TP 3 sur la régression logistique

1. On considère la base de données **diabetes**. Cette base de données est accessible sur ma page web et peut être importée directement sur R en tapant la commande `read.table("diabetes.csv",sep=",",header=TRUE)`. Elle contient des informations sur les pimas, un peuple nord-amérindien originaire du Mexique et du Sonora. Le but est d'identifier les facteurs de risque du diabète et de prédire le risque d'être diabétique dans cette population. La base de données contient les variables suivantes :

- pregnant : le nombre de fois où l'individu a été enceinte
- glucose : la concentration de plasma glucose
- pressure : la pression sanguine diastolique (mm Hg)
- skin thickness : l'épaisseur du pli cutané du triceps (mm)
- insulin : l'insuline (mu U/ml)
- BMI : l'Indice de Masse Corporelle (poids en kg/(taille en mètres)^2)
- diabetes pedigree function : score mesurant le risque familial du diabète
- age : l'âge
- outcome : vaut 1 si la personne a du diabète, 0 sinon

1. Faire une analyse descriptive de la base de données.
2. Ajuster un modèle de régression logistique à l'aide de la fonction `glm` avec l'option `family="binomial"`.
3. Afficher les résultats à l'aide de la commande `summary`
4. Calculer les odds-ratios et leurs intervalles de confiance. On pourra obtenir ce résultat de deux façons :
   - à partir de la fonction `confint`
   - à l'aide de la fonction `tidy` de la librairie `broom`, que l'on combinera avec la fonction `ggplot` pour obtenir une représentation graphique des odds-ratios et de leurs intervalles de confiance.
5. Interpréter les sorties. Quelles semblent être les facteurs de risque du diabète ? Proposer une méthode de sélection de variables permettant de construire le meilleur sous-modèle au sens de la prédiction du diabète. On gardera ce sous modèle dans la suite de l'exercice.
6. Visualiser l'effet de chaque variable sur le diabète à l'aide de la fonction `ggpredict` de la librairie `ggeffects`. Retrouver ces résultats par le calcul. Exhiber des profils d'individus particulièrement à risque d'être diabétiques.
7. Construire un échantillon d'apprentissage contenant 80% des données et un échantillon de test contenant les 20% des données restantes. Construire une règle de classification à partir de l'échantillon d'apprentissage et évaluer le taux de mauvaise classification sur les données test. On pourra pour cela utiliser la fonction `predict` permettant de prédire pour tout individu de l'échantillon test sa probabilité d'être diabétique.
8. A l'aide d'une boucle **for**, faire varier le seuil utilisé dans le critère de classification et calculer la sensibilité et la spécificité pour chacune des valeurs de seuil possible. A partir de ce résultat, tracer la courbe ROC du modèle.
9. Retrouver la courbe ROC en utilisant la fonction `roc` du package `pROC`.