Cytomégalovirus humain en Ouganda

Projet de biostatistiques L3

Guillaume Louvel

Contexte

Le cytomégalovirus humain, (HCMV) ou *Human Herpesvirus 5* est un virus de la famille des herpès virus très fréquent dans la population humaine. Il est transmis par contact via sécrétions corporelles. Il passe le plus souvent inaperçu mais peut être responsable de surdité ou cécité chez les enfants. Chez les adultes, certaines études ont détecté des liens entre HCMV et risques cardio-vasculaires, ainsi que des complications telles que des rétinites en cas de co-infection avec le VIH. Des études récentes suggèrent également une susceptibilité accrue à la tuberculose. Ce TP est basé sur le jeu de données généré par Stockdale et al. 2018, légèrement modifié pour des besoins pédagogiques.

Objectif de l'étude

Pour comprendre par quels facteurs le HCMV peut être responsable d'une mortalité accrue, on veut caractériser l'interaction du HCMV avec d'autres infections transmissibles comme le VIH et la tuberculose, ainsi qu'avec des facteurs de risques cardiovasculaires.

Échantillon

L'échantillon observé est une cohorte suivie en Ouganda, dont 2174 individus ont été retenus dans cette étude. Le taux d'immunoglobuline anti-HCMV a été mesuré et la séropositivé établie. De même on dispose des résultats de tests à la tuberculose effectués entre 2011 et 2014, et des résultats du test du VIH. Les mesures cardio-vasculaires (pression, rythme cardiaque, cholesterol. . .) ont été obtenues pour les individus de plus de 12 ans.

1 Étude descriptive

Télécharger le tableau de données : www.normalesup.org/~glouvel/data/projet_HCMV.csv

1.1 Description des variables

Nom	Définition
`sex`	Sexe des individus (valeurs [`"Male"`/`"Female"`]).
`cmv`	Dosage d'anticorps au HCMV.
`age`	Âge en année.
`cmvstatus`	Séropositivité au HCMV (Valeurs [`"positive"`/`"negative"`]).
`hiv`	Séropositivité au HIV.
`TB`	Résultat du test salivaire de la tuberculose.
`m_syst`	Durée moyenne de systole.
`m_diast`	Durée moyenne de diastole.
`R22_bpgroup`	Hypertension.
`BMI`	Indice de Masse Corporelle (IMC).
`bmi`	Catégorisation des individus en 4 groupes d'IMC :

Nom	Définition
	$< 18,4$: "underweight", $18,5 - 24,9$: "normal", $25 - 29,9$: "overweight", > 30 : "obese", selon l'OMS.
`cholesterol`	Taux sanguin de cholestérol.
`high_d_lipid`	Taux sanguin de HDL (Lipoprotéines de haute densité).
`low_d_lipid`	Taux sanguin de LDL (Lipoprotéines de basse densité : risque cardiovasculaire).
`hba1c`	Taux d'hémoglobine glyquée (marqueur de diabète).
`cmv_tertile`	Catégorisation en 3 quantiles de `cmv`.

1.2 Introduction, lecture et description du fichier de données

1. Charger le tableau de données `hcmv.csv` et le nommer `df`. Vérifier que l'objet produit (un `data.frame`) est correct en affichant le début du tableau. Quelles sont ses dimensions ? [`read.csv`, `head`, `dim`, `help`]
2. Afficher le nom et la classe de chacune des variables présentes dans le *data frame* `df`. [`data.class`, `$`, `sapply`]

1.3 Description des variables qualitatives

1. Quelles sont les variables qualitatives contenues dans `df` ?
2. Considérons les variables `sex`, `hiv`, `cmvstatus` et `TB` ? Donner les effectifs des différentes observations. Représenter ces distributions sous forme d'un diagramme en bâton. [`table`, `plot`]

1.4 Description des variables quantitatives

1. Quelles sont les variables quantitatives contenues dans `df` ?
2. Quelles sont les informations que l'on peut obtenir sur chacune de ces variables ? [`summary` et beaucoup d'autres. . .]
3. Quelles variables quantitatives contiennent des valeurs manquantes ? Combien ? [`is.numeric`, `is.na`]
4. Quel type de variable est `cmv_tertile` ? Quels *niveaux* peut prendre cette variable ? [`is.factor`, `levels`]
5. Considérons la variable `age`. Quel est son type ? Pour l'étude, nous allons grouper les âges. Créer une variable `age_cat` qui représente les catégories d'âges avec les seuils suivants (borne droite exclue) : $0, 2, 4, 6, 11, 16, 21, 31, 41, 51, 61$. [`cut`, `as.factor`]

Représentations graphiques

6. Regarder les distributions de `cmv` et `BMI` avec une représentation en histogramme. Superposer à ces histogrammes, la courbe de la densité de probabilité d'une loi normale (ses paramètres seront estimés sur l'échantillon considéré). Attention, nous allons superposer une densité de probabilité à un histogramme qui propose par défaut des effectifs. Il y a donc conflit et une transformation s'impose lors de la construction de l'histogramme. Consulter l'aide des fonctions si besoin. [`hist`, `seq`, `dnorm`, `lines`]
7. La variable `m_syst` est-elle normalement distribuée ? Que pensez-vous de sa distribution après passage au logarithme ?
8. Afficher les graphes "x,y" de chaque paire possible de variables quantitatives [`pairs`]. Quelles informations cela vous donne-t-il ?

1.5 Croisement de deux variables

1. Séparer les valeurs de la variable [`cmv`] pour chaque valeur de la variable `sex`. Afficher la nouvelle variable et représenter ces distributions sous forme d'histogramme. [`split`]

2. Que semble suggérer un affichage en boîte à moustache des valeurs de ces 2 groupes ? [`boxplot`]

2 Tests statistiques

2.1 Tests d'indépendance

1. Pour comparer les répartitions d'individus dans les catégories de `cmvstatus` et `hiv`, construire la *table de contingence* de ces 2 variables. [`table`]
2. Y a-t-il une interaction entre la séropositivité au HCMV (`cmvstatus`) et l'infection par le VIH ? [`chisq.test`]
3. Que se passe-t-il lorsque l'on considère cette fois le *niveau* d'anticorps au HCMV (`cmv_tertile`) ?
4. Y a-t-il une interaction entre l'infection au HCMV et l'infection à la tuberculose ? [`fisher.test`]

2.2 Comparaison de deux distributions

1. Faire une analyse de la moyenne de la variable `cmv` chez les hommes et les femmes (tracer les distributions, calculer la moyenne). [`split`] On peut faire selon la variable `hiv` aussi mais attention les effectifs sont très déséquilibrés.
2. Réaliser un test de comparaison de ces deux distributions. Quel test choisissez-vous, et pourquoi ?

2.3 Comparaison de plusieurs distributions

1. On s'intéresse maintenant au lien entre HCMV et risques cardiovasculaires. Pour cela on *sélectionne* uniquement les individus *de plus de 12 ans* (exclus), et uniquement ceux *qui n'ont pas le VIH* (connu pour augmenter ces risques).
   Étudier la corrélation entre la variable `cmv_tertiles` et les variables liées aux risques cardiovasculaires : `m_syst`, `m_diast`, `R22_bpgroup`, `BMI`, `bmi`, `hba1c`. Choisir le test adapté selon la nature des variables considérées. [`aov`, `chisq.test`, `kruskal.wallis`, `fisher.test`, `cor.test`]
2. Refaire ces tests, mais pour chaque catégorie d'âge (utiliser la nouvelle variable `age_cat` séparant les individus par paliers).
3. Que se passe-t-il lorsque que l'on a pris en compte l'âge et le sexe ?
4. Quel est le problème avec la stratégie proposée dans les questions 2 et 3 ci-dessus ?

2.4 Question bonus

1. Que dire de l'effet combiné du HIV et de la tuberculose sur le taux d'anticorps au HCMV ?

Évaluation

Les questions ci-dessus doivent vous permettre de vous familiariser avec le jeu de données mais vous ne serez pas noté sur vos réponses à chacune de celles-ci ;

l'évaluation se fera sur une seule analyse, vous devrez donc :

— choisir ***une* question d'intérêt** (parmi les questions ci-dessus ou non),
— l'analyser et proposer une réponse,
— et surtout effectuer une **synthèse**.

Vous devrez rendre un rapport court (2 pages maximum), ainsi que le script `.R` avec les commandes utilisées. Vous présenterez dans un deuxième temps votre analyse sous forme synthétique à l'oral.

Références

Stockdale, Lisa et al. (2019). Data from : *Human cytomegalovirus epidemiology and relationship to tuberculosis and cardiovascular disease risk factors in a rural Ugandan cohort* [Dataset]. Dryad. https://doi.org/10.5061/dryad.d1k17