TD3 régression logistique (correction)

Marie Fourcot

15/03/2022

```
library(dplyr)
library("ggplot2")
library(ROCR)
library("plotROC")
```

Chargement des données :

```
load("prema.RData")
```

Pour rappel : Dans le cadre d'une étude sur les facteurs prénataux liés à un accouchement prématuré chez les femmes déjà en travail prématuré, on dispose de 13 variables explicatives sur 388 femmes incluses dans l'étude.

La variable à expliquer (PREMATURE) est l'accouchement prématuré.
L'objectif est de définir les facteurs prédictifs d'un accouchement prématuré (Y). Pour chaque modèle considéré, on notera π la probabilité d'un accouchement prématuré sachant les variables $X_1, \cdots, X_p$ incluses.
Les données contiennent les variables suivantes :

Var	Description	Commentaire
GEST	l'âge gestationnel à l'entrée dans l'étude	en semaine
DILATE	la dilatation du col utérin	en cm
EFFACE	l'effacement du col	en %
CONSIS	la consistance du col	1 : mou
		2 : moyen
		3 : ferme
CONTR	la présence de contractions	1 : oui
		2 : non
MEMBRAN	état des membranes	1 : rupturées
		2 : non rupturées
		3 : incertain
AGE	l'âge de la mère	en années
STRAT	période de la grossesse	1-4
GRAVID	la gestité	nombre de grossesses antérieures, y compris celle en cours

Var	Description	Commentaire
PARIT	la parité	nombre de grossesses à terme antérieures
DIAB	diabète	1 : présence
		2 : absence
TRANSF	le transfert vers un hôpital en soins spécialisés	1 : oui
		2 : non
GEMEL	type de grossesse	1 : simple
		2 : multiple

Pour remplir cet objectif, nous avons d'abord construit deux modèles :

- un premier modèle avec comme variable explicative une variable binaire, la variable GEMEL
- un deuxième, avec comme variable explicative une variable quantitative, la variable EFFACE.

Puis nous avons construit un modèle complet que nous avons affiné et évalué.

Nous allons maintenant construire des modèles affinés et permettre une meilleure évaluation de ceux-ci.

27. Imputer les données manquantes de la variable `DIAB` par la réponse majoritaire.

```
ind.na <- which(is.na(prema), arr.ind = TRUE)[,1]
summary(prema$DIAB)
```

```
## Oui Non
##  11 377
```

```
prema[ind.na,'DIAB'] <- 'Non'
```

28. Séparer le jeu de données en jeu d'apprentissage et jeu de test. Avec 70% des données pour l'apprentissage et 30% pour le test. Pour cela, utiliser les fonctions `slice_sample` et `anti_join` du package `dplyr` .

```
prema$ind <- c(1:nrow(prema))
train <- slice_sample(prema, prop = 0.7)
test <- anti_join(prema, train, by="ind")

train <- train[,-15]
test <- test[,-15]
```

La sélection des données pour le jeu d'apprentissage et de test étant aléatoire, deux exécutions ne conduiront pas à la meme sélection, et par la suite pas nécessairement au meme modèle.

29. Estimer le modèle complet des données d'apprentissage.
    Évaluer la significativité de chaque coefficient de ce modèle.

```
logit.fit <- glm(PREMATURE ~ ., family = "binomial", data = train)
```

```
summary(logit.fit)
```

```
##
## Call:
## glm(formula = PREMATURE ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4468  -0.6866   0.2034   0.6300   2.1679
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.703508   4.258216   0.165 0.868777
## GEST              0.255043   0.167748   1.520 0.128412
## DILATE            0.400574   0.195302   2.051 0.040262 *
## EFFACE            0.023105   0.006544   3.531 0.000415 ***
## CONSISMoyen      -0.254772   0.559148  -0.456 0.648647
## CONSISFerme      -0.240644   0.574289  -0.419 0.675195
## CONTRNon         -0.469996   0.756614  -0.621 0.534479
## MEMBRANNon       -3.311242   0.742806  -4.458 8.28e-06 ***
## MEMBRANIncertain -2.489791   1.147931  -2.169 0.030087 *
## AGE               0.021418   0.034990   0.612 0.540457
## STRAT2           -1.107444   1.302578  -0.850 0.395217
## STRAT3           -2.915751   1.593547  -1.830 0.067291 .
## STRAT4           -4.607962   2.077730  -2.218 0.026569 *
## GRAVID            0.151381   0.192636   0.786 0.431961
## PARIT            -0.704941   0.263448  -2.676 0.007455 **
## DIABNon          -2.441442   1.206375  -2.024 0.042992 *
## TRANSFNon        -1.019989   0.377803  -2.700 0.006938 **
## GEMELMultiple     1.357908   0.694376   1.956 0.050515 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 340.18  on 270  degrees of freedom
## Residual deviance: 220.23  on 253  degrees of freedom
## AIC: 256.23
##
## Number of Fisher Scoring iterations: 6
```

30. Réaliser la sélection automatique de variables dans le modèle.

```
logit.reduced <- step(logit.fit)
```

```
summary(logit.reduced)
```

```
##
## Call:
## glm(formula = PREMATURE ~ GEST + DILATE + EFFACE + MEMBRAN +
##     STRAT + PARIT + DIAB + TRANSF + GEMEL, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3638  -0.6655   0.2051   0.6552   2.1034
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.457062   4.117236   0.354 0.723419
## GEST              0.242978   0.166638   1.458 0.144809
## DILATE            0.367077   0.188745   1.945 0.051796 .
## EFFACE            0.023259   0.006103   3.811 0.000138 ***
## MEMBRANNon       -3.234524   0.715471  -4.521 6.16e-06 ***
## MEMBRANIncertain -2.459688   1.129795  -2.177 0.029472 *
## STRAT2           -1.009766   1.252397  -0.806 0.420089
## STRAT3           -2.807809   1.578668  -1.779 0.075306 .
## STRAT4           -4.431584   2.064345  -2.147 0.031815 *
## PARIT            -0.518018   0.170967  -3.030 0.002446 **
## DIABNon          -2.522621   1.182623  -2.133 0.032919 *
## TRANSFNon        -0.928378   0.359807  -2.580 0.009874 **
## GEMELMultiple     1.299303   0.689551   1.884 0.059528 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 340.18  on 270  degrees of freedom
## Residual deviance: 222.01  on 258  degrees of freedom
## AIC: 248.01
##
## Number of Fisher Scoring iterations: 6
```

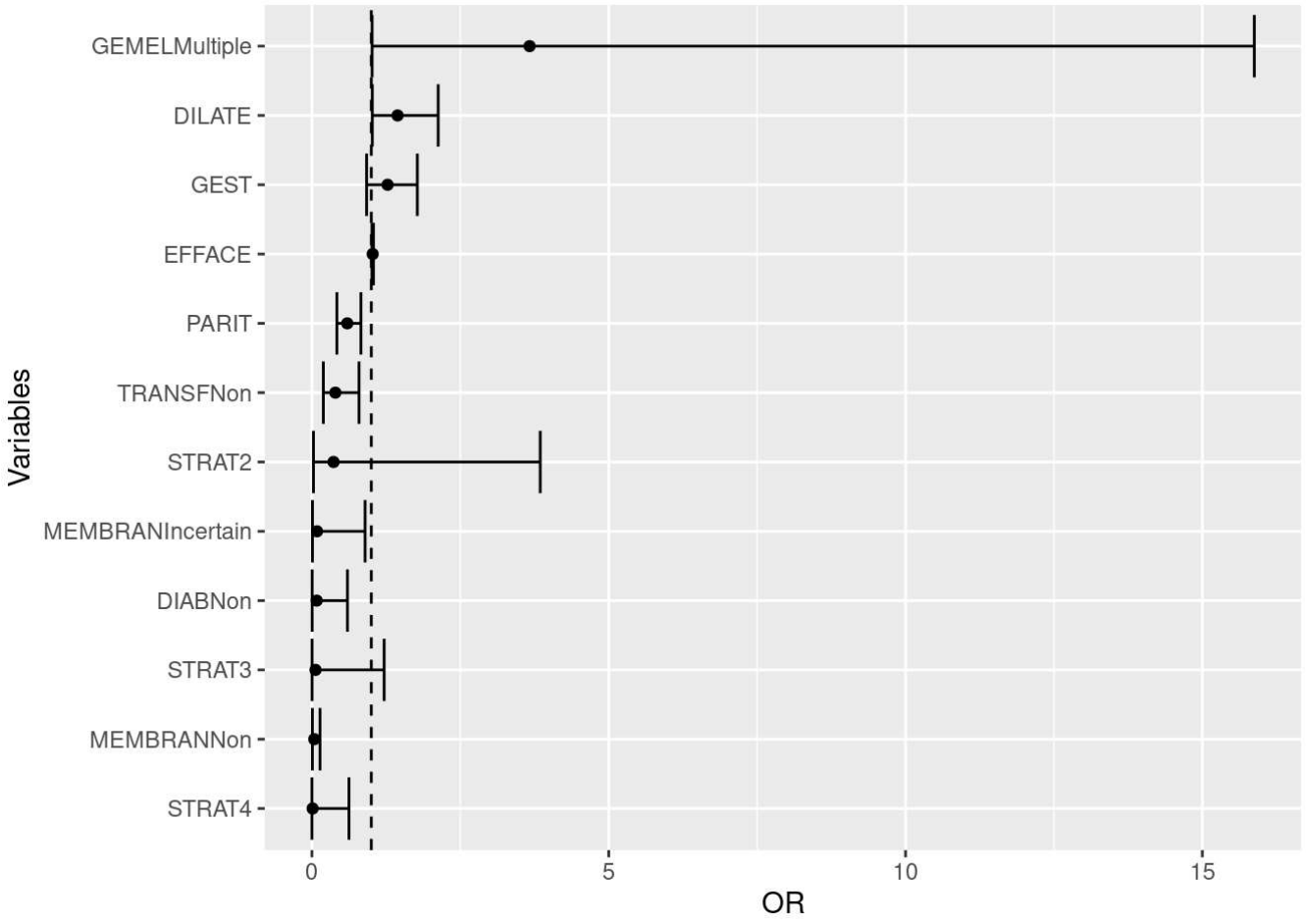
31. Calculer et interpréter les odds ratio significatifs.

```
odds <- data.frame(cbind(exp(coef(logit.reduced)), exp(confint(logit.reduced))))
names(odds) <-c('OR', 'lower', 'upper')
knitr::kable(odds)
```

	OR	lower	upper
(Intercept)	4.2933257	0.0014089	1.564794e+04
GEST	1.2750409	0.9212934	1.775569e+00
DILATE	1.4435086	1.0168982	2.127008e+00
EFFACE	1.0235319	1.0116983	1.036313e+00
MEMBRANNon	0.0393790	0.0076569	1.361606e-01

	OR	lower	upper
MEMBRANIncertain	0.0854616	0.0089943	8.958228e-01
STRAT2	0.3643043	0.0258535	3.846039e+00
STRAT3	0.0603370	0.0023344	1.216511e+00
STRAT4	0.0118956	0.0001810	6.244394e-01
PARIT	0.5957001	0.4205238	8.258108e-01
DIABNon	0.0802490	0.0037971	5.987331e-01
TRANSFNon	0.3951942	0.1925157	7.937150e-01
GEMELMultiple	3.6667393	1.0138213	1.587555e+01

```
odds$vars<-row.names(odds)
odds <- odds[-1,]
ggplot(odds, aes(y= OR, x = reorder(vars, OR))) +
  geom_point() +
  geom_errorbar(aes(ymin=lower, ymax=upper)) +
  geom_hline(yintercept = 1, linetype = 2) +
  coord_flip() +
  labs(x = 'Variables', y = 'OR')
```



Un odds ratio est significatif s'il ne contient pas la valeur 1. C'est un facteur protecteur s'il est inférieur à 1 et un facteur de risque s'il est supérieur à 1.
L'odds ratio de `GEMEL` au niveau multiple est supérieur à 1, son intervalle de confiance ne comprend pas la valeur 1, donc il est significatif. Une grossesse multiple multiplie par 3,67 [1,01;15,88] le risque d'accouchement prématuré.
L'odds ratio associé à la variable `DILATE` est significatif mais difficilement interprétable car c'est une mesure en cm de la dilatation du col utérin. Les variables `EFFACE` et `GEST` ont un odd-ratio qui n'est pas significatif, il comprend 1. De plus, l'odds ration de ces deux variables quantitatives serait compliqué à interpréter.

Pour la variable `PARIT` , l'odd ratio une grossesse antérieure à terme multiplie le risque d'accouchement prématuré par 0,6 [0,42;0,83]. C'est un facteur protecteur, compréhensible bien qu'il s'agisse d'une variable quantitative.
Que la patiente ne soit pas diabétique multiplie le risque d'accouchement prématuré par 0,08 [0,004;0,6]. L'interprétation de l'odds ratio dans ce sens est un peu étrange, il serait plus logique de mettre l'absence de diabète comme référence pour interpréter un facteur de risque en cas de diabète.
La non-rupture des membranes multiple le risque d'accouchement prémature par 0,04 [0,01;0,14] par rapport à des membranes rupturées. Il serait aussi plus logique de considérer la non-rupture des membranes comme valeur de référence car c'est l'état normal.

32. Changer la valeur de référence pour les variables DIAB et MEMBRAN à l'aide de la fonction `relevel` .

```
train$DIAB <- relevel(train$DIAB, 'Non')
train$MEMBRAN <- relevel(train$MEMBRAN, 'Non')
logit.reduced2 <- step(glm(PREMATURE ~ ., family = "binomial", data = train), trace =
0)
summary(logit.reduced2)
```

```
##
## Call:
## glm(formula = PREMATURE ~ GEST + DILATE + EFFACE + MEMBRAN +
##     STRAT + PARIT + DIAB + TRANSF + GEMEL, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3638  -0.6655   0.2051   0.6552   2.1034
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.300083   3.834316  -1.121 0.262086
## GEST              0.242978   0.166638   1.458 0.144809
## DILATE            0.367077   0.188745   1.945 0.051796 .
## EFFACE            0.023259   0.006103   3.811 0.000138 ***
## MEMBRANOui        3.234524   0.715471   4.521 6.16e-06 ***
## MEMBRANIncertain  0.774835   0.922669   0.840 0.401034
## STRAT2           -1.009766   1.252397  -0.806 0.420089
## STRAT3           -2.807809   1.578668  -1.779 0.075306 .
## STRAT4           -4.431584   2.064345  -2.147 0.031815 *
## PARIT            -0.518018   0.170967  -3.030 0.002446 **
## DIABOui           2.522621   1.182623   2.133 0.032919 *
## TRANSFNon        -0.928378   0.359807  -2.580 0.009874 **
## GEMELMultiple     1.299303   0.689551   1.884 0.059528 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 340.18  on 270  degrees of freedom
## Residual deviance: 222.01  on 258  degrees of freedom
## AIC: 248.01
##
## Number of Fisher Scoring iterations: 6
```

```
odds <- data.frame(cbind(exp(coef(logit.reduced2)), exp(confint(logit.reduced2))))
```

```
## Waiting for profiling to be done...
```

```
names(odds) <-c('OR', 'lower', 'upper')
knitr::kable(odds)
```

	OR	lower	upper
(Intercept)	0.0135674	0.0000068	24.5973275
GEST	1.2750409	0.9212934	1.7755685
DILATE	1.4435086	1.0168982	2.1270079

	OR	lower	upper
EFFACE	1.0235319	1.0116983	1.0363128
MEMBRANOui	25.3942748	7.3442669	130.6019889
MEMBRANIncertain	2.1702350	0.4054333	17.5588573
STRAT2	0.3643043	0.0258535	3.8460385
STRAT3	0.0603370	0.0023344	1.2165112
STRAT4	0.0118956	0.0001810	0.6244394
PARIT	0.5957001	0.4205238	0.8258108
DIABOui	12.4612139	1.6701932	263.3623036
TRANSFNon	0.3951942	0.1925157	0.7937150
GEMELMultiple	3.6667393	1.0138213	15.8755524

La rupture des membranes est associée à une multiplication du risque d'accouchement prématuré par 25,39 [7,34;130,6]. Pour calculer cet odds ratio, nous aurions pu aussi inverser l'odds ratio précédemment obtenu.
Le diabète multiplie par 12,46 [1,67;263,36] le risque d'accouchement prématuré.

33. Prédire l'accouchement prématuré sur le jeu de test.

```
S <- predict(logit.reduced2, newdata = test, type="response")

seuil <- 0.5
Y_hat=as.factor(ifelse(S >= seuil, "positif", "negatif"))

MatConf=table(Y_reel = test$PREMATURE,Y_predit = Y_hat)
MatConf
```

```
##          Y_predit
## Y_reel    negatif positif
##   negatif      21      15
##   positif      21      60
```

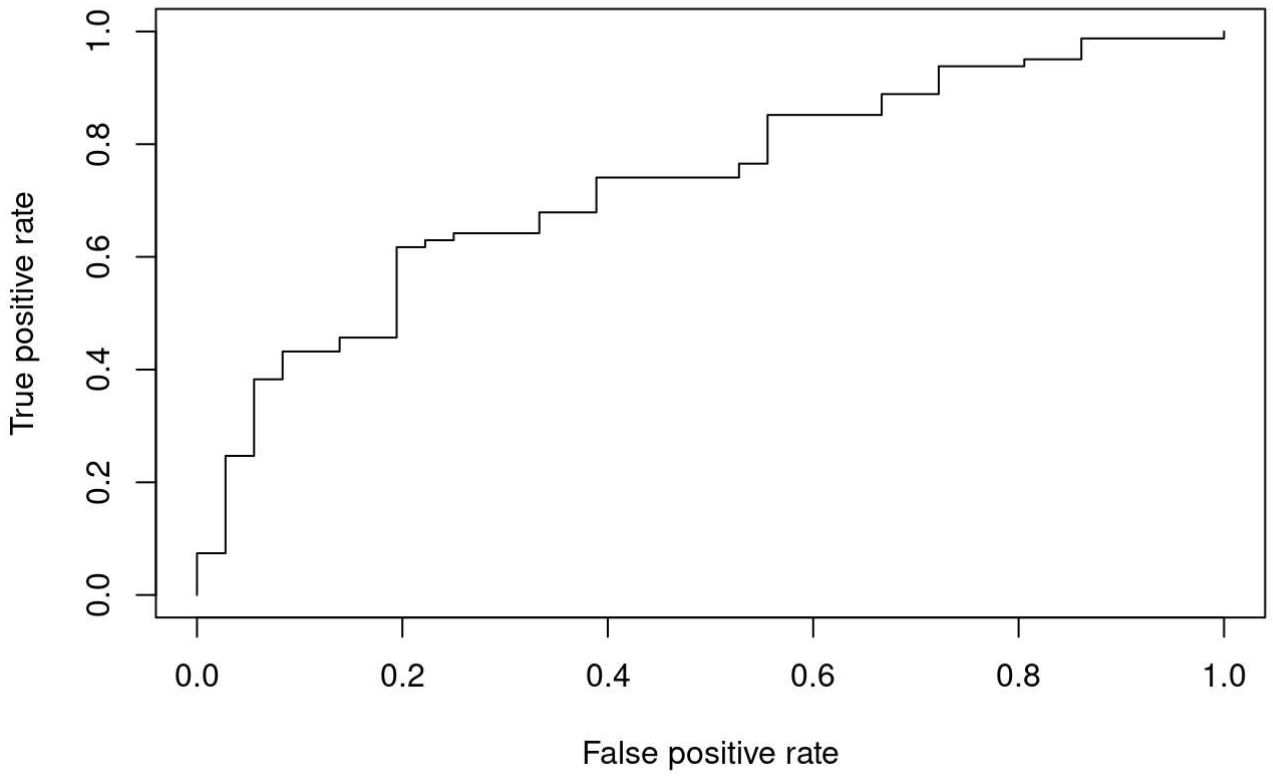
```
PBC = mean(Y_hat == test$PREMATURE, na.rm = TRUE)
PBC
```

```
## [1] 0.6923077
```

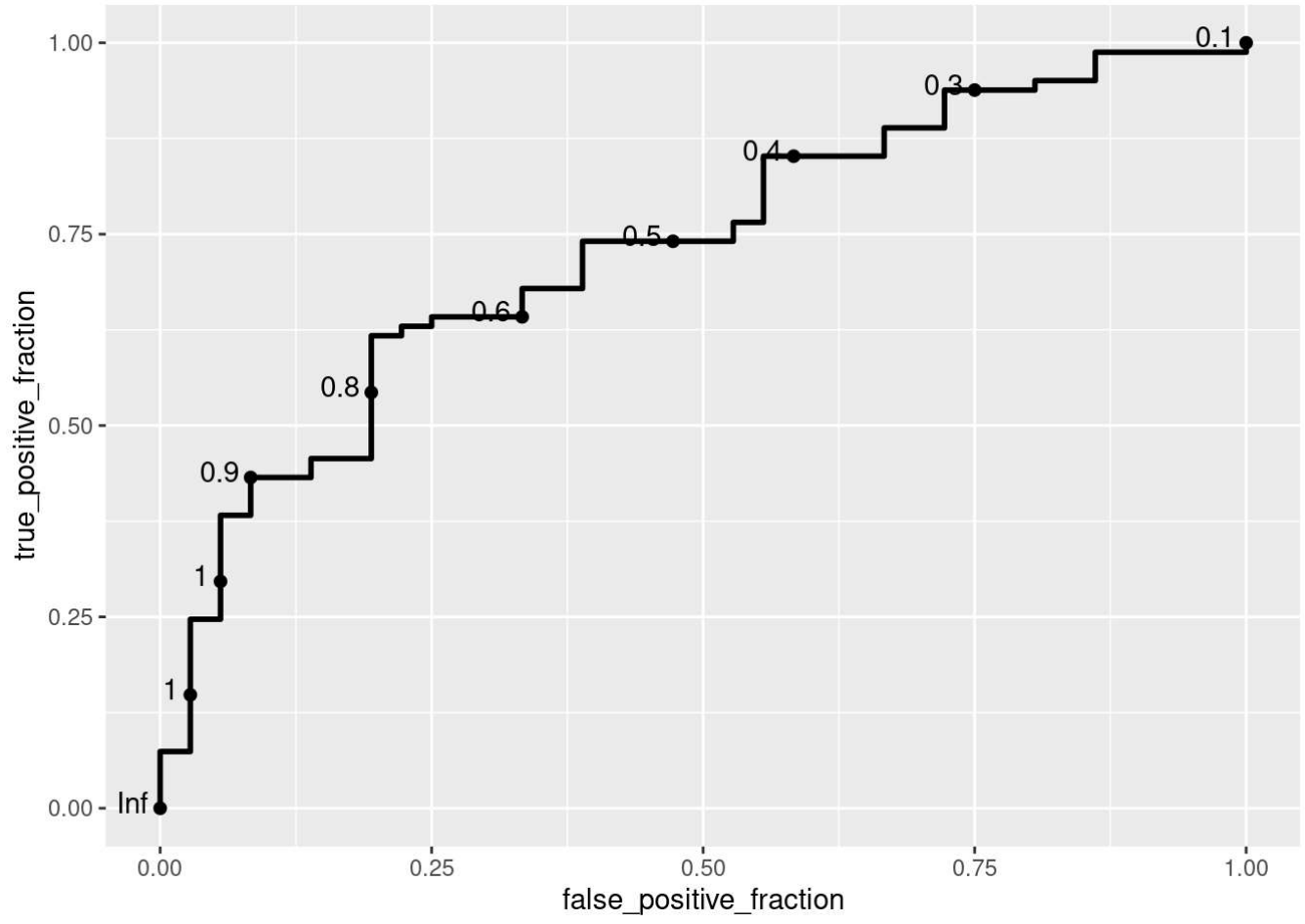
Avec ce modèle, nous avons 69% de bien classés sur le jeu de test.

34. Tracer la courbe ROC.

```
pred=prediction(S, test$PREMATURE)
perf <- performance(pred, "tpr", "fpr")
plot(perf)
```




```
Score = cbind.data.frame(Y = ifelse(test$PREMATURE == "positif",1,0),S = S)
ggplot(Score, aes(d = Y, m = S)) + geom_roc()
```



35. Calculer l'aire sous la courbe ROC.

```
AUC = performance(pred, "auc")
attr(AUC, "y.values")[[1]]
```

```
## [1] 0.7311385
```

L'AUC obtenue de 0,73 est correcte, puisque comme vu en cours, une AUC entre 0,7 et 0,8 est acceptable.