TP : Régression linéaire

Exercice 1 : Clinique

Explications

Ce premier exemple est inspiré d'un projet que nous avons eu, les données sont simulées.

Contexte :

La cohorte dont nous disposons est constituée de 600 patients atteints de la maladie de Parkinson, tous porteurs d'une mutation spécifique parmi un total de six mutations différentes. Dans le cadre de notre étude, nous suivons ces patients et mesurons leur score UPDRS, un indicateur cognitif largement utilisé pour évaluer la sévérité de la maladie de Parkinson. Nous avons également recueilli d'autres informations, notamment la durée de la maladie, l'âge au moment du diagnostic, le sexe, et l'âge au moment de la participation à l'étude.

Problématiques :

Critère de jugement principal :
Nous visons à comprendre les facteurs qui influencent l'UPDRS (Unified Parkinson's Disease Rating Scale) qui est indicateur de la gravité de la maladie de Parkinson et à éventuellement le prédire en utilisant des variables telles que la durée de la maladie (disease duration). Nous avons des raisons de croire que la gravité de la maladie varie en fonction des mutations spécifiques (mutation senescence vs non-senescence), et nous prévoyons de prendre en compte cette variation. De plus, nous supposons que l'âge au moment de la participation à l'étude peut influencer l'UPDRS, et nous avons l'intention de corriger cette influence. En outre, il est possible que le sexe des patients ait également un impact, que nous examinerons.

Critère de jugement secondaire :
Nous cherchons également à évaluer dans quelle mesure la mutation peut influencer l'âge auquel le diagnostic de la maladie de Parkinson est posé et le moment où la maladie se manifeste.

Description de la base de donnée :

Le jeu de données se compose de 600 patients, et les variables incluses sont les suivantes :

- "Mutation" : indiquant la mutation spécifique du patient, avec un total de 6 modalités : "m/s", "f/s", "s/s", “f/f”, “m/m”, “f/m”.

- "Age_at_onset" : représentant l'âge au moment du diagnostic.
- "Sexe" : spécifiant le sexe du patient.
- "Disease_duration" : décrivant la durée de la maladie au moment de l'étude.
- "Âge" : indiquant l'âge au moment de la participation à l'étude.
- "UPDRS" : reflétant la valeur du score UPDRS, un indicateur de la sévérité de la maladie de Parkinson.

Questions :

Data management :

1. Importer le jeu de données "Exercice1.txt" dans R et afficher les premières lignes du jeu de données pour avoir un aperçu de son contenu.
   Indication : utiliser les fonctions read.table, head

2. Après avoir importé le jeu de données, assurez-vous que les variables "Sex" et "Mutations" sont correctement reconnues comme des facteurs.

3. Créez une nouvelle variable appelée "Mutation_bis" qui comporte deux catégories : "senescence" et "non-senescence". La catégorie "senescence" regroupe les modalités "m/s", "f/s" et "s/s", tandis que la catégorie "non-senescence" regroupe "m/m", "f/f" et "m/f" de la variable “Mutation”.

Statistiques univariés :

4. Créez des histogrammes pour les variables UPDRS, Âge, Âge au moment du diagnostic (Age_at_onset), et Durée de la maladie (Disease_duration). Commentez la forme des distributions.
   Indication : utiliser geom_histogram

5. Créez un barplot pour la variable 'Sex’.
   Indication : utiliser geom_barplot

6. Calculez la moyenne et l'écart-type pour les variables numériques, ainsi que les pourcentages pour les variables catégorielles, et présentez ces mesures dans un tableau de données.
   Indication : utiliser la fonction table1 du package furniture

Statistiques bi-variés :

7. Créez la même table que précédemment, mais cette fois en segmentant les résultats en fonction de la variable “Sex”.
   Indication : utiliser les options splitby et test de la fonction table1 du package furniture

8. On souhaite déterminer si la durée de la maladie (variable 'disease_duration') varie en fonction du type de mutation (variable Mutation).
    - Écrivez l'équation du modèle correspondant.
    - Lancer ce modèle en R. *Indication utiliser la fonction lm, summary, Anova.*
    - Faire une figure afin de représenter les résultats. *Indication utiliser geom_boxplot et stat_compare_means.*

9. On souhaite déterminer si l'âge au moment du diagnostic (variable 'Age_at_onset') varie en fonction du type de mutation (variable Mutation).
    - Écrivez l'équation du modèle correspondant.
    - Lancer ce modèle en R. Indication utiliser la fonction lm, summary, Anova.
    - Faire une figure afin de représenter les résultats. *Indication utiliser geom_boxplot et stat_compare_means*.
    - Effectuer des comparaisons post hoc pour identifier les différences d'âge au moment du diagnostic entre les différentes mutations. *Indication : utiliser le package emmeans*
    - Répondez au critère secondaire.

10. On souhaite déterminer si l'UPDRS (variable 'UPDRS') varie en fonction du type de mutation (variable Mutation).
    - Écrivez l'équation du modèle correspondant.
    - Lancer ce modèle en R. Indication utiliser la fonction lm, summary, Anova.
    - Faire une figure afin de représenter les résultats. Indication utiliser geom_boxplot et stat_compare_means.
    - Effectuer des comparaisons post hoc pour identifier les différences UPDRS au moment du diagnostic entre les différentes mutations. Indication : utiliser le package emmeans.

<u>Régression linéaire simple :</u>

11. En utilisant le critère principal, identifiez les variables explicatives et déterminez la variable à expliquer.

12. Effectuez une régression linéaire simple entre UPDRS et l'âge
    - Lancer le modèle en R. *Indication : utiliser la fonction lm*
    - Analysez la sortie du modèle, notamment les coefficients de la droite de régression, l'ordonnée à l'origine, le R^2. En fonction des résultats, évaluez s'il existe une association significative entre ces deux variables.
      Indication : Utiliser la fonction summary et Anova pour obtenir ces informations.

- Faire une figure afin de représenter les résultats. Indication : *geom_point, geom_smooth, stat_summary.*

13. Effectuez une régression linéaire simple entre UPDRS et la durée de la maladie (Disease_duration)
    - Lancer le modèle en R. *Indication : utiliser la fonction lm*
    - Analysez la sortie du modèle, notamment les coefficients de la droite de régression, l'ordonnée à l'origine, le R^2. En fonction des résultats, évaluez s'il existe une association significative entre ces deux variables.
      Indication : Utiliser la fonction summary et Anova pour obtenir ces informations.
    - Faire une figure afin de représenter les résultats. *Indication : geom_point, geom_smooth, stat_summary.*

<u>Régression linéaire multiple :</u>
Réponse au critère de jugement principal.

14. Écrire l'équation qui correspond au critère de jugement principal.

15. Créez une représentation graphique du modèle (UPDRS en fonction de Disease_duration), en utilisant ggplot. Divisez la variable 'Age' en catégories basées sur ses quantiles en ‘Age_discrete’, utilisez 'geom_point' et 'geom_smooth', segmentez le graphique Age avec 'facet_wrap(~Age_discrete)', et attribuez des couleurs différentes en fonction de 'Mutation_bis' en utilisant l'option 'color’.

16. Lancer le modèle en R à l’aide de la fonction lm.

17. Analysez la sortie du modèle, notamment les coefficients de la droite de régression, l'ordonnée à l'origine, le R^2. Interpréter le modèle.
    Indication : Utiliser la fonction summary et Anova pour obtenir ces informations.

18. Effectuer une analyse des contrastes afin d’évaluer l'impact des différentes mutations sur notre variable continue, nous souhaitons effectuer des comparaisons post hoc tout en prenant en compte une variable continue.
    Indications : utiliser le package emmeans

19. Créez une figure synthétique qui résume notre modèle.
    Indication : utiliser le package interactions et la fonction interact_plot

20. Effectuez une vérification des hypothèses de validité du modèle.
    Indication : utiliser le package ggResidpanel et la fonction resid_panel

21. La colinéarité dans une régression linéaire multiple se produit lorsque deux ou plus de deux variables indépendantes dans le modèle sont fortement corrélées, ce qui rend difficile la distinction de leur impact individuel sur la variable dépendante. Cela peut entraîner une instabilité dans les estimations des coefficients de régression.
    Vérifier la colinéarité.
    Indication : Utiliser la fonction vif du package car

Prédiction :

22. Utiliser le modèle que vous avez construit afin de prédire sur les mêmes données la valeur d'UPDRS.
    Indications : Utiliser la fonction predict.