Projet sur les tests statistiques et la régression linéaire

Ce projet est à réaliser (mais pas à rendre) afin de préparer le QCM n°3 du 16 décembre dont les questions et les sorties `R` correspondantes seront issues de ce projet. Tous les tests sont à réaliser avec un risque α de 5 %. Même si les questions vous incitent à faire un test unilatéral, ne faites que des tests bilatéraux. Donnez pour chaque test les hypothèses nulles et alternatives ainsi que le nom du test utilisé. Pensez à vérifier les conditions de validité du test et à donner une conclusion précise.

Contexte : le fichier `medecins.txt` contient des données pour l'année 1999 d'un échantillon aléatoire de 500 médecins spécialistes de Midi-Pyrénées (pour information, la population des médecins spécialistes comptait 2082 membres en Midi-Pyrénées en 1999). Les médecins peuvent appartenir à deux secteurs différents : le secteur 1 où le montant de la consultation est fixe et le secteur 2 où le médecin choisit librement le montant de la consultation en pratiquant un dépassement (« avec tact et mesure »). Les médecins du secteur 2 sont donc autorisés à pratiquer des dépassements d'honoraires de façon habituelle, alors que ceux du secteur 1 peuvent dépasser occasionnellement (jours fériés, actes exceptionnels…). Les variables auxquelles vous allez vous intéresser sont les suivantes :

- $genre$: genre du médecin (F pour féminin et M pour masculin),
- $spec$: spécialité du médecin (en trois catégories : chirurgicale, médicale et mixte),
- $secteur$: secteur tarifaire du médecin (en deux catégories : secteur 1 et secteur 2),
- $nbpatients$: nombre total de patients vus par le médecin en 1999,
- $honor$: honoraires totaux du médecin pour l'année 1999 (en milliers d'euros),
- $moins16$: proportion de patients âgés de moins de 16 ans dans la patientèle du médecin,
- $60a69$: proportion de patients âgés de 60 à 69 ans dans la patientèle du médecin,
- $plus70$: proportion de patients âgés de plus de 70 ans dans la patientèle du médecin.

L'objectif de l'étude est de comprendre les liaisons qui peuvent exister entre les différentes variables et d'identifier en particulier les facteurs qui jouent un rôle sur les honoraires.

Partie 1 : analyse univariée et tests statistiques

1. Etude descriptive univariée

   Commencer l'étude statistique par une étude descriptive de toutes les variables du fichier. Pour chaque variable, décrire sa distribution à l'aide des outils appropriés en fonction de son type (résumés numériques et graphiques).

2. Différences hommes/femmes

   (a) Le choix de la spécialité par le médecin est-il lié à son genre ?

   (b) Les médecins spécialistes masculins gagnent-ils mieux leur vie que leurs consœurs ?

3. Influence du secteur tarifaire

   (a) La répartition par secteur est-elle la même dans les trois types de spécialités ?

   (b) Le secteur tarifaire auquel il appartient a-t-il un effet sur les honoraires du médecin ?

4. Effet des autres variables sur les honoraires

   (a) Les honoraires d'un médecin dépendent-ils de la proportion de personnes âgées dans sa patientèle ?

   (b) Le type de spécialité a-t-il une influence sur les honoraires des médecins ? Si oui, déterminer à l'aide de tests quelles spécialités diffèrent significativement en termes d'honoraires.

Partie 2 : régression linéaire

A Régression linéaire simple

Problématique : les honoraires d'un médecin peuvent-ils être modélisés (linéairement) en fonction du nombre de ses patients ?

1. Réaliser le nuage de points des honoraires en fonction du nombre total de patients et calculer le coefficient de corrélation linéaire de ces deux variables. Commenter.

2. Ecrire le modèle de régression linéaire simple théorique correspondant.

3. Estimer le modèle et commenter les résultats (R^2, test de validité globale du modèle, test de significativité du paramètre associé à la variable `nbpatients` et interprétation du paramètre estimé). Représenter la droite de régression des honoraires sur le nombre total de patients sur le nuage de points.

4. Vérifier la normalité des résidus (histogramme + QQ-plot) et donner un graphique permettant de repérer d'éventuels points aberrants.

B Régression linéaire multiple

Problématique : quels sont les déterminants des honoraires d'un médecin ?

1. Effectuer la régression linéaire multiple des honoraires en fonction de toutes les variables explicatives disponibles.

   Indication importante : on prendra la modalité `Mixte` comme modalité de référence pour la variable `spec`. Il vous faut pour cela créer les deux variables indicatrices des modalités `Chirurgicale` et `Médicale` et les mettre dans votre formule du modèle (à la place de `spec`).

Ecrire le modèle théorique correspondant. Commenter les résultats obtenus (R^2, test de validité globale du modèle) et dire quelles sont les variables qui ont un effet significatif sur les honoraires. On ne demande pas, dans cette question, de commenter leurs effets.

2. Certains coefficients de la régression précédente n'étant pas significatifs, recommencer un ajustement de régression linéaire multiple par la méthode de pas à pas vue en TP de façon à avoir à la fin un modèle ne contenant que des coefficients significatifs à 5 %. Donner toutes les sorties `R` de ce pas à pas. Pour le modèle final obtenu par le pas à pas :

   - écrire le modèle de régression linéaire théorique correspondant,
   - commenter le R^2 et le test de validité globale du modèle,
   - interpréter la valeur des paramètres estimés.

3. Vérifier la normalité des résidus dans ce dernier modèle (histogramme + QQ-plot) et donner un graphique permettant de repérer d'éventuels points aberrants.

Conclusion de l'étude

Faire une courte synthèse de votre étude en précisant les limites éventuelles quant à la fiabilité de vos résultats.