Devoir 1

Consignes : Remettre ce devoir dans la boîte de dépôt sur le Portail, au format pdf. Vous devez également fournir tout votre code R, commenté de façon à ce que je puisse m'y retrouver. Soignez la présentation de votre travail, et notez que la qualité de la langue sera également évaluée.

Question 1 (10 pts)

L'Université du Michigan publiait le 19 mai 2003 la nouvelle suivante : *Student Drug Testing Not Effective in Reducing Drug Use*. On y présentait une étude de 76 000 étudiants répartis dans 722 écoles (497 écoles secondaires et 225 *middle schools*') aux États-Unis [1]. On demandait aux étudiants sélectionnés d'indiquer s'ils consommaient des drogues ou pas. On a ensuite trouvé que parmi les 5623 étudiants des écoles avec des tests de drogue la proportion d'étudiants consommateurs était 37% alors qu'elle était de 36% parmi les 17 437 étudiants des écoles sans test de drogue.

a) Quel est le type de collecte de données dans cette étude ?

b) Les données semblent montrer que tester pour la drogue augmente en fait la consommation dans les écoles. Qu'en pensez-vous ?

Question 2 (10 pts)

Les responsables de la santé d'une région s'intéressaient aux avantages possibles sur la santé d'être propriétaire d'un chien. Ils ont donc fait les deux études suivantes.

Étude 1 : En utilisant les registres de la région, un échantillon aléatoire simple de 50 résidents de la région qui possèdent un chien a été sélectionné et un échantillon aléatoire de 50 résidents de la région qui ne possèdent pas de chien a été sélectionné. La pression artérielle a été mesurée pour chacune de ces 100 personnes.

Étude 2 : Cent résidents de la région qui ne possédaient pas de chien et qui se sont portés volontaires pour participer à l'étude ont été assignés au hasard à l'un de deux groupes. Ceux du groupe 1 ont reçu un chien adulte. Ceux du groupe 2 n'ont pas reçu de chien et se sont engagés à ne pas devenir propriétaires de chien durant l'année suivante. Après un an, la pression artérielle a été mesurée pour chacune de ces 100 personnes.

Dans les deux cas, les fonctionnaires ont conclu à l'aide des données que la pression artérielle moyenne pour les propriétaires de chiens est significativement inférieure à la pression artérielle moyenne pour ceux qui ne possèdent pas de chien.

a) L'une ou l'autre de ces études permet-elle aux autorités du de la région de conclure que la possession d'un chien est la cause de la différence de pression artérielle observée ? Justifiez votre réponse.

b) L'une ou l'autre de ces études permet-elle aux fonctionnaires de la région de généraliser les résultats obtenus à tous les résidents de la région ? Justifiez votre réponse.

1. L'étude en question a été publiée par Yamaguchi R. et al. dans *Journal of School Health*, vol 73, pp 159-164, 2003.

Question 3 (70 pts)

Je vous demande ici de faire une étude de simulation pour tester diverses méthodes de traitement des données manquantes.

Jeu complet :
Simuler 250 observations avec deux variables : Y_1 et Y_2. Les variables Y_1 et Y_2 suivent une loi normale bivariée. La variable Y_1 a une moyenne de 100 et une variance de 169. La variable Y_2 a une moyenne de 12 et une variance de 9. La covariance entre les deux variables est de $19,5$; leur corrélation est donc $0,5$. (Utilisez la fonction `mvnorm` de la librarie `MASS` si désiré.)

On considère trois mécanismes de non-réponse :
- MCAR : Remplacer au hasard 50% des valeurs de Y_2 par NA.
- MAR : Remplacer par NA les valeurs de Y_2 pour la moitié des observations avec les plus petites valeurs de Y_1.
- NMAR : Remplacer par NA les valeurs de Y_2 pour la moitié des observations avec les plus petites valeurs de Y_2.

On considère quatre méthodes de traitement de la non-réponse :
- Analyse des cas complets
- Imputation simple par la moyenne
- Imputation simple par la régression linéaire
- Imputation simple par la régression linéaire, version stochastique

Créez d'abord un jeu de données complet. Puis, appliquez les trois mécanismes de non-réponse pour obtenir trois jeux de données incomplets. Créez ensuite quatre jeux de données imputés pour chaque jeu de données incomplet, en utilisant les quatre méthodes de traitement de la non-réponse. Répétez ces étapes 500 fois.

Présentez dans un tableau la valeur moyenne des 500 estimés des cinq paramètres de la normale bivariée (deux moyennes, deux variances et une corrélation) pour chacune des combinaisons de mécanisme de non-réponse et méthode de traitement de la non-réponse.

Commentez brièvement sur les résultats obtenus. Dans quelles circonstances peut-on obtenir une estimation sans biais du paramètre d'intérêt.

Question 4 (20 pts)

On a dit en classe que l'imputation simple stochastique par la régression mène à une sous-estimation de la variance des estimateurs, même sous MCAR. Illustrez ce fait à l'aide d'un exemple simple.

Aide : Vous pouvez utiliser le jeu de données complet de la question précédente et considérer l'estimation de la moyenne de la variable Y_2. Répétez B fois les étapes suivantes : simuler un jeu de données complet, ajouter de la non-réponse MCAR à Y_2, imputer par régression stochastique les valeurs manquantes, estimer la moyenne de Y_2 du jeu imputé, calculer l'estimateur de la variance de votre estimateur de la variance. Pour cette dernière étape, utilisez simplement s^2/n comme vous le feriez pour un jeu de données complet. Vous pourrez alors comparer vos B estimateurs de la variance de la moyenne estimée de Y_2 à la variance empirique de vos B moyennes estimées pour conclure.