**TP de classification supervisée : Fiche 4 (TP4-5)**

# 1 Etude des données prématurés

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 (1=oui ; 0 = non)
Les données contiennent les variables suivantes :

Table 1 – Variables disponibles pour les données sur les accouchements

| Var | Description | Commentaire |
|---|---|---|
| `GEST` | l'âge gestationnel à l'entrée dans l'étude | en semaines |
| `DILATE` | la dilatation du col | 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` | les membranes | 1 : rupturées |
| | | 2 : non rupturées |
| | | 3 : incertain |
| `AGE` | âge de la mère | en années |
| `STRAT` | période de grossesse | 1-4 |
| `GRAVID` | la gestité | nombre de grossesses antérieures y compris celle en cours |
| `PARIT` | la parité | nombre de grossesses à terme antérieures |
| `DIAB` | problème de 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 |

L'objectif est de définir les facteurs prédictifs d'un accouchement prématuré ($Y$). Pour chaque modèle considéré, on notera $\pi$ la probabilité d'un accouchement prématuré sachant les variables $X_1, \cdots, X_p$ incluses.

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()`

```
load("prema.RData")
str(prema)
prema$DIAB = as.factor(prema$DIAB)
attach(prema)
```

## Etude d'une variable binaire

2. Construire le tableau de contingence PREMATURE/GEMEL.
3. Calculer la probabilité d'accoucher prématurément lors d'une grossesse multiple.
4. Ajuster le modèle expliquant l'accouchement prématuré par le type de grossesse GEMEL (`model1`).

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

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

## Etude d'une variable quantitative

6. Quel est l'effacement moyen du col chez les patientes ayant accouché prématurément ? chez les autres ? (vous pourrez vous aider de la fonction `by`)
7. Ajuster le modèle expliquant l'accouchement prématuré par l'effacement du col (`model2`).
8. Exprimer $\pi(x) = P(\text{PREMATURE} = 1/\text{EFFACE} = x)$ en fonction de $x$ et écrire une fonction R permettant de réaliser ce calcul.
9. Quelle est la probabilité d'accoucher prématurément quand le col est effacé à 60% ?
10. Utiliser la fonction précédemment écrite pour calculer le score $\pi$ 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
```

11. Faire un graphique permettant d'illustrer la dépendance entre l'effacement du col et l'accouchement prématuré. On pourra par exemple tracer deux densités correspondant aux distributions du score dans les deux groupes en utilisant les commandes suivantes :

```
library(lattice)
gS = densityplot(~pi_hat, data = data.frame(prema, pi_hat), groups = PREMATURE,
    plot.points = FALSE, ref = TRUE, auto.key = list(columns = 1))
print(gS)
```

## Etude de plusieurs variables explicatives

12. Ajuster le modèle expliquant l'accouchement prématuré par le type de grossesse et l'effacement du col (`model3`) :

```
model3 <- glm(PREMATURE ~ GEMEL + EFFACE, family = "binomial",
    data = prema)
summary(model3)
```

13. Comparer les deux modèles `model2` et `model3` en utilisant le test du rapport de vraisemblance

```
anova(model2, model3, test = "LRT")
```

14. Quel modèle gardez-vous ?
15. Estimer le modèle complet (`fullmodel`) :

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

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

16. Évaluer la significativité de chaque coefficient de `fullmodel`. Utilisez la fonction `step` pour la sélection automatique de variables dans le modèle et interpréter. On appelle `reduced` le modèle réduit aux variables sélectionnées. Comparer les deux modèles (complet et réduit).
17. Interpréter les coefficients de `reduced`. Quels sont les facteurs de risque pour l'accouchement prématuré ? Quels sont les facteurs protecteurs ?
    Aide :

```
exp(cbind(OR=coef(reduced), confint(reduced)))
```

## Evaluation de la règle de décision

18. Calculer les valeurs prédites des probabilités d'intérêt en utilisant la fonction `predict` ou le champ `fitted.values` de `reduced`. On nommera ce nouveau score S. Visualiser et commenter la qualité de prédiction (tracer par exemple des boites à moustache).
19. Calculer la matrice de confusion pour un seuil de décision à $0,5$.
20. On décide arbitrairement d'affecter toutes les valeurs qui ont un score S supérieur au score de la dernière ligne au groupe 1 et les autres au groupe 0. Calculer alors sensibilité et spécificité pour ce seuil.
21. Tracer la courbe ROC associée au score S en utilisant les fonctions `prediction` et `performance` du package `ROCR`

```
library(ROCR)
pred = prediction(S, prema$PREMATURE)
perf = performance(pred, "tpr", "fpr")
plot(perf)
```

22. Explorer les objets qui permettent de calculer la courbe ROC :

```
perf@x.values[[1]]
perf@y.values[[1]]
perf@alpha.values[[1]]
```

23. Calculer l'aire sous la courbe ROC en utilisant les commandes suivantes :

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

24. Calculer le seuil le plus proche du point idéal pour la courbe ROC liée au score S. Calculer la nouvelle matrice de confusion associée à ce seuil de décision.