Travaux Pratiques

Master 1 Bioinformatique/Biostatistique - 2019/2020

Enoncé TP - Gestion des données manquantes

PARTIE I

Arthur Chatton (arthur.chatton@etu.univ-nantes.fr)

Note:

1. Les fichiers sont disponibles sur Madoc.
2. Merci d'effacer votre dossier de travail de l'ordinateur à la fin de la séance.
3. Pensez à enregistrer vos données.

Les objectifs des TP sur les données manquantes sont:

1. d'apprendre à manipuler des données manquantes dans R
2. de savoir imputer des données manquantes par des méthodes d'imputation simple (application sur des données fictives d'un essai contrôlé randomisé visant à comparer deux traitements anti-hypertenseur)

1 Manipulation des données manquantes dans R

Les données manquantes dans R sont représentées par NA sans différence pour les valeurs numériques ou les chaines de caractères.

1. Saisir les deux vecteurs suivants:

   num <- c(-10,0,10, 20, NA, 30, 40)
   str<-c("M1","M2",".",NA,"NA")

   Tester la présence de données manquantes dans ces vecteurs avec les 3 commandes suivantes : ==NA, puis is.na(), puis !is.na(). Observer si des erreurs sont générées et le résultat de ces commandes.

2. Remplacer la cinquième valeur du vecteur *str* par une donnée manquante.

3. Identifier où sont placées les valeurs manquantes (identifiants) dans les vecteurs *num* et *str*, en utilisant les fonctions which() et is.na().

4. Observer comment sont traitées les données manquantes avec les fonctions usuelles suivantes.

```
factor(num)
factor(num, exclude=NULL)
table(num)
table(num, exclude=NULL)
summary(num)
mean(num)
mean(num,na.rm=TRUE)
sd(num)
sd(num,na.rm=TRUE)
sum(num)
sum(num,na.rm=TRUE)
```

2 Données HTA

Chez les adultes, l'hypertension artérielle (HTA) se définit comme une pression artérielle systolique (PAS) $\geq$ 140 ou diastolique (PAD) $\geq$ 90 mm Hg pour au moins 2 mesures au cabinet médical distancées de plusieurs minutes et dans des conditions précises (position assise, après 5 minutes de repos,...).
Un essai randomisé a été mis en place pour comparer l'efficacité de deux traitements anti-HTA (Trt1: Nouveau traitement anti-HTA, Trt2: traitement anti-HTA habituellement administré). Pour être inclus, les patients devaient être des adultes de moins de 80 ans, suivis dans l'un des 10 cabinets participant à l'étude, avec un diagnostic d'HTA établi lors de la visite d'inclusion (M0), et n'étant pas sous traitement anti-HTA lors de l'inclusion. Une fois son consentement signé, le patient était randomisé dans l'un des bras (tirage aléatoire pour l'un des deux antihypertenseurs). Il était prévu d'inclure 330 patients (165 dans chaque bras) puis d'évaluer leur tension artérielle 3 mois puis 6 mois après le début du traitement.
Pour juger de la supériorité d'un traitement par rapport à l'autre, trois critères seront comparés: l'évolution de la PAS entre l'inclusion et après 6 mois de traitement, l'évolution de la PAD entre l'inclusion et après 6 mois de traitement, le diagnostic d'HTA à 6 mois. A l'inclusion, les critères suivants étaient également recueillis: âge (en années), sexe (1=Homme/0=Femme), fumeur (1=oui/0=non), taille (en cm), poids (en kg). Les variables premat et premat.cause renseignaient si le patient était sorti prématurément de l'étude c'est-à-dire avant 6 mois (1=oui/0=non) et la cause en cas de sortie prématurée. Les variables "pas.M0","pad.M0","pas.M3","pad.M3","pas.M6","pad.M6" correspondent aux mesures de pression artérielle aux différentes visites (moyenne de deux mesures en position assise).

1. Importer les données "donneesinit" sous R.

2. Calculer la variable "bmi" correspondant au poids divisé par la taille au carré (taille en mètre), et les variables "hta.M0" et "hta.M6" correspondant au diagnostic d'HTA à l'inclusion et 6 mois après le début du traitement (1=oui/0=non).

3. Supprimer tous les sujets ne vérifiant pas les critères d'inclusion.

4. On s'intéresse au taux de données manquantes pour chaque variable. Calculer le % de sujets sans mesure manquante à M3. Calculer le % de sujets sans mesure manquante à M6. Calculer le % de sujets sans mesure manquante au cours du suivi.

5. Y-a-t il des sujets avec des mesures manquantes à M3 mais pas à M6 (structure des données manquantes = intermittente)? Y-a-t il des sujets sans mesures après M0 ou après M3 (sorties prématurées) ? Observer les causes de sorties prématurées.

6. Créer les variables pas.M6.missing, pad.M6.missing et hta.M6.missing codant si les mesures de PAS, de PAD et du diagnostic d'HTA sont manquantes à M6 (1=oui,0=non).

7. Créer les variables diffpasM3 et diffpadM3 calculant la différence M3-M0 pour la PAS et la PAD. Créer les variables diffpasM6 et diffpadM6 calculant la différence M6-M0 pour la PAS et la PAD.

8. Effectuer la régression logistique de présence de données manquantes à M6 pour la PAS en fonction des valeurs de PAS à l'inclusion. Effectuer la régression logistique de présence de données manquantes à M6 pour la PAS en fonction de l'évolution des mesures entre l'inclusion et M3. L'hypothèse de données manquantes complètement aléatoirement (MCAR) pour la PAS à M6 est-elle rejetée ? Quelles sont les limites de ces tests? Faire le même raisonnement pour la PAD.

9. Calculer les moyennes de PAS ainsi que les IC95% dans chacun des groupes à M0 et à M6.