Analyse de données médico-administratives avec R: Exploration des données OpenCCAM de Scan Santé

Nolwenn Le Meur - EHESP

Sequence 1 - Mai 2023

Vous vous intéressez aux procédures qui peuvent être effectués en ambulatoire ou en chirurgie conventionnelle.

Vous disposez d'un premier d'un jeu données qui liste par établissement de soins les actes réalisés en France en 2020, le nombre d'actes, la DMS, l'activité ambulatoire, le département et la région. Ces données sont issues de ScanSante et librement disponibles sur https://www.scansante.fr/open-ccam/open-ccam-2020

I. Lecture et nettoyage des données

Dans un premier temps, il vous est demandé de lire les données, de vérifier leur qualité, et de sélectionner votre population d'intérêt.

a. Lire le fichier *"Open_ccam_20.csv"* avec la fonction *read.csv()* dans R pour créer l'objet R *ccam20* et vérifiez le type des variables. Attention les variables sont séparées pas des ";" et les données "manquantes" sont notées "." .

b. Utliser la fonction skim() de la librarie skimr pour un premier diagnostic.

c. Que font les lignes de codes ci-dessous? Executez pas à pas les lignes pour mieux les comprendre.

```
## region et département
ccam20$reg <- as.factor(ccam20$reg)
ccam20$dep <- as.factor(ccam20$dep)
# acte
ccam20$acte <- substr(ccam20$acte, 1, 7)
ccam20$acte <- as.factor(ccam20$acte)

# FINESS
ccam20$FinessGeo<- ifelse(nchar(ccam20$FinessGeo)==8,
                          paste("0",ccam20$FinessGeo,sep=""), ccam20$FinessGeo)
```

Vous souhaitez faire le focus sur un groupe de procédures. Nous vous proposons de travailler sur les actes endovasculaires mais vous pouvez utiliser votre propre liste d'actes réalisés en ambulatoire et en conventionnel.

d. Vous devez pour ce faire lire le fichier "liste_ccam2020.csv" (séparateur de colonnes ",") qui comporte les actes endovasculaires d'intérêt.

e. Puis vous devez sélectionner dans l'objet *ccam20* les lignes correspondant aux actes endovasculaires que vous venez de lire pour créer une sous table *endo*. Executez la ligne de commande ci-dessous pour comprendre le rôle de la function *%in%*.

```
# Sélection des actes endovasculaires dans la base ccam20
endo <- ccam20[ccam20$acte%in%ccam_endo$CDC_ACT, ]
```

f. Essayez de trouver une manière de faire la même chose avec la fonction filter de librairie *dplyr*

g. Sauvegarder cette nouvelle base de donnée au format *.Rdata avec la fonction save()

II. Statistiques descriptives

Pour l'exercice je vous invite à suavegarder votre script R et fermer votre application pour la réouvrir. Vous redémarrez ainsi comme après un long week-end…

Votre base de données *endo* n'est sans doute pas chargée dans votre environemment, vous devez l'importer avec la fonction *load()*.

1. Quels établissements sont au-dessus de la DMS nationale pour les actes d'intérêts?

a. Calculez les indicateurs statistiques usuels pour la DMS en utilisant la fonction *summary()* et la fonction *ci()* de la library *epiDisplay*. Commentez les résultats.

b. Calculez les indicateurs statistiques usuels pour la dms par région en utilisant la fonction *by()* ou la syntaxe de *dplyr*.

c. Observez l'utilisation de la fonction *ifelse()* ci-dessous. Que fait cette ligne de commmande.

```
endo$Top <- ifelse(endo$dms_globale > endoci$upper95ci, 1, 0)
```

e. Avec la function *table()* créer la table de contingence qui dénombre les établissements au-dessus la borne supérieure de l'intervalle de confiance de la DMS nationale pour 1 acte dans chaque région.

f. Utilisez la fonction *prop.table()* pour obtenir la proportion de ces établissements par région.

2. Quel est le mode de financement des établissements qui dépassent la DMS nationale pour les actes endovasculaires?

Pour obtenir le mode de financement des hopitaux (mft - mode fixation tarifs) nous devons apparier nos données CCAM aux données du répertoire Finess. Ces informations sont librement disponibles sur le site Open-data.gov.

(source: https://www.data.gouv.fr/fr/datasets/finess-extraction-du-fichier-des-etablissements/)

a. Lire le fichier *"etalab_stock_et_20201231.csv"* dans R pour créer l'objet R *finess*

b. Vérifier le type de la variable *mft* qui code le mode de financement des établissements.

c. Vérifier la longueur des codes *FinessGeo* dans les 2 bases à fusionner

d. Apparier les données CCAM pour les actes endovasculaires aux données Finess (variables *nofinesset* et *mft* uniquement) dans un objet nommé *endofiness* grâce à la fonction *merge()*.

```
## Jointure entre base avec clé de jointure de nom différent
## Si les 2 clés de jointures ont le même nom pas besoin de by.x et by.y mais juste by
endofiness <- merge(endo, finess[, c("nofinesset", "mft")], by.x="FinessGeo",by.y="nofinesset", all.x=TR
## colnames(endofiness)
```

e. Trouver une autre façon de le faire avec la librarie dplyr et sa syntaxe

f. Que fait le code chunk suivant?

```
# useNA = "always pour visualiser si il y a de NA
table(endofiness$mft, useNA = "always")
levels(endofiness$mft)
endofiness$mft <- droplevels(endofiness$mft)
table(endofiness$mft, useNA = "always")
endofiness$mft <- factor(endofiness$mft,
                         labels=c("public", "public","non lucratif", "privé",
```

```
                                    "non lucratif", "privé",
                                    "public","privé","indéterminé"))
```

g. Utilisez les fonctions *table* et *prop.table()* pour obtenir le dénombrement et la proportion des établissements au-dessus de la DMS nationale (variable *Top*) par type de financements (variable *mft*).

3. Est-ce que les établissements dépassant la dms nationale pour les actes endovasculaires proposent aussi cette intervention en ambulatoire?

a. Créez une variable binaire *Ambu* dans le table *endofiness* avec la modalité 1 lorsque l'établissement présente des informations dans la variable *nb_sej_0_nuit* et 0 autrement.

b. Utilisez les fonctions *table* et *prop.table()* pour obtenir le dénombrement et la proportion des établissements au-dessus de la DMS nationale pour 1 acte (variable *Top*) selon la présence ou nom d'une activité ambulatoire pour l'acte.

III. Statistiques inferentielles (optionel)

1. Analyses univariées

a. Réalisez un test de Chi2 ou un test Exact de Fisher pour tester l'hypothèse d'indépendence entre dépassement de la DMS nationale et la pratique de cette intervention en ambulatoire.

Vous vous demandez si la DMS des établissements est fonction de l'activité soit la variable *nb_actes*?

b. Créez la variable binaire *nbactes_eleve* sur la base de la médiane de la distribution des *nb_actes* des établissements.

c. Calculez les indicateurs statistiques usuels pour la DMS selon la variable *nbactes_eleve*.

d. Testez l'hypothèse de l'indépendence entre la DMS et un nombre d'actes élevés avec le test statistique adapté. Interpréter les résultats.

2. ANOVA and Co

Vous vous interrogez ensuite sur l'existence de différences de DMS entre établissements aux modes de financements différents.

a. Calculez la moyenne des DMS par mode de financement.

b. Appliquez le test statistique adapté pour explorer l'hypothèse d'une indépendence entre mode de financement et DMS. Vous pouvez utiliser la librarie *rstatix* pour une variété de tests.

3. Régressions

A partir de vos données uniquement, vous chercher ensuite à expliquer les raisons possibles d'un dépassement de la DMS nationale.

a. Utilisez la méthode *glm()* pour réaliser une régression logistique et explorer les possibles liens entre DMS au dessus de la DMS nationale (variable *Top*) et les variables *nb_actes*, *mft*, *Ambu*.

b. Utilisez la fonction *logistic.display()* de la libary *epiDisplay* pour un affichage synthetique des OR.

IV. Représentations graphiques et cartes

R dispose de nombreuses fonctionnalités graphiques de base mais la libraire *ggplot2* s'impose aujourd'hui grâce à l'optimisation des paramètres pour un respect des règles de technique de visualisation.

1. Graphiques

a. Utilisez la fonction *boxplot()* pour représenter les distributions des DMS par région avec des boîtes à moustaches

b. Utilisez la library *ggplot2* pour représenter les distributions des DMS par région avec des boîtes à moustaches

c. Représentez le nombre d'actes en fonction de la DMS avec la syntaxe graphique incluant *geom_point()*

d. Représentez le nombre d'actes en fonction de la DMS et colorez les points selon les statuts juridiques

e. Utilisez une représentation en boîtes à moustaches pour représenter les distributions des DMS par région et par mode de financement en utilisant la syntaxe *facet_wrap()*

f. Représentez les distributions des DMS par mode de financement en utilisant *geom_density()*

Extra: Diagramme en barre de la répartition des établissements selon leur mode de financement

2. Cartes

Vous souhaitez représenter les DMS par région.

a. Calculez la moyenne des DMS par département et stocker le résultat sous la forme du data.frame du nom de *dmsdep*

Pour construire une carte vous devez obtenir les coordonnées géographiques. Vous utilisez les library *ggmap* et *maps*.

b. Installer et charger ces libraries dans R puis avec la fonction *map_data()* charger la carte de France dans un objet R nommé *france*

c. Vous avez besoin de la correspondance entre les numéros de départements et leurs libellés disponibles dans le fichier *dep2reg.csv*. Lire le fichier dans R et nommer votre objet *dep2reg*.

d. Effectuer les jointures entre les tables *dmsdep* et *dep2reg* puis avec la table *france* pour créer l'objet *dmsdep2map*

e. Executer le code ci-dessous et commenter les résultats

```
dmsdep2map <- dmsdep2map[order(dmsdep2map$order), ]
ggplot(data = dmsdep2map) +
   geom_polygon(aes(x = long, y = lat, fill = dms, group = group)) +
  scale_fill_gradientn(colours =heat.colors(7, alpha=0.8, rev = T))
   coord_fixed(1.3)
```