TD1

Econométrie des variables qualitatives 1
Modèle logit binaire et ordonné

Exercice : L'hypertension artérielle et ses facteurs explicatifs

Vous disposez pour 500 personnes de leur statut en termes d'hypertension (Press_arter) ainsi que des caractéristiques sur leur comportement. Les données sont disponibles dans la Base_pression_arterielle.xls sous Madoc.

- Press_arter : la pression systolique se répartit en 4 classes ordonnées

Classe	Modalités (millimètre de mercure, mmHg)
1 : Pression artérielle normale	< 140
2 : Hypertension artérielle de grade 1	[140-159[
3 : Hypertension artérielle de grade 2	[160-179[
4 : Hypertension artérielle de grade 3	≥ 179

Genre = 1 si la personne est un homme, 0 sinon
Fumer = 1 si la personne fume, 0 sinon
Sport = 1 si la personne pratique le sport de manière intensive, 0 sinon
Age : Age de la personne
Alcool =1 si la personne boit de l'alcool de manière excessive, 0 sinon
IMC : indice de masse corporelle
Stress = 1 si la personne est stressée, 0 sinon
Sel = 1 si l'alimentation de la personne est très salée, 0 sinon

Il existe également dans la base de données la variable Pression = 1 si la pression artérielle de la personne est supérieure à 140 (supérieure à la normale), 0 dans le cas contraire

Questions :

1) Importer la base sous le logiciel R. Nommer la base : Pression
2) Vérifier la corrélation entre les différentes explicatives quantitatives. Qu'en concluez-vous ?
3) Transformer les variables de type qualitatif en facteur.
4) Réaliser les différentes statistiques. Représenter également les boites à moustache des variables Age et IMC ainsi que leur histogramme respectif de telle manière que les deux boites à moustache soient sur la première ligne et les histogrammes sur la seconde ligne. L'histogramme de l'âge doit être de couleur rouge, celui de la variable IMC en bleu. Les axes des abscisses des histogrammes doivent être systématiquement nommés. Les deux histogrammes doivent être avoir un titre. Qu'en concluez-vous ?

5) Quelles sont les variables qui permettent (et ne permettent pas) d'expliquer de manière significative la probabilité que les personnes aient une hypertension artérielle supérieure à la normale ?

   Pensez à regarder :

   - La multicolinéarité entre les variables explicatives utilisées dans l'estimation du modèle
   - L'hypothèse de nullité de l'ensemble des coefficients des variables explicatives du modèle
   - Les effets marginaux pour les variables explicatives quantitatives
   - Les odds-ratios pour les variables explicatives qualitatives
   - Le tableau de prédiction et par conséquent le taux d'erreur du modèle estimé
   - Le taux de sensibilité et de spécificité du modèle estimé, le ROC
   - La qualité d'ajustement du modèle estimé
   - L'existence (ou non) d'observations influençant de manière significative l'estimation
   - l'effet de l'IMC en fonction de la pratique du sport sur la probabilité d'avoir une pression artérielle supérieure à la normale
   - l'hypothèse d'homoscédasticité des erreurs du modèle estimé

6) Quelles sont les variables qui permettent d'expliquer de manière significative la probabilité que les personnes soient dans telle ou telle catégorie d'hypertension artérielle? Conserver l'ensemble des variables explicatives pour l'estimation du modèle.

   Pour répondre à cette question, vous devez en utilisant la fontion polr et la fonction vglm:
   - vérifier l'hypothèse de nullité de l'ensemble des coefficients des variables explicatives du modèle.
   - calculer les Odd ratios pour les différentes variables explicatives
   - calculer la qualité de prévision et d'ajustement du modèle estimé.

   Dans le cas de la fonction polr, vous pouvez également utiliser la fonction Effect pour la variable IMC

7) A partir de la fonction vglm, vérifiez l'hypothèse d'égalité des pentes pour le modèle incluant comme variables explicatives IMC,Fumer,Sport,Alcool. Est-elle vérifiée ?

8) Rechercher à partir du modèle précédant les variables explicatives vérifiant cette hypothèse et ré-estimer le modèle en ne conservant que les variables vérifiant cette hypothèse ?

9) A partir du modèle de la question 8 et de la fonction oglmx, calculer les effets marginaux au niveau moyen de l'échantillon (en considérant les variables explicatives binaires comme des dummies dans le calcul de ces effets)

10) Prendre en compte (ou non) l'hétéroscédasticité des erreurs du modèle de la question 8 en supposant que c'est la variable Alcool qui influence la variance des erreurs. Estimer le modèle le cas échéant. Interpréter les résultats du modèle.

11) A partir du modèle de la question 10, calculer les effets marginaux au niveau moyen de l'échantillon (en considérant les variables explicatives binaires comme des dummies dans le calcul de ces effets). Interpréter les résultats obtenus.