TD régression logistique (correction)

Marie Fourcot

03/2021

```
library(ggplot2)
#devtools::install_github("haleyjeppson/ggmosaic")
library(ggmosaic)
```

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

Var	Description	Commentaire
GRAVID	la gestité	nombre de grossesses antérieures, y compris celle en cours
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

1 Étude des données prématurés

1. Charger le jeu de données dans un tableau prema, obtenir le résumé et vérifier que les variables qualitatives nominales sont bien des facteurs (nécessaire pour la régression logistique). Au besoin, utiliser la commande `as.factor()`.

Chargement des données :

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

```
## 'data.frame':    388 obs. of  14 variables:
##  $ GEST     : int  31 28 31 27 28 33 32 30 33 28 ...
##  $ DILATE   : int  3 8 3 2 6 2 4 1 0 0 ...
##  $ EFFACE   : int  100 0 100 75 75 100 75 50 25 0 ...
##  $ CONSIS   : Factor w/ 3 levels "Mou","Moyen",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ CONTR    : Factor w/ 2 levels "Oui","Non": 1 1 2 2 2 1 1 1 1 1 ...
##  $ MEMBRAN  : Factor w/ 3 levels "Oui","Non","Incertain": 2 2 2 2 2 1 2 2 2 2
...
##  $ AGE      : int  26 25 28 27 17 25 25 29 22 25 ...
##  $ STRAT    : Factor w/ 4 levels "1","2","3","4": 3 3 3 2 3 4 4 3 4 3 ...
##  $ GRAVID   : int  1 1 2 2 1 2 2 2 1 3 ...
##  $ PARIT    : int  0 0 0 1 0 0 1 1 0 1 ...
##  $ DIAB     : Factor w/ 2 levels "Oui","Non": 2 2 2 2 2 2 2 2 2 2 ...
##  $ TRANSF   : Factor w/ 2 levels "Oui","Non": 2 1 1 1 1 1 1 1 2 2 ...
##  $ GEMEL    : Factor w/ 2 levels "Simple","Multiple": 1 2 1 2 1 1 1 1 2 1 ...
##  $ PREMATURE: Factor w/ 2 levels "negatif","positif": 2 2 2 2 2 2 2 2 1 2 ...
```

Résumé des données :

```
summary(prema)
```

```
##       GEST           DILATE          EFFACE         CONSIS    CONTR
##  Min.   :20.00   Min.   :0.000   Min.   :  0.00   Mou  : 54   Oui:355
##  1st Qu.:28.00   1st Qu.:0.000   1st Qu.:  0.00   Moyen:126   Non: 33
##  Median :31.00   Median :1.000   Median : 50.00   Ferme:208
##  Mean   :30.32   Mean   :1.242   Mean   : 43.95
##  3rd Qu.:33.00   3rd Qu.:2.000   3rd Qu.: 75.00
##  Max.   :35.00   Max.   :8.000   Max.   :100.00
##       MEMBRAN         AGE        STRAT       GRAVID           PARIT
##  Oui      : 91   Min.   :15.00   1: 14   Min.   : 0.000   Min.   :0.0000
##  Non      :283   1st Qu.:23.00   2: 52   1st Qu.: 1.000   1st Qu.:0.0000
##  Incertain: 14   Median :26.00   3:153   Median : 2.000   Median :1.0000
##                  Mean   :26.35   4:169   Mean   : 2.302   Mean   :0.7809
##                  3rd Qu.:30.00           3rd Qu.: 3.000   3rd Qu.:1.0000
##                  Max.   :42.00           Max.   :13.000   Max.   :7.0000
##    DIAB     TRANSF         GEMEL       PREMATURE
##  Oui : 11   Oui:187   Simple  :349   negatif:123
##  Non :374   Non:201   Multiple: 39   positif:265
##  NA's:  3
##
##
##
```

La variable DIAB comporte 3 valeurs manquantes, cela peut poser des problèmes par la suite dans les modèles comportant cette variable. En effet, si on lance une régression logistique prenant en compte cette variable, R supprimera (sans vous en informer …) les trois individus avec des données manquantes pour réaliser la régression logistique, en effet toutes les données doivent êtes complètes pour utiliser la régression logistique. Diverses solutions existent pour gérer ce problème des données manquantes :

- Supprimer les individus avec des données manquantes
- Supprimer la variable si elle comporte trop de valeurs manquantes
- Imputer les valeurs manquantes : c'est-à-dire leur affecter arbitrairement une valeur (moyenne, mode, espérance conditionnelle)

2 Étude d'une variable binaire

2. Construire le tableau de contingence PREMATURE/GEMEL.

```
knitr::kable(table(prema$GEMEL,prema$PREMATURE))
```

	negatif	positif
Simple	119	230
Multiple	4	35

3. Calculer la probabilité d'accoucher prématurément lors d'une grossesse multiple.

Probabilité d'accouchement prématuré lors d'une grossesse multiple se lit dans le tableau précédent.

```
35/(35+4)
```

```
## [1] 0.8974359
```

On peut aussi faire un tableau des profils ligne (distribution de PREMATURE sachant GEMEL)

```
knitr::kable(prop.table(table(prema$GEMEL,prema$PREMATURE),margin=1))
```

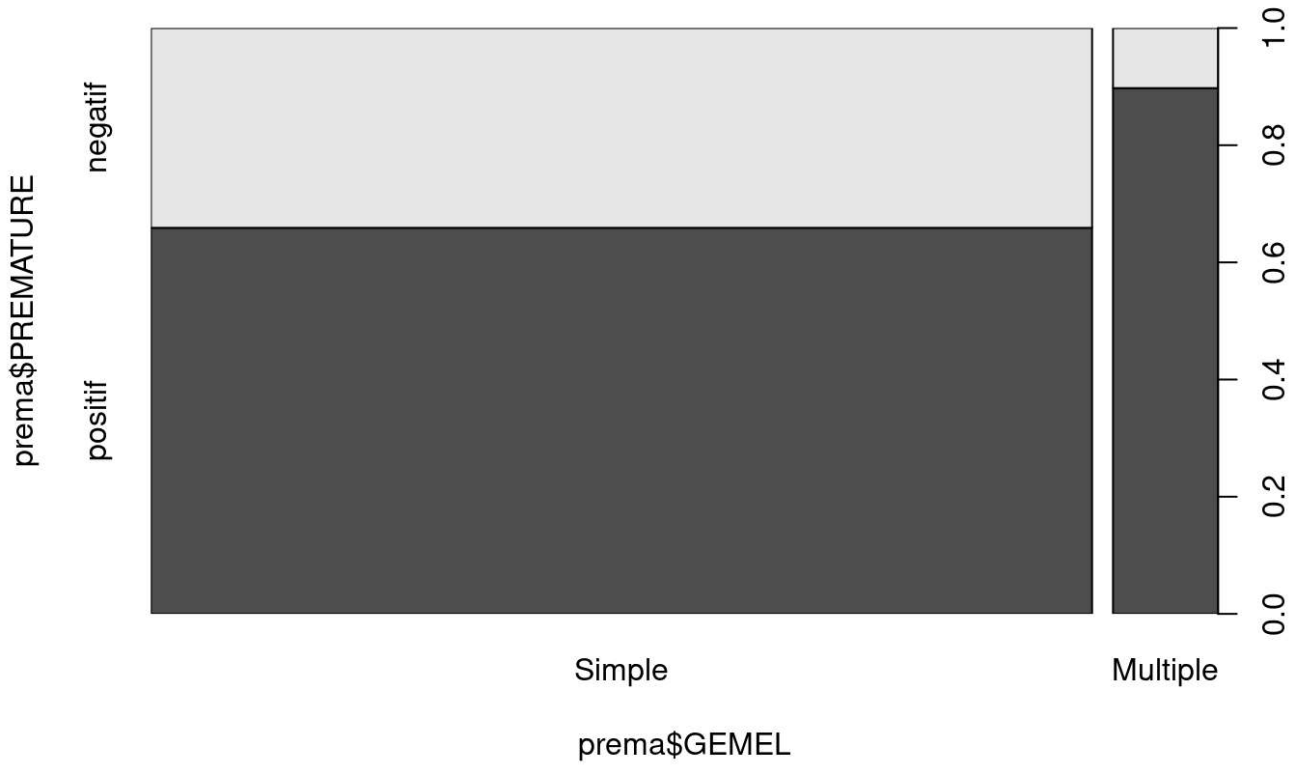
	negatif	positif
Simple	0.3409742	0.6590258
Multiple	0.1025641	0.8974359

À partir du tableau de contingence d'accouchement prématuré lors d'une grossesse multiple, on fait un prop.table, avec l'argument `margin = 1` pour que les calculs soient faits sur les lignes.

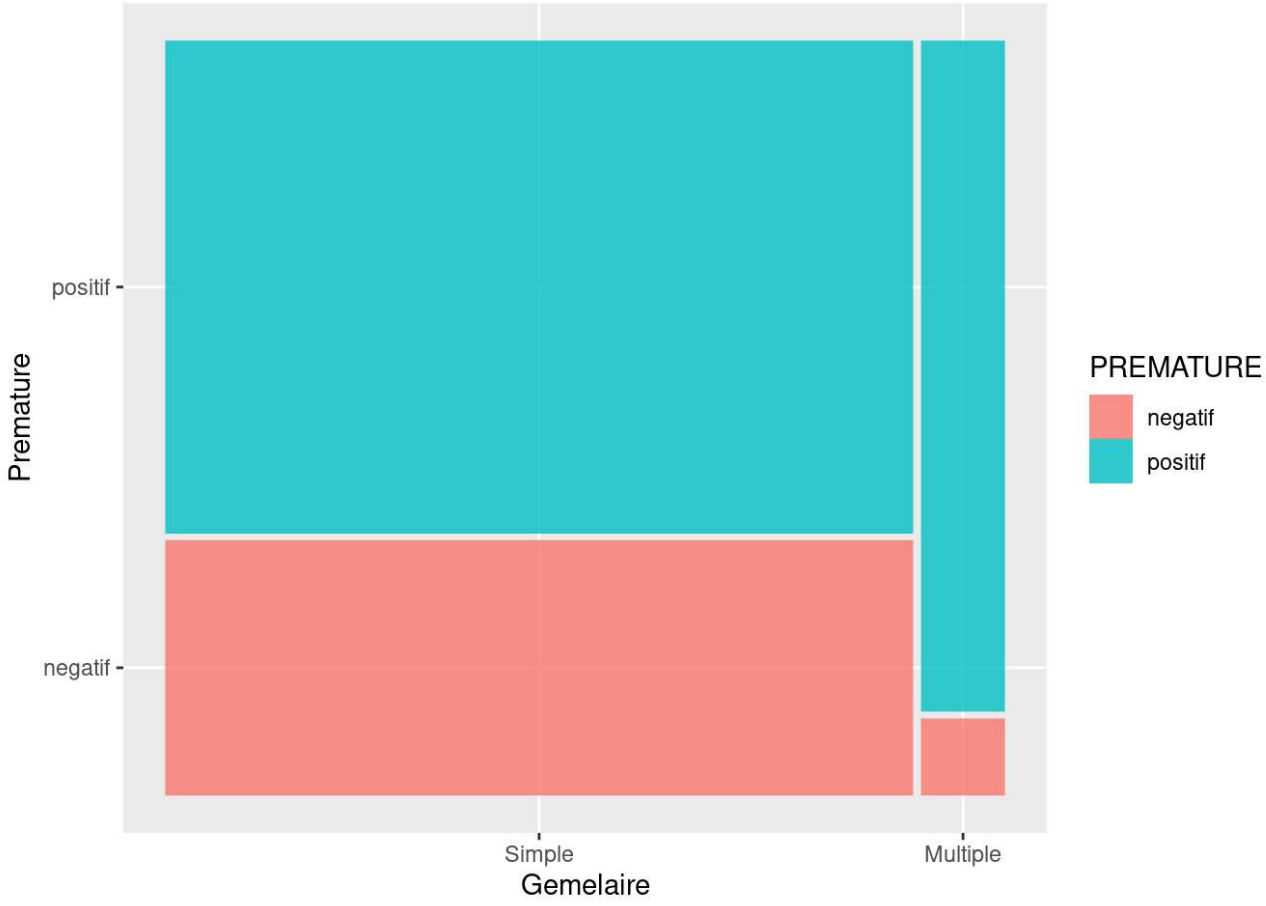
Dans les deux cas, on obtient une probabilité de 90% d'accouchement prématuré lors d'une grossesse multiple.

4. Représenter graphiquement la dépendance entre l'accouchement prématuré et le type de grossesse (multiples solutions).

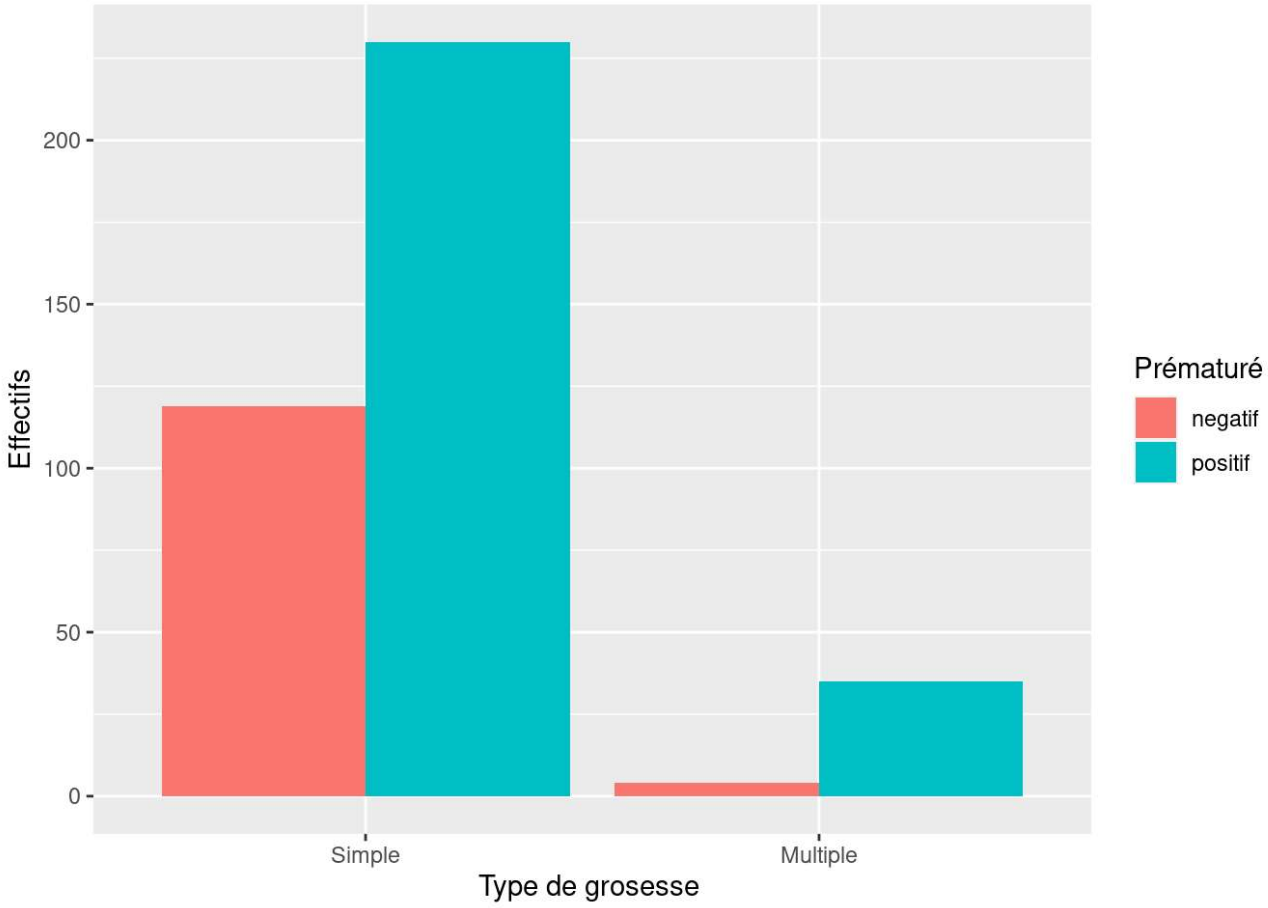
```
plot(prema$PREMATURE ~ prema$GEMEL)
```



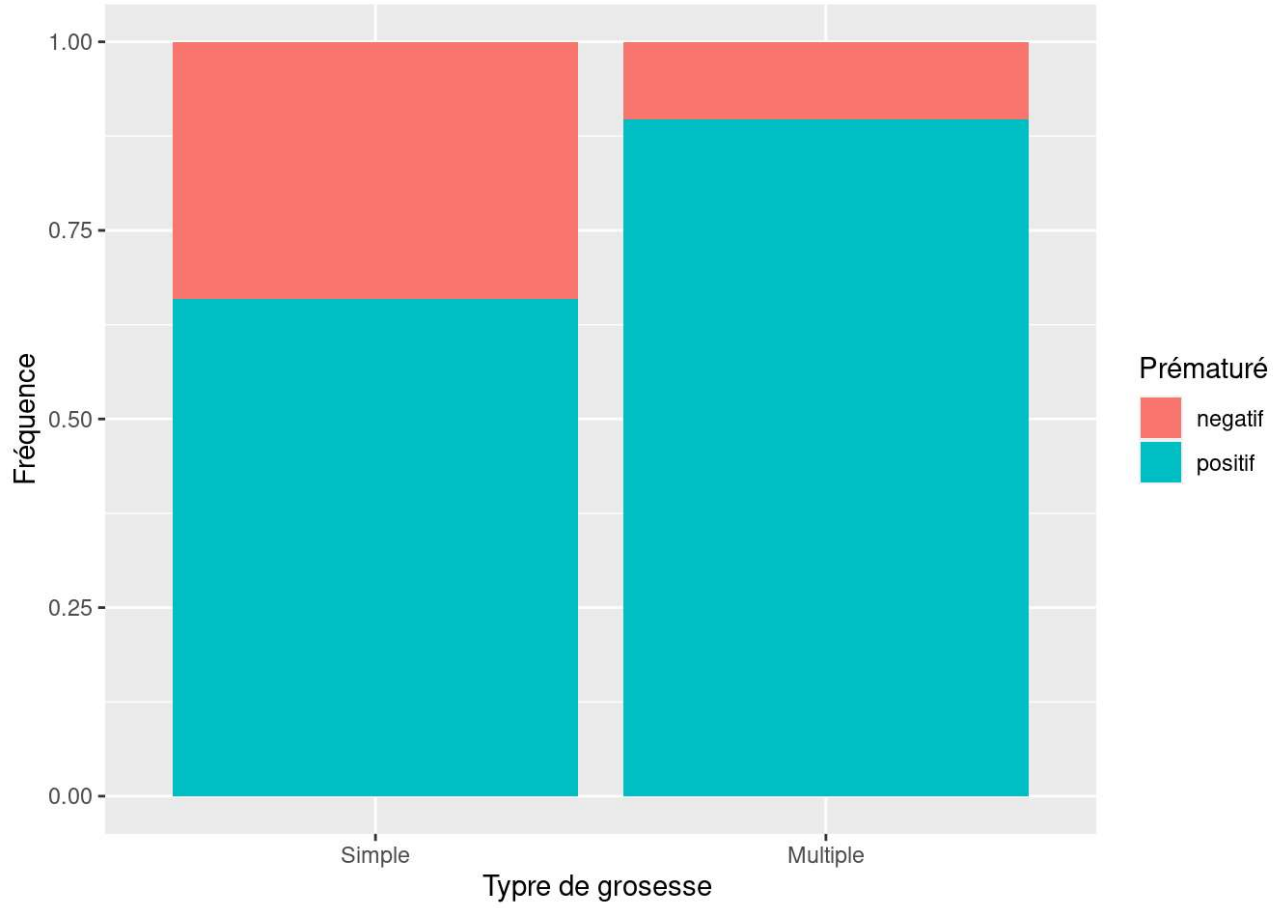
```
ggplot(data = prema) +
    geom_mosaic(aes(x = product(PREMATURE, GEMEL), fill = PREMATURE)) +
    xlab("Gemelaire") +
    ylab("Premature")
```




```
ggplot(prema) +
    aes(x = GEMEL, fill = PREMATURE) +
    geom_bar(position = "dodge") +
    xlab("Type de grosesse ") +
    ylab("Effectifs") +
    labs(fill = "Prématuré")
```




```
ggplot(prema) +
    aes(x = GEMEL, fill = PREMATURE) +
    geom_bar(position = "fill") +
    xlab("Typre de grosesse ") +
    ylab("Fréquence") +
    labs(fill = "Prématuré")
```



Les deux premiers graphes permettent de donner une représentation graphique des proportions de chaque classes (reportées sur les axes) et de chaque croisement de classe (aire). Les deux graphes suivant donnent une représentation en histogramme des données en termes d'effet d'abord, puis de fréquence.

5. Ajuster le modèle expliquant l'accouchement prématuré par le type de grossesse GEMEL. Pour cela utilisez la fonction `glm` avec `family="binomial"` .

```
model1  <- glm(PREMATURE ~ GEMEL, family="binomial", data=prema)
```

Les arguments utilisés sont :

- `PREMATURE ~ GEMEL` indique qu'on modélise la variable PREMATURE par la variable GEMEL
- `family="binomial"` indique l'utilisation d'une régression logistique, aussi appelée régression binomiale
- `data = prema` sélectionne le jeu de données prema.

```
model1
```

```
##
## Call:  glm(formula = PREMATURE ~ GEMEL, family = "binomial", data = prema)
##
## Coefficients:
##   (Intercept)  GEMELMultiple
##         0.659          1.510
##
## Degrees of Freedom: 387 Total (i.e. Null);  386 Residual
## Null Deviance:       484.7
## Residual Deviance: 473.7     AIC: 477.7
```

Ici :

- Call : affiche le modèle qui a été ajusté
- Coefficients : affiche les coefficients estimés, ici $\hat{\beta}_0 = 0,659$ et $\hat{\beta}_1 = 1,510$
- Degrees of Freedom :
  - 387 Total : Pour le modèle Null (sans variable explicative), nombre de données 388 - 1 car estimation de la proportion de grossesses prématurées
  - 386 Residual : 388 - 2 car estimation de deux paramètres
- Null Deviance: 484,7, déviance du modèle Null c'est-à-dire $D_0 = -2\ell_0$ avec ℓ_0 la log-vraisemblance du modèle Null
- Residual Deviance: 473,7, déviance du modèle c'est-à-dire $D = -2\ell$ avec ℓ la log-vraisemblance du modèle
- AIC: 477,7, critère AIC : $AIC = -2\ell + 2\nu = D + 2\nu$ où ν est le nombre de paramètres du modèle

Ici les modalités "negative" de PREMATURE et "Simple" de GEMEL servent de modalités de références (premiers niveaux de la variable).

```
levels(prema$PREMATURE)
```

```
## [1] "negatif" "positif"
```

```
levels(prema$GEMEL)
```

```
## [1] "Simple"   "Multiple"
```

La fonction `relevel` permet au besoin de redéfinir la modalité de référence.

Pour faire le lien avec les notations du cours, l'ajustement précédent est équivalent à

```
Y = ifelse(prema$PREMATURE == "positif", 1,0)
X = ifelse(prema$GEMEL == "Multiple",1,0)
knitr::kable(head(cbind.data.frame(Y,prema$PREMATURE,X,prema$GEMEL)))
```

Y	prema$PREMATURE	X	prema$GEMEL
1	positif	0	Simple

Y	prema$PREMATURE\|X\|prema$GEMEL	
1	positif	1 Multiple
1	positif	0 Simple
1	positif	1 Multiple
1	positif	0 Simple
1	positif	0 Simple

```
glm(Y ~ X, family = "binomial") # Fidèle aux notations du cours
```

```
##
## Call:  glm(formula = Y ~ X, family = "binomial")
##
## Coefficients:
## (Intercept)            X
##       0.659        1.510
##
## Degrees of Freedom: 387 Total (i.e. Null);  386 Residual
## Null Deviance:       484.7
## Residual Deviance: 473.7     AIC: 477.7
```

En fait un recodage binaire des variables est implicitement opéré par l'appel à glm.

Enfin le modèle permet de calculer les différentes probabilités

```
b0 = model1$coefficients[1]
b1 = model1$coefficients[2]
b0
```

```
## (Intercept)
##   0.6589558
```

```
b1
```

```
## GEMELMultiple
##      1.510098
```

```
# P(PREMATURE = positif | GEMEL = Simple) = 0,65
exp(b0)/(1+exp(b0))
```

```
## (Intercept)
##   0.6590258
```

```
# P(PREMATURE = positif | GEMEL = Multiple) = 0,89
exp(b0+b1)/(1+exp(b0+b1))
```

```
## (Intercept)
##   0.8974359
```

6. Le coefficient associé à la variable `GEMEL` est-il significatif ? Retrouver de deux manières différentes l'odd-ratio associé. L'interpréter.

```
summary(model1)
```

```
##
## Call:
## glm(formula = PREMATURE ~ GEMEL, family = "binomial", data = prema)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1341  -1.4669   0.9132   0.9132   0.9132
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.6590     0.1129   5.836 5.36e-09 ***
## GEMELMultiple   1.5101     0.5397   2.798  0.00514 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 484.69  on 387  degrees of freedom
## Residual deviance: 473.69  on 386  degrees of freedom
## AIC: 477.69
##
## Number of Fisher Scoring iterations: 4
```

Si on regarde les résultats du modèle à l'aide d'un `summary` , notre variable `GEMEL` est significative avec une p-valeur <0,01.

Pour le calcul de l'odds-radio, on peut le faire : - à partir du coefficient estimé

```
exp(b1)
```

```
## GEMELMultiple
##      4.527174
```

- à partir du tableau de contingence (pour vérifier la cohérence)

```
cotegrmultiple=35/4 #(35/39)/(4/39)
cotegrsimple=230/119
OR=cotegrmultiple/cotegrsimple
OR
```

```
## [1] 4.527174
```

La cote de l'évènement "l'accouchement est prématuré" est multipliée par 4,53 quand on passe de la modalité "Simple" à la modalité "Multiple" pour la variable GEMEL.

3 Étude d'une variable quantitative

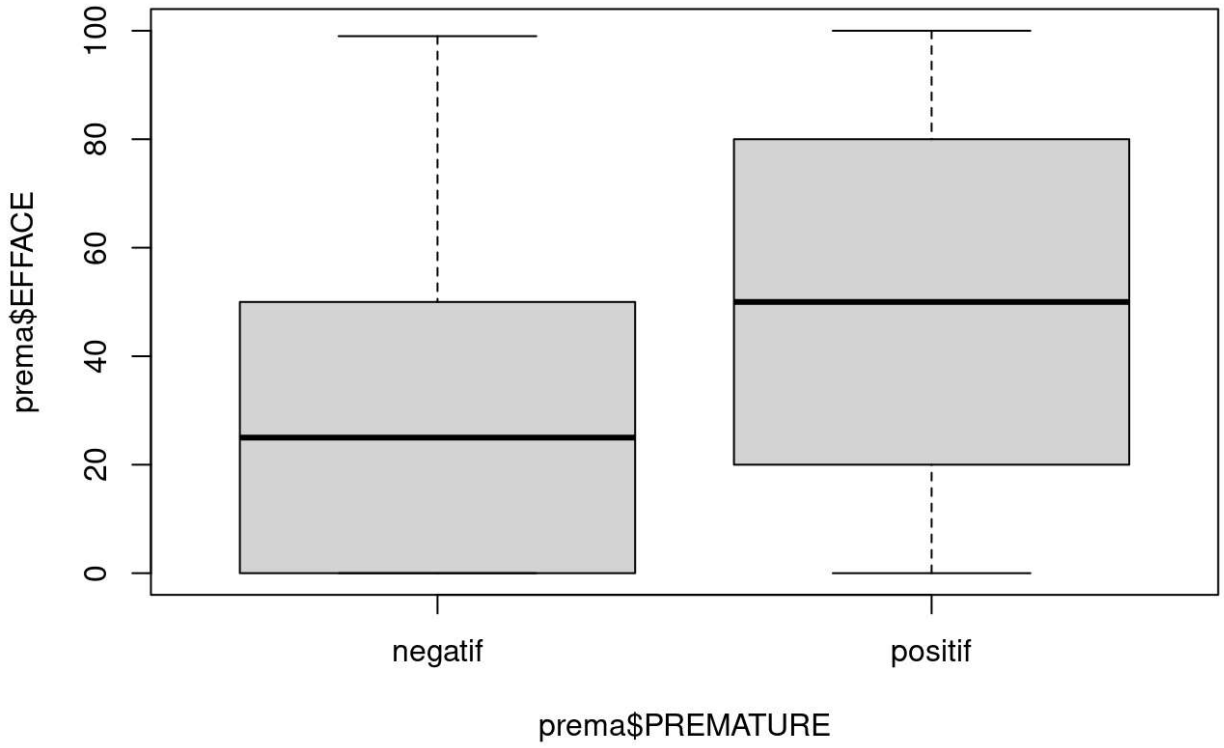
7. Quel est l'effacement moyen du col chez les patientes ayant accouché prématurément ? chez les autres ? La variable concernée est EFFACE.
   Vous pourrez vous aider de la fonction by.

```
by(prema$EFFACE, prema$PREMATURE, mean)
```

```
## prema$PREMATURE: negatif
## [1] 27.02439
## ------------------------------------------------------------
## prema$PREMATURE: positif
## [1] 51.80377
```

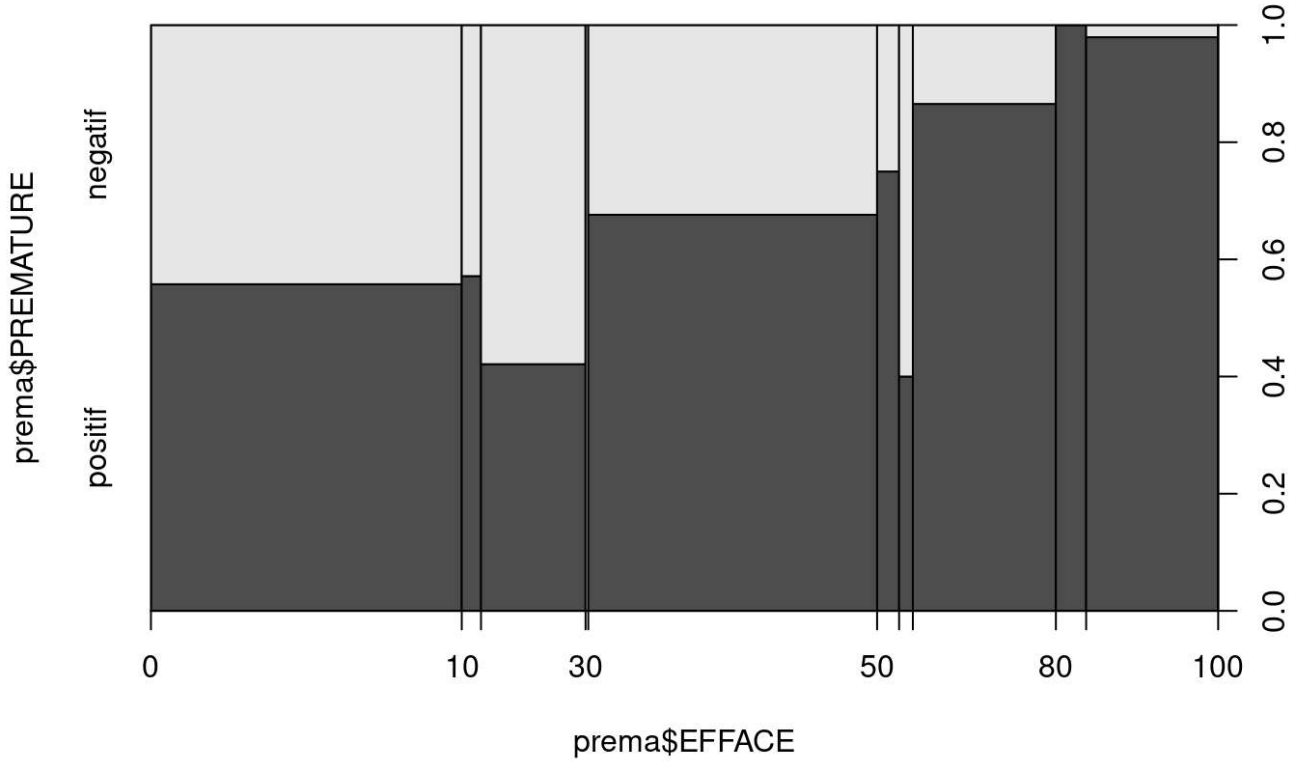
8. Faire des graphiques permettant d'illustrer la dépendance entre l'effacement du col et l'accouchement prématuré. On pourra par exemple utiliser les commandes suivantes :

```
plot(prema$EFFACE ~ prema$PREMATURE)
```



et

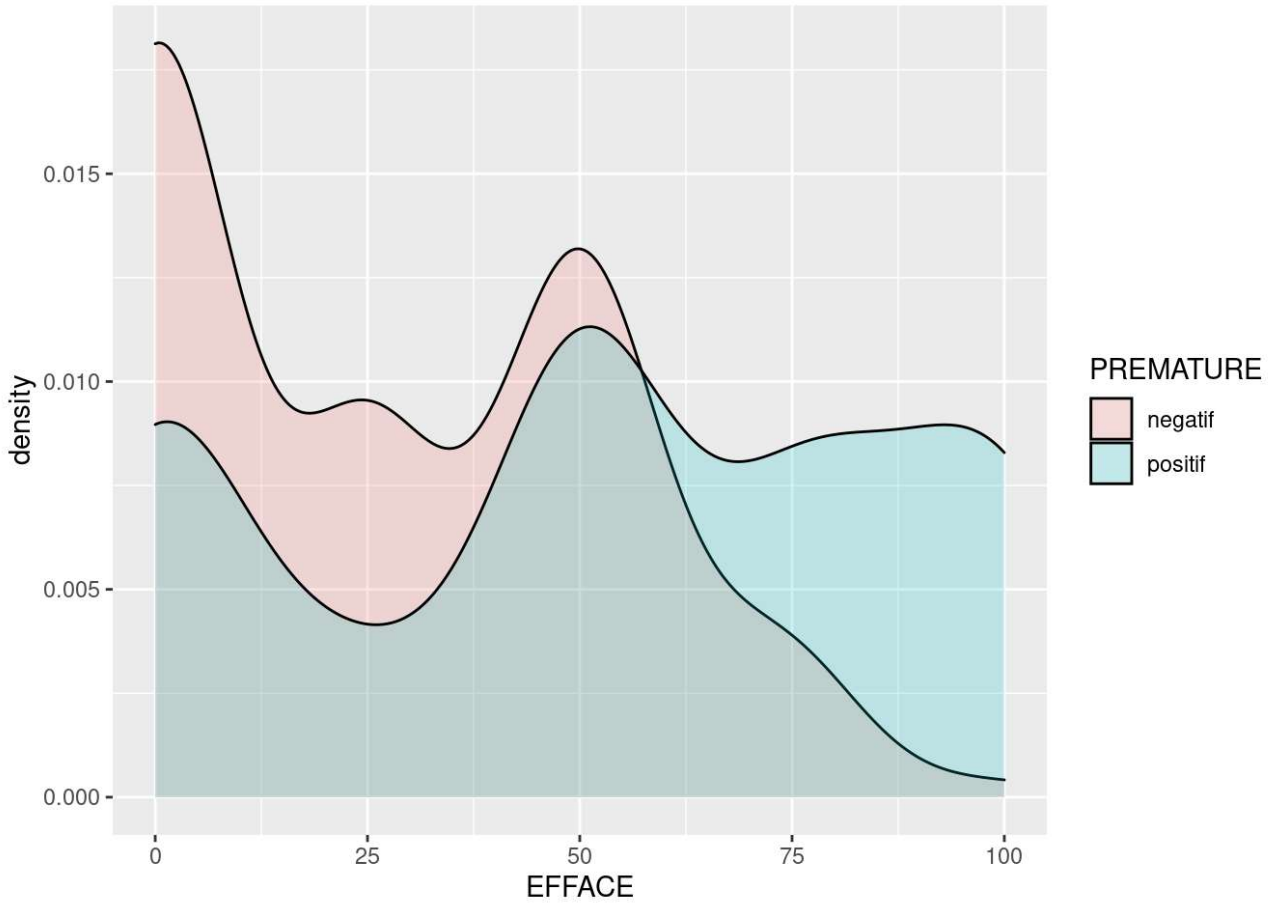
```
plot(prema$PREMATURE ~ prema$EFFACE)
```



qui lui est basé sur un découpage en classe de la variable EFFACE.

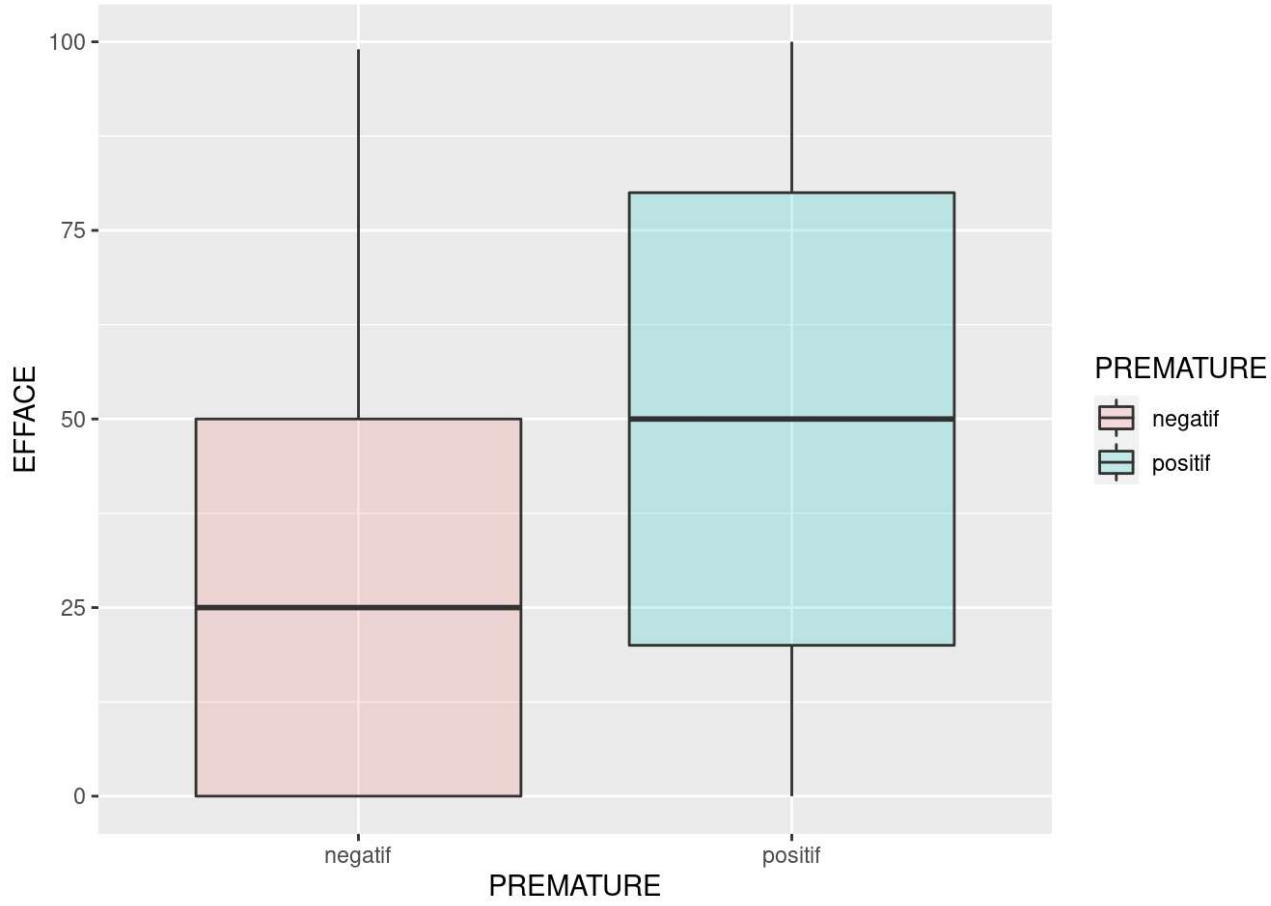
On peut aussi visualiser la distribution de EFFACE dans chacune des classes comme suit :

```
ggplot(prema, aes(EFFACE, fill = PREMATURE)) + geom_density(alpha = 0.2)
```



Une représentation en boite à moustache avec ggplot :

```
ggplot(prema, aes(x=PREMATURE,y=EFFACE, fill = PREMATURE)) + geom_boxplot(alpha =
0.2)
```



On remarque donc que plus l'effacement du col est grand, et plus l'accouchement a de risque d'être prématuré.

9. Ajuster le modèle expliquant l'accouchement prématuré par l'effacement du col (model2).

```
model2 <- glm(PREMATURE ~ EFFACE, family="binomial", data=prema)
summary(model2)
```

```
##
## Call:
## glm(formula = PREMATURE ~ EFFACE, family = "binomial", data = prema)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1201  -1.1264   0.5786   0.7852   1.2293
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.121256   0.170024  -0.713    0.476
## EFFACE       0.022799   0.003634   6.273 3.53e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 484.69  on 387  degrees of freedom
## Residual deviance: 439.51  on 386  degrees of freedom
## AIC: 443.51
##
## Number of Fisher Scoring iterations: 3
```

En sortie nous avons :

- `Call`
- `Coefficiens` : les coefficients, leur variance, la valeur Z du test de Wald, sa p-valeur et un codage en étoiles pour indiquer la significativité
- et comme pour un affichage du modèle sans la fonction summary la déviance et l'AIC

10. Exprimer π(x) = P (PREMATURE = 1/EFFACE = x) en fonction de x et écrire une fonction R permettant de réaliser ce calcul.

```
b0=coef(model2)[1]
b1=coef(model2)[2]
calculpi=function(x){
  exp(b0+b1*x)/(1+exp(b0+b1*x))
}
```

11. Quelle est la probabilité d'accoucher prématurément quand le col est effacé à 60%

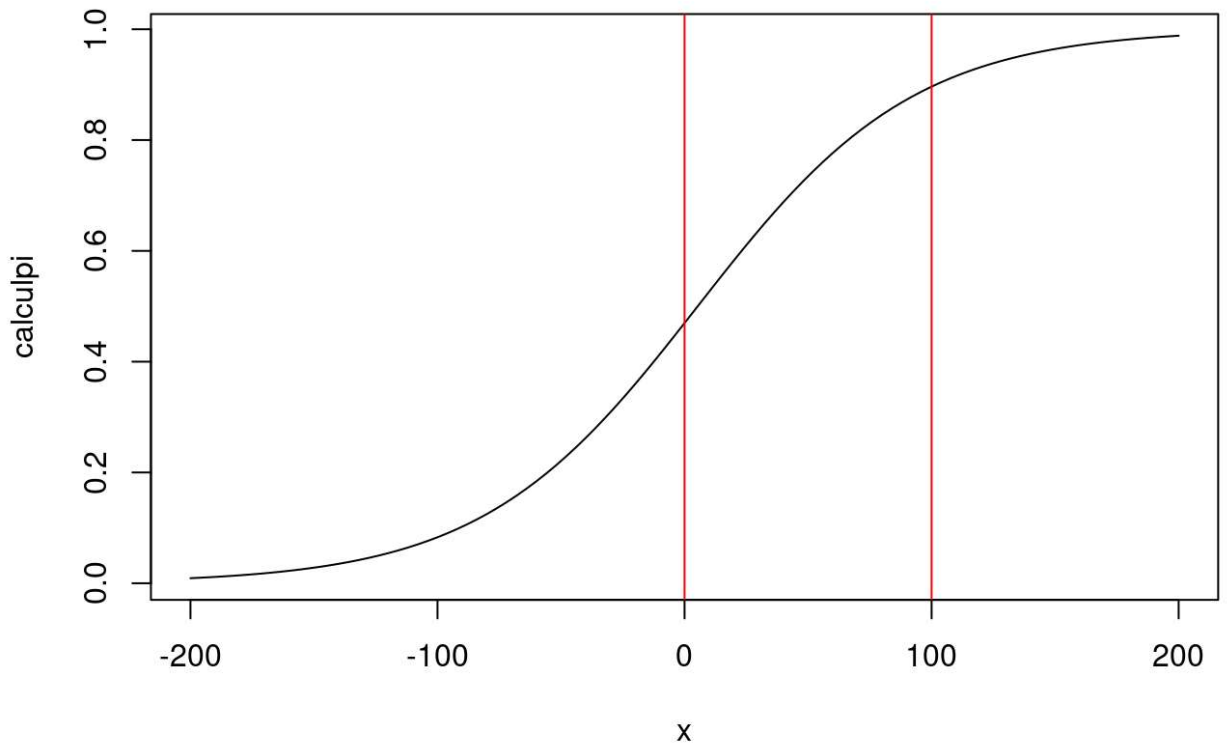
```
calculpi(60)
```

```
## (Intercept)
##    0.776724
```

La probabilité d'accouchement prématuré sachant EFFACE = 60 est donc de 78%.

L'allure de la fonction est la suivante

```
plot(calculpi,-200,200)
abline(v = c(0,100),col = "red")
```



On retrouve donc la forme sigmoïde; cependant, dans le cadre d'étude x ne varie qu'entre 0 et 100 (zone limitée par les traits rouges).

Ici on a bien entendu croissance de la probabilité de grossesse prématurée en fonction de la variable EFFACE (coefficient $\hat{\beta}_1$ positif).

12. Utiliser la fonction précédemment écrite pour calculer le score π associé aux femmes de l'étude. Comparer ce score aux résultats renvoyés par les commandes suivantes

```
pi_hat=predict(model2, prema, type="response")
model2$fitted.values
```

```
head(calculpi(prema$EFFACE))
```

```
## [1] 0.8964726 0.4697230 0.8964726 0.8304275 0.8304275 0.8964726
```

```
pi_hat=predict(model2, prema, type="response")
head(pi_hat)
```

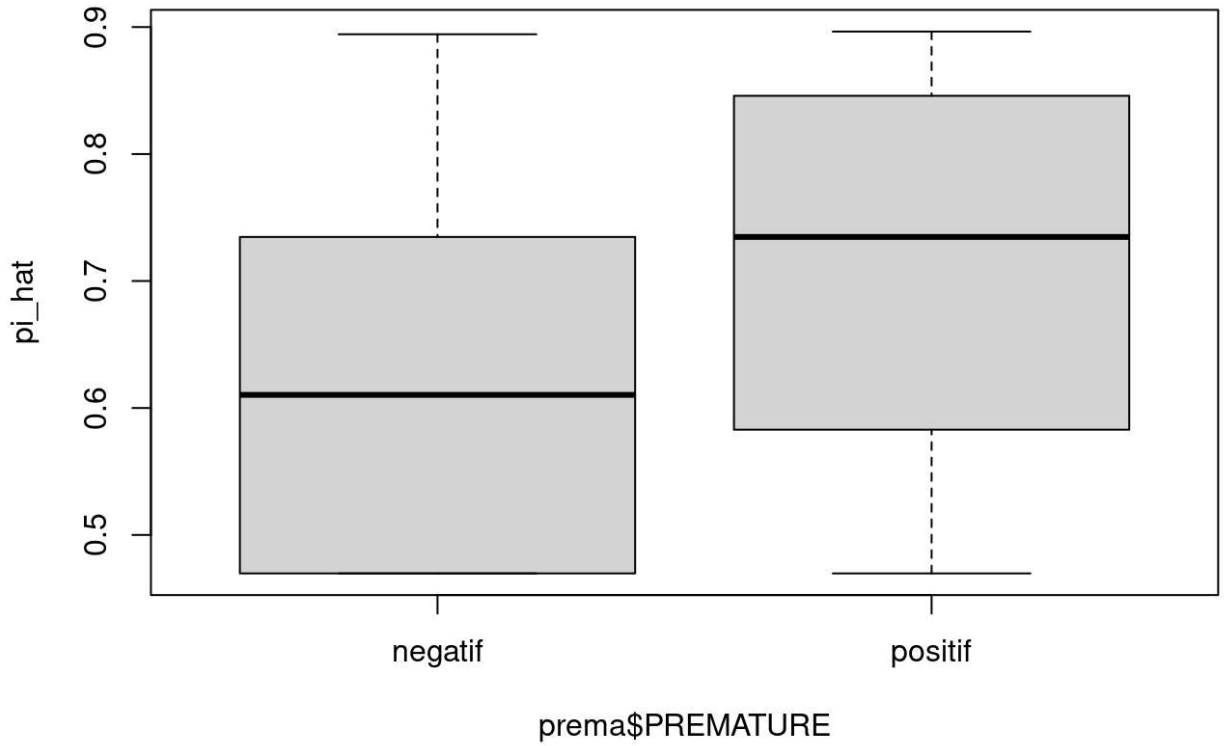
```
##         1         2         3         4         5         6
## 0.8964726 0.4697230 0.8964726 0.8304275 0.8304275 0.8964726
```

```
head(model2$fitted.values)
```

```
##         1         2         3         4         5         6
## 0.8964726 0.4697230 0.8964726 0.8304275 0.8304275 0.8964726
```

Comparaison de la distribution des probabilités calculées entre les patientes ayant accouché prématurément et les autres.

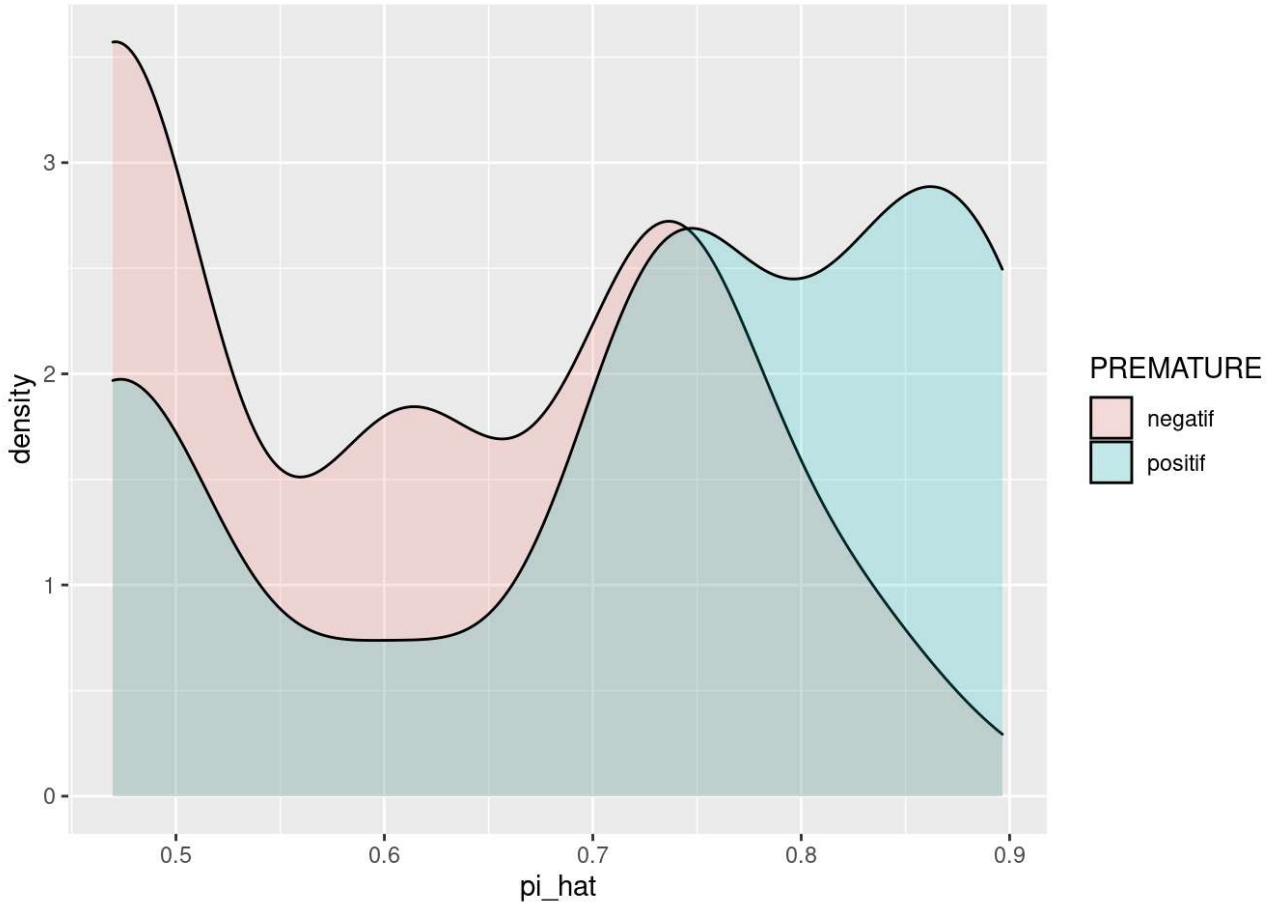
```
boxplot(pi_hat ~ prema$PREMATURE)
```



Ici on a bien des probabilités en moyenne plus petites pour les patientes avec grossesses prématurées que pour les autres. Cependant la séparation est assez faible (boxplot chevauchantes)

On peut comparer la densité de $\pi(x)$ dans chacune des classes :

```
ggplot(data.frame(prema, pi_hat), aes(pi_hat, fill = PREMATURE)) + geom_density(a
lpha = 0.2)
```


density
pi_hat
0
1
2
3
0.5
0.6
0.7
0.8
0.9
PREMATURE
negatif
positif