

## TP régression logistique « Pima ».

L'objectif est d'étudier les facteurs favorisant le diabète chez les indiens Pima vivant près de Phoenix (Arizona).

La variable à expliquer est « test » valant 1 si des signes de diabète sont observés et 0 sinon.

Les variables explicatives sont

- pregnant : nombre de grossesses
- glucose : concentration de glucose dans le sang après 2 heures
- diastolic : pression diastolique
- triceps : épaisseur de la peau du triceps replié (sert de mesure de graisse dans le corps)
- insulin : taux d'insuline dans le sang 2 heures après le repas.
- bmi : indice de masse corporelle
- diabetes : mesure de l'influence des antécédents familiaux de diabète
- age : âge

1. Charger la table.
2. Regarder les effectifs de la variable « test »
3. Quelle est la prévalence pour ce jeu de données ?
4. Étudier la variable « diastolic ». Que remarquez-vous ?
5. Mettre les données codées 0 en NA pour toutes les variables (sauf pour la variable test !)
6. Faire une régression logistique.
7. Choisir le meilleur modèle en sélectionnant les variables par la méthode AIC, puis BIC.
8. Pour le meilleur modèle choisi, donner les OR. Comment les interpréter ?
9. Donner les intervalles de confiance des OR à 95 %
10. Quel est l'ODD ratio lorsque le BMI augmente de +10. Donner un IC.
11. Tracer une courbe ROC