Statistique et Informatique — UFR STAPS
Master 1 — Université de Paris X - Nanterre

TD 7&8. Régression

1 Données

Nous allons travailler sur l'enquête **EPO2009 Fraude à l'université**. Reprenez votre script et chargez les données en mémoires dans `donnees`.

2 Modèle choisit à la main

2.1 Régression linéaire

Pour commencer, on va chercher à expliquer la variable `[ScoreTriche]` (que vous avez créé lors du TD *Mesurer la Subjectivité*.

1. Choisissez 8 variables qui selon vous sont les plus à même d'avoir un lien avec le `ScoreTriche`.
2. Construisez le modèle. Pour cela, utilisez l'instruction :
   `model <- lm(ScoreTriche Variable1+variable2+...+Variable8,data=donnees)`.
3. L'instruction précédente a construit le modèle mais ne l'a pas affiché. Pour l'afficher, utilisez `summary(model)`.
4. Combien vaut le r^2 ajusté ? Le modèle est-il bon ?
5. Commentez votre modèle. Quels sont les variables qui influencent `ScoreTriche`, quelles sont celles sans influence ?

2.2 Régression logistique

On cherche à savoir si les hommes et les femmes se comportent de la même manière face à la triche. On va donc faire une régression logistique sur la variable `sexe`.

6. Choisissez 8 variables qui selon vous sont les plus à même d'avoir un lien avec le `Sexe`.
7. Construisez le modèle. Pour cela, utilisez l'instruction :
   `model <- glm(Sexe Variable1+variable2+...+Variable8,data=donnees,family='binomial')`.
8. L'instruction précédente a construit le modèle mais ne l'a pas affiché. Pour l'afficher, utilisez `summary(model)`.
9. Combien vaut le r^2 ajusté ? Le modèle est-il bon ?
10. Commentez votre modèle. Quels sont les variables qui influencent `ScoreTriche`, quelles sont celles sans influence ?

3 Modèle selon les tests

Le modèle précédent à été fait en choisissant les variables de manière assez imprécise. On peut faire mieux en n'intégrant dans le modèle que les variables qui ont un lien en bivarié avec la variable à expliquer.

11. Pour chacune des 8 variables que vous avez choisis, tester si elle a un lien avec `ScoreTriche`. Note : ne faites pas le diagnostic, faites simplement un test paramétrique sans vous souciez de sa validité.
12. Choisissez les variables qui sont le plus fortement associées avec `ScoreTriche`. Combien vaut le r^2 ajusté ?

On peut faire le même travail pour la variable `Sexe`

13. Mêmes questions pour la régression logistique.

4 Modèle de sélection descendante

Beaucoup d'autres méthodes existent pour construire un modèle :

14. Considérez le modèle dans lequel sont intégrés les 8 variables. Quel est le r^2 ajusté ?
15. Supprimez la variable pour laquelle le petit p est le plus élevé. Vous obtenez ainsi un modèle à 7 variables. Quel est le r^2 ajusté ?
16. Supprimez la variable pour laquelle le petit p est le plus élevé. Vous obtenez ainsi un modèle a 6 variables. Quel est le r^2 ajusté ?
17. Continuez jusqu'a n'avoir plus que 2 variables. Le meilleur modèle sera celui pour lequel le r^2 ajusté est le meilleur.

On peut faire le même travail pour la variable `Sexe`

18. Mêmes questions pour la régression logistique.