TD 2 — Régression logistique (R)

Année universitaire 2025–2026 ▪ Parcours Économie de la santé & Développement durable

Pierre Beaucoral

Questions TD

1. Télécharger le fichier bankloanT.xls depuis l'ENT, puis importer les données dans R

2. Etudier la distribution de la variable ed. Créer une variable catégorielle, puis une variable ne contenant que 4 classes.

3. Etudier la matrice des corrélations entre les 7 variables quantitatives présentes.

4. Recoder la variable à expliquer (default) en variable quantitative puis réaliser une première régression logistique incluant toutes les variables explicatives quantitatives.

5. Réaliser une deuxième régression logistique incluant aussi la variable educ en classe. Enregistrer aussi les résultats de ce modèle, qui est le modèle le plus complexe que nous appliquerons à ces données.

6. Tester l'ajustement de ce modèle complet, grâce à la commande hoslem.test, où g est le nombre de groupes de niveaux différents de fonction prédictive utilisé pour le test. Varier les valeurs de g pour vérifier la robustesse du résultat.

7. Grâce à la commande anova, réaliser un test de rapport de vraisemblance entre les deux modèles ajustés, et conclure sur la significativité de la variable educ.

8. Ôter la variable la moins significative du modèle retenu. Vérifier que la P-value du test de rapport de vraisemblance entre les modèles avec et sans cette variable est égale ou très proche de la P-value du test bilatéral de nullité du coefficient associé à la variable.

9. Une variable est à nouveau très peu significative. Ajuster un nouveau modèle sans cette variable. Sauvegarder les valeurs prédites par ce modèle.

10. Etablir le tableau de contingence des individus bien ou mal classés avec une règle de coupure à 0,5

11. Etablir la courbe ROC pour le « meilleur » modèle, puis calculer les “probabilités prédites” et étudier leur distribution

12. Refaire la même modélisation avec un modèle Probit et observer les différences et ressemblances avec la modélisation Logit.

Pour aller plus loin...

Supposant qu'un individu ne remboursant pas son emprunt coûte en moyenne 100000$, et qu'un individu payant son emprunt rapporte en moyenne 40000$, on peut calculer (…) qu'il est optimal de n'accorder un prêt qu'aux individus ayant une probabilité de rembourser estimée à 0,7 ou plus.

13. Règle de décision (probabilité de remboursement >= 0.7)

14. Etablir le tableau de contingence des individus bien ou mal classés en considérant comme défaillants potentiels tous les individus ayant moins de 70% de chances de rembourser. Commenter ce tableau.