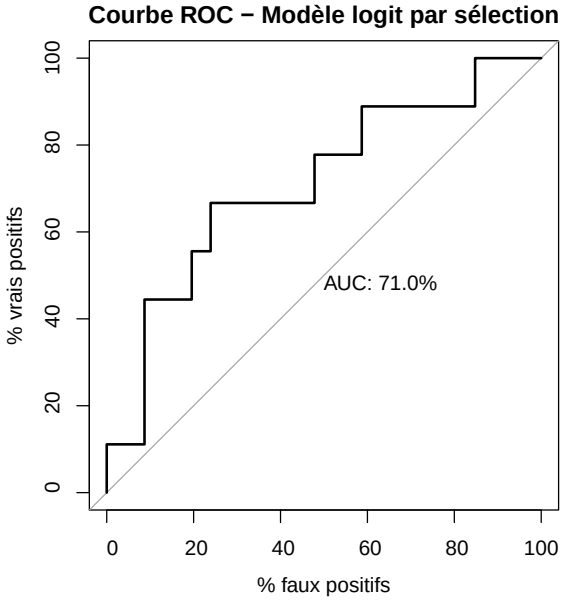
```
##
## Call:
## roc.default(response = data.logit.Test$rechute24, predictor = pred_proba$coxph.aic,     percent = T,
##
## Data: pred_proba$coxph.aic in 46 controls (data.logit.Test$rechute24 FALSE) < 9 cases (data.logit.Tes
## Area under the curve: 74.15%
```




```
##
## Call:
## roc.default(response = data.logit.Test$rechute24, predictor = pred_proba$coxph.cmp,     percent = T,
##
## Data: pred_proba$coxph.cmp in 46 controls (data.logit.Test$rechute24 FALSE) < 9 cases (data.logit.Tes
## Area under the curve: 76.81%
```

```
##
## Call:
## roc.default(response = data.logit.Test$rechute24, predictor = pred_proba$logit.cmp,     percent = T,
##
## Data: pred_proba$logit.cmp in 46 controls (data.logit.Test$rechute24 FALSE) < 9 cases (data.logit.Tes
## Area under the curve: 80.92%
```






```
##
## Call:
## roc.default(response = data.logit.Test$rechute24, predictor = pred_proba$logit.aic,     percent = T,
##
## Data: pred_proba$logit.aic in 46 controls (data.logit.Test$rechute24 FALSE) < 9 cases (data.logit.Tes
## Area under the curve: 71.01%
```

1.8 Conclusion

En terme d'accruancy (erreur de prévision par rapport aux données de test) les modèles sont relativement similaires. Le fait est que l'on n'est pas parvenu à obtenir des résultats stables au niveau des critères AUC et d'accruancy. Probablement dû à la manière dont les jeux de données Train et Test ont été construits. Il est très possible que ce soit la conséquence d'une mauvaise stratification. Et donc le choix de modèle est rendu difficile par l'instabilité des résultats. Le seul résultat stable se situe au niveau de la prédiction de survie qui est nettement supérieure dans le cas des modèles de Cox d'environ (10% à 15%) par rapport au modèle logit.

1.9 Annexes

1.9.1 comparaison des jeux de données complet, train et test

- Donnnées complètes

```
summary(data.cox)
```

```
##        id          recur             time              V1
##  8423   :  1   Min.   :0.0000   Min.   :  1.00   Min.   :10.95
##  85715  :  1   1st Qu.:0.0000   1st Qu.: 14.25   1st Qu.:15.05
##  86208  :  1   Median :0.0000   Median : 39.50   Median :17.29
##  86517  :  1   Mean   :0.2371   Mean   : 46.94   Mean   :17.40
##  87112  :  1   3rd Qu.:0.0000   3rd Qu.: 73.00   3rd Qu.:19.58
##  87163  :  1   Max.   :1.0000   Max.   :125.00   Max.   :27.22
##  (Other):188
##        V2              V3               V4               V5
##  Min.   :10.38   Min.   : 71.90   Min.   : 361.6   Min.   :0.07497
##  1st Qu.:19.34   1st Qu.: 98.16   1st Qu.: 702.5   1st Qu.:0.09390
##  Median :21.80   Median :113.70   Median : 929.1   Median :0.10220
##  Mean   :22.30   Mean   :114.78   Mean   : 969.1   Mean   :0.10277
##  3rd Qu.:24.78   3rd Qu.:129.65   3rd Qu.:1193.5   3rd Qu.:0.11138
##  Max.   :39.28   Max.   :182.10   Max.   :2250.0   Max.   :0.14470
##
##        V6               V7               V8               V9
##  Min.   :0.04605   Min.   :0.02398   Min.   :0.02031   Min.   :0.1308
##  1st Qu.:0.10985   1st Qu.:0.10608   1st Qu.:0.06376   1st Qu.:0.1741
##  Median :0.13175   Median :0.15205   Median :0.08607   Median :0.1893
##  Mean   :0.14264   Mean   :0.15631   Mean   :0.08681   Mean   :0.1929
##  3rd Qu.:0.17220   3rd Qu.:0.20050   3rd Qu.:0.10393   3rd Qu.:0.2095
##  Max.   :0.31140   Max.   :0.42680   Max.   :0.20120   Max.   :0.3040
##
##       V10              V11              V12              V13
##  Min.   :0.05025   Min.   :0.1938   Min.   :0.3621   Min.   : 1.153
##  1st Qu.:0.05672   1st Qu.:0.3882   1st Qu.:0.9245   1st Qu.: 2.743
##  Median :0.06171   Median :0.5407   Median :1.1820   Median : 3.782
##  Mean   :0.06274   Mean   :0.6040   Mean   :1.2737   Mean   : 4.259
##  3rd Qu.:0.06681   3rd Qu.:0.7509   3rd Qu.:1.4688   3rd Qu.: 5.213
##  Max.   :0.09744   Max.   :1.8190   Max.   :3.5030   Max.   :13.280
##
##       V14              V15               V16               V17
##  Min.   : 13.99   Min.   :0.002667   Min.   :0.007347   Min.   :0.01094
##  1st Qu.: 35.37   1st Qu.:0.005016   1st Qu.:0.019803   1st Qu.:0.02687
##  Median : 58.45   Median :0.006209   Median :0.027975   Median :0.03691
##  Mean   : 70.29   Mean   :0.006796   Mean   :0.031328   Mean   :0.04093
##  3rd Qu.: 92.48   3rd Qu.:0.007991   3rd Qu.:0.038532   3rd Qu.:0.04897
##  Max.   :316.00   Max.   :0.031130   Max.   :0.135400   Max.   :0.14380
##
##       V18               V19               V20               V21
##  Min.   :0.005174   Min.   :0.007882   Min.   :0.001087   Min.   :12.84
##  1st Qu.:0.011423   1st Qu.:0.014807   1st Qu.:0.002753   1st Qu.:17.59
##  Median :0.014280   Median :0.017945   Median :0.003719   Median :20.52
##  Mean   :0.015151   Mean   :0.020609   Mean   :0.004004   Mean   :20.99
```

```
##  3rd Qu.:0.017680   3rd Qu.:0.022880   3rd Qu.:0.004636   3rd Qu.:23.73
##  Max.   :0.039270   Max.   :0.060410   Max.   :0.012560   Max.   :35.13
##
##       V22              V23              V24              V25
##  Min.   :16.67   Min.   : 85.1   Min.   : 508.1   Min.   :0.08191
##  1st Qu.:26.21   1st Qu.:117.9   1st Qu.: 940.6   1st Qu.:0.12932
##  Median :30.30   Median :136.5   Median :1295.0   Median :0.14175
##  Mean   :30.18   Mean   :140.1   Mean   :1401.8   Mean   :0.14392
##  3rd Qu.:33.62   3rd Qu.:159.9   3rd Qu.:1694.2   3rd Qu.:0.15445
##  Max.   :49.54   Max.   :232.2   Max.   :3903.0   Max.   :0.22260
##
##       V26               V27               V28               V29
##  Min.   :0.05131   Min.   :0.02398   Min.   :0.02899   Min.   :0.1565
##  1st Qu.:0.24755   1st Qu.:0.32215   1st Qu.:0.15222   1st Qu.:0.2759
##  Median :0.35045   Median :0.40115   Median :0.17850   Median :0.3103
##  Mean   :0.36457   Mean   :0.43601   Mean   :0.17845   Mean   :0.3223
##  3rd Qu.:0.42368   3rd Qu.:0.55017   3rd Qu.:0.20713   3rd Qu.:0.3585
##  Max.   :1.05800   Max.   :1.17000   Max.   :0.29030   Max.   :0.6638
##
##       V30            tumor_size         lymph
##  Min.   :0.05504   Min.   : 0.400   Min.   : 0.000
##  1st Qu.:0.07637   1st Qu.: 1.500   1st Qu.: 0.000
##  Median :0.08654   Median : 2.500   Median : 1.000
##  Mean   :0.09078   Mean   : 2.868   Mean   : 3.211
##  3rd Qu.:0.10178   3rd Qu.: 3.500   3rd Qu.: 4.000
##  Max.   :0.20750   Max.   :10.000   Max.   :27.000
##
```

- Données Train

```
summary(dataTrain)
```

```
##        id          recur             time              V1
##  86208  :  1   Min.   :0.0000   Min.   :  1.00   Min.   :10.95
##  87112  :  1   1st Qu.:0.0000   1st Qu.: 14.00   1st Qu.:15.05
##  87163  :  1   Median :0.0000   Median : 38.00   Median :17.29
##  87164  :  1   Mean   :0.2446   Mean   : 44.15   Mean   :17.35
##  87880  :  1   3rd Qu.:0.0000   3rd Qu.: 68.50   3rd Qu.:19.55
##  89122  :  1   Max.   :1.0000   Max.   :125.00   Max.   :27.22
##  (Other):133
##        V2              V3               V4               V5
##  Min.   :11.89   Min.   : 71.90   Min.   : 361.6   Min.   :0.07497
##  1st Qu.:19.64   1st Qu.: 98.71   1st Qu.: 703.1   1st Qu.:0.09378
##  Median :22.02   Median :113.40   Median : 929.4   Median :0.10150
##  Mean   :22.69   Mean   :114.33   Mean   : 963.1   Mean   :0.10253
##  3rd Qu.:25.16   3rd Qu.:129.50   3rd Qu.:1192.0   3rd Qu.:0.11125
##  Max.   :39.28   Max.   :182.10   Max.   :2250.0   Max.   :0.14470
##
##        V6                V7                V8                V9
##  Min.   :0.05131   Min.   :0.02398   Min.   :0.02307   Min.   :0.1424
##  1st Qu.:0.10700   1st Qu.:0.10037   1st Qu.:0.06124   1st Qu.:0.1742
##  Median :0.13280   Median :0.14910   Median :0.08520   Median :0.1893
##  Mean   :0.14025   Mean   :0.15424   Mean   :0.08540   Mean   :0.1922
```

```
##  3rd Qu.:0.17175   3rd Qu.:0.20405   3rd Qu.:0.10125   3rd Qu.:0.2092
##  Max.   :0.28670   Max.   :0.42680   Max.   :0.20120   Max.   :0.2678
##
##       V10               V11               V12               V13
##  Min.   :0.05025   Min.   :0.1938   Min.   :0.3621   Min.   : 1.153
##  1st Qu.:0.05679   1st Qu.:0.3902   1st Qu.:0.9853   1st Qu.: 2.696
##  Median :0.06149   Median :0.5299   Median :1.1930   Median : 3.705
##  Mean   :0.06249   Mean   :0.5989   Mean   :1.2962   Mean   : 4.196
##  3rd Qu.:0.06685   3rd Qu.:0.7378   3rd Qu.:1.4935   3rd Qu.: 5.134
##  Max.   :0.08243   Max.   :1.8190   Max.   :3.1200   Max.   :13.280
##
##       V14               V15                V16                V17
##  Min.   : 13.99   Min.   :0.002667   Min.   :0.007347   Min.   :0.01094
##  1st Qu.: 35.89   1st Qu.:0.004973   1st Qu.:0.020115   1st Qu.:0.02806
##  Median : 54.16   Median :0.005960   Median :0.027220   Median :0.03582
##  Mean   : 69.58   Mean   :0.006632   Mean   :0.030681   Mean   :0.04061
##  3rd Qu.: 93.27   3rd Qu.:0.007985   3rd Qu.:0.038570   3rd Qu.:0.04947
##  Max.   :253.80   Max.   :0.023330   Max.   :0.100600   Max.   :0.12780
##
##       V18                V19                V20                V21
##  Min.   :0.005297   Min.   :0.007882   Min.   :0.001087   Min.   :12.84
##  1st Qu.:0.011095   1st Qu.:0.015250   1st Qu.:0.002736   1st Qu.:17.55
##  Median :0.013920   Median :0.018780   Median :0.003742   Median :20.38
##  Mean   :0.014945   Mean   :0.020697   Mean   :0.003990   Mean   :20.85
##  3rd Qu.:0.018110   3rd Qu.:0.022695   3rd Qu.:0.004698   3rd Qu.:23.51
##  Max.   :0.036600   Max.   :0.060410   Max.   :0.011300   Max.   :33.12
##
##       V22              V23              V24               V25
##  Min.   :16.67   Min.   : 85.1   Min.   : 508.1   Min.   :0.08191
##  1st Qu.:26.39   1st Qu.:117.7   1st Qu.: 941.5   1st Qu.:0.12735
##  Median :30.36   Median :134.9   Median :1272.0   Median :0.14080
##  Mean   :30.66   Mean   :139.0   Mean   :1382.4   Mean   :0.14283
##  3rd Qu.:33.98   3rd Qu.:157.6   3rd Qu.:1658.0   3rd Qu.:0.15350
##  Max.   :49.54   Max.   :220.8   Max.   :3432.0   Max.   :0.22260
##
##       V26               V27               V28               V29
##  Min.   :0.05131   Min.   :0.02398   Min.   :0.02899   Min.   :0.1565
##  1st Qu.:0.24880   1st Qu.:0.32120   1st Qu.:0.14895   1st Qu.:0.2798
##  Median :0.34980   Median :0.39950   Median :0.17320   Median :0.3074
##  Mean   :0.35845   Mean   :0.43261   Mean   :0.17429   Mean   :0.3216
##  3rd Qu.:0.42205   3rd Qu.:0.53670   3rd Qu.:0.20410   3rd Qu.:0.3561
##  Max.   :1.05800   Max.   :1.17000   Max.   :0.27330   Max.   :0.5774
##
##       V30            tumor_size         lymph
##  Min.   :0.05504   Min.   : 0.400   Min.   : 0.000
##  1st Qu.:0.07711   1st Qu.: 1.500   1st Qu.: 0.000
##  Median :0.08465   Median : 2.500   Median : 1.000
##  Mean   :0.09018   Mean   : 2.817   Mean   : 3.281
##  3rd Qu.:0.10120   3rd Qu.: 3.500   3rd Qu.: 4.000
##  Max.   :0.20750   Max.   :10.000   Max.   :27.000
##
```

- Données Test

```r
summary(dataTest)
```

```
##        id         recur             time              V1
##  8423   : 1   Min.   :0.0000   Min.   :  4.00   Min.   :11.42
##  85715  : 1   1st Qu.:0.0000   1st Qu.: 16.00   1st Qu.:15.04
##  86517  : 1   Median :0.0000   Median : 56.00   Median :17.29
##  88607  : 1   Mean   :0.2182   Mean   : 53.98   Mean   :17.54
##  91485  : 1   3rd Qu.:0.0000   3rd Qu.: 80.00   3rd Qu.:19.70
##  842517 : 1   Max.   :1.0000   Max.   :123.00   Max.   :25.73
##  (Other):49
##        V2              V3               V4               V5
##  Min.   :10.38   Min.   : 77.58   Min.   : 386.1   Min.   :0.07840
##  1st Qu.:18.70   1st Qu.: 97.13   1st Qu.: 698.8   1st Qu.:0.09581
##  Median :21.00   Median :114.40   Median : 928.8   Median :0.10350
##  Mean   :21.31   Mean   :115.93   Mean   : 984.2   Mean   :0.10340
##  3rd Qu.:23.53   3rd Qu.:130.20   3rd Qu.:1218.0   3rd Qu.:0.11140
##  Max.   :30.98   Max.   :174.20   Max.   :2010.0   Max.   :0.14250
##
##        V6                V7               V8                V9
##  Min.   :0.04605   Min.   :0.0311   Min.   :0.02031   Min.   :0.1308
##  1st Qu.:0.11325   1st Qu.:0.1181   1st Qu.:0.06607   1st Qu.:0.1749
##  Median :0.13040   Median :0.1527   Median :0.08665   Median :0.1894
##  Mean   :0.14869   Mean   :0.1615   Mean   :0.09037   Mean   :0.1945
##  3rd Qu.:0.18190   3rd Qu.:0.1962   3rd Qu.:0.10960   3rd Qu.:0.2118
##  Max.   :0.31140   Max.   :0.3579   Max.   :0.19130   Max.   :0.3040
##
##       V10               V11              V12              V13
##  Min.   :0.05175   Min.   :0.2130   Min.   :0.5914   Min.   : 1.534
##  1st Qu.:0.05654   1st Qu.:0.3792   1st Qu.:0.8650   1st Qu.: 2.781
##  Median :0.06216   Median :0.5692   Median :1.0730   Median : 3.854
##  Mean   :0.06339   Mean   :0.6170   Mean   :1.2169   Mean   : 4.419
##  3rd Qu.:0.06655   3rd Qu.:0.8115   3rd Qu.:1.4425   3rd Qu.: 5.766
##  Max.   :0.09744   Max.   :1.7300   Max.   :3.5030   Max.   :11.560
##
##       V14              V15               V16               V17
##  Min.   : 18.52   Min.   :0.002826   Min.   :0.009105   Min.   :0.01311
##  1st Qu.: 32.67   1st Qu.:0.005327   1st Qu.:0.019405   1st Qu.:0.02681
##  Median : 58.53   Median :0.006455   Median :0.028630   Median :0.03863
##  Mean   : 72.10   Mean   :0.007212   Mean   :0.032962   Mean   :0.04172
##  3rd Qu.: 90.70   3rd Qu.:0.007987   3rd Qu.:0.037625   3rd Qu.:0.04691
##  Max.   :316.00   Max.   :0.031130   Max.   :0.135400   Max.   :0.14380
##
##       V18                V19               V20               V21
##  Min.   :0.005174   Min.   :0.01013   Min.   :0.001286   Min.   :14.91
##  1st Qu.:0.012470   1st Qu.:0.01417   1st Qu.:0.002804   1st Qu.:17.77
##  Median :0.014790   Median :0.01717   Median :0.003643   Median :20.99
##  Mean   :0.015674   Mean   :0.02039   Mean   :0.004041   Mean   :21.34
##  3rd Qu.:0.017290   3rd Qu.:0.02321   3rd Qu.:0.004511   3rd Qu.:23.98
##  Max.   :0.039270   Max.   :0.05963   Max.   :0.012560   Max.   :35.13
##
##       V22              V23               V24              V25
##  Min.   :17.33   Min.   : 96.75   Min.   : 567.7   Min.   :0.1084
##  1st Qu.:25.09   1st Qu.:120.90   1st Qu.: 951.4   1st Qu.:0.1356
```

```
##  Median :30.25   Median :139.80   Median :1349.0   Median :0.1482
##  Mean   :28.97   Mean   :142.98   Mean   :1450.6   Mean   :0.1467
##  3rd Qu.:32.29   3rd Qu.:161.20   3rd Qu.:1750.0   3rd Qu.:0.1550
##  Max.   :41.05   Max.   :232.20   Max.   :3903.0   Max.   :0.2098
##
##       V26               V27               V28               V29
##  Min.   :0.09866   Min.   :0.1547   Min.   :0.06575   Min.   :0.1603
##  1st Qu.:0.23905   1st Qu.:0.3286   1st Qu.:0.16765   1st Qu.:0.2686
##  Median :0.35150   Median :0.4251   Median :0.18570   Median :0.3138
##  Mean   :0.38002   Mean   :0.4446   Mean   :0.18896   Mean   :0.3240
##  3rd Qu.:0.44695   3rd Qu.:0.5646   3rd Qu.:0.21125   3rd Qu.:0.3599
##  Max.   :0.93270   Max.   :0.8488   Max.   :0.29030   Max.   :0.6638
##
##       V30            tumor_size        lymph
##  Min.   :0.06091   Min.   :0.400   Min.   : 0.000
##  1st Qu.:0.07527   1st Qu.:1.500   1st Qu.: 0.000
##  Median :0.09067   Median :2.500   Median : 1.000
##  Mean   :0.09227   Mean   :2.996   Mean   : 3.036
##  3rd Qu.:0.10245   3rd Qu.:4.000   3rd Qu.: 3.500
##  Max.   :0.17300   Max.   :9.000   Max.   :20.000
##
```