TP noté, durée 3h30

Tous les documents sont autorisés

On reprend le jeu de données utilisé lors de la SAE "Régression sur données réelles" (BUT1, semestre 2) :

En 2020, une étude internationale a été menée dans plus de 90 pays afin d'obtenir des informations sur les courbes de croissance des filles et des garçons âgés de 5 à 19 ans. La méthodologie de cette étude est la suivante : pour chaque pays, 200 médecins (généralistes ou pédiatres, hospitaliers ou libéraux) ont été sélectionnés au hasard sur l'ensemble du territoire. Chaque médecin a ensuite sélectionné au hasard 20 patients (10 de chaque sexe) âgés de 5 à 19 ans et a enregistré leur âge et leur taille. L'âge a été enregistré par tous les praticiens avec un arrondi d'un chiffre après la virgule (en années décimales), et la taille avec un arrondi d'un chiffre après la virgule (en cm). Pour chaque pays de l'étude, l'âge et la taille de 4000 enfants/jeunes adultes (2000 de chaque sexe) ont donc été relevés.

Dans le jeu de donnée dont vous disposez, vous devez analyser les résultats de l'étude pour deux pays.

Le rendu peut être sous forme R Mardown, word ou pdf. Seules les parties importantes du code (ou même seulement le nom des fonctions utilisées) doivent apparaître. La notation se fera principalement sur le rendu graphique et sur les commentaires associés.

Préliminaire.

Pour lire le jeu de donnée (par exemple ici Data1), utiliser la commande
`Data1<-read.csv(file="Data1.csv")`

Bien que ce ne soit pas indispensable, il est conseillé de créer des sous-jeux de données de la façon suivante (ici pour les pays Germany et Mali du jeu de donnée Data1) :

```
Ger=Data1[(Data1$Pays=="Germany"),]
GerG=Ger[(Ger$Sexe=="Girls"),]
GerB=Ger[(Ger$Sexe=="Boys"),]
Mali=Data1[(Data1$Pays=="Mali"),]
MaliG=Mali[(Mali$Sexe=="Girls"),]
MaliB=Mali[(Mali$Sexe=="Boys"),]
```

Pour les deux pays, tracer les deux nuages de points de la taille en fonction de l'âge.
Pour les deux pays, tracer le nuage de point de la taille en fonction de l'âge pour les filles et pour les garçons (faites attention aux titres des graphiques et aux labels des axes, cette recommandation étant valable pour tout le TP). Que constatez vous ?

Partie 1 : Distribution de la Taille.

1. Pour chaque pays, utiliser le choix de classe automatique pour l'histogramme (HISTSELECT2) pour représenter la distribution de la Taille des enfants/jeunes adultes. Par rapport au choix par défaut de R, quelle différence constatez-vous ?
2. Sur les deux graphiques de la question 1, ajouter l'estimation de la densité de la Taille obtenu par Noyau (on utilisera la fenêtre de validation croisée bw.bcv). Commentez les graphiques obtenus. En particulier, comment expliquez-vous le mode très prononcé ?

3. Tracer sur un même graphique les estimations de la densité de la Taille dans les deux pays obtenues par Noyau. Que constatez-vous ?
4. On veut tester l'hypothèse "H_0 : les distributions de la Taille sont identiques dans les deux pays" contre "H_1 : les distributions de la Taille sont différentes dans les deux pays". Quel test peut-on utiliser ? Les résultats du test sont-ils en accord avec ce que vous avez observé en question 3 ?
5. Pour chaque pays, tracer sur un même graphique les estimations de la densité de la Taille des filles et de la Taille des garçons obtenues par Noyau. Que constatez-vous ?
6. Pour chaque pays, tester l'égalité des distributions de la Taille des filles et de la Taille des garçons (on précisera le nom du test utilisé). Les résultats des tests sont-ils en accord avec ce que vous avez observé en question 5 ? Les variables Sexe et Taille sont-elles indépendantes ?

Partie 2 : Estimation de la fonction de régression de la Taille en fonction de l'Age.

1. Pour chaque pays, tracer sur le nuage de point le régressogramme et la courbe de régression de la Taille en fonction de l'Age (on fixera le nombre de classes à 25). Commenter.
2. Tracer sur un même graphique les deux courbes de régression obtenues en question 1. Que constatez-vous ?
3. Pour chaque pays, tracer sur un même graphique les courbes de régression de la Taille des filles en fonction de l'Age et de la Taille des garçons en fonction de l'Age (on fixera le nombre de classes à 25). Que constatez vous ?
4. Sur les deux graphiques de la question 3, ajoutez l'ajustement polynomial de degré 6 (pour expliquer la Taille en fonction de l'Age) que vous obtenez par la méthode des moindres carrés. Que constatez-vous ?
5. Pour chaque pays : les variables Age et Taille semblent-elles indépendantes ? Les résultats du test d'indépendance de deux variables continues sont-ils en accord avec cette observation ? On donnera aussi les résultats du test de corrélation de Pearson du package robusTest.
6. Reprendre la question 1, en effectuant une selection automatique de partition avec Capushe (on prendra une partition maximale de taille $N = 100$). Indiquer la taille de la partition obtenue et commenter.