Imen RACHED	Mathematics For Data Sciences Examen TP 2023-2024 ADIF72-TP	17/10/2023 ADIF72-TP 23-24-S7

IMPORTANT

Lisez attentivement les instructions avant de commencer votre travail

Instructions:

— Ce devoir peut être fait seul OU en binôme.

—Dans toutes vos réponses, utilisez trois chiffres significatifs.

- Le rapport

Une attention particulière sera portée à la présentation du rapport final. Le rapport peut être rédigé en anglais ou en français. Dans tous les cas, la qualité de l'écriture sera appréciée. **Les phrases doivent être clairs, explicites et bien compréhensibles.**

Le rapport doit contenir une page de couverture, une table des matières, une introduction, le corps du rapport (code, résultats, figures, tableaux, interprétations, commentaires, etc.), la conclusion et l'annexe pour le code. Le nombre de pages ne doit pas dépasser 10 pages (annexe non incluse). Vous pouvez utiliser un Rmarkdown mais dans ce cas le rendu doit être (le Rmarkdown et le rapport en pdf).

La page de couverture doit contenir le prénom, le nom et le numéro d'identification de l'étudiant de tous les auteurs.

- Remise du rapport

Vous trouverez une boîte de dépôt nommée TP noté dans Moodle. Le rapport final au **format pdf** doit être téléchargé dans ce coffre au plus tard le 19/10/2022 à 23h55. Vous pouvez télécharger plus de 2 fichiers (si vous souhaitez inclure le script de code, **ne les compressez pas** tous dans un seul fichier au format zip. Les noms de fichiers doivent être les suivants :

NomPrénomÉtudiant1 NomPrénomÉtudiant2.pdf.

Un seul dépôt par groupe doit être effectuée !

---

1. Analyse des données

1.1. Description du jeu de données

L'ensemble de données prostate.txt contient des informations sur les patients atteints d'un cancer de la prostate. La prostate est une glande du système reproducteur masculin. Le cancer se développe à partir des tissus de la prostate lorsque les cellules mutent et se multiplient de manière incontrôlable. Celles-ci peuvent ensuite se propager (métastaser) en migrant de la prostate vers d'autres parties du corps. Comme les différents types de cancer, plus il est détecté tôt, plus le patient a de chances de guérison. L'antigène spécifique de la

prostate est une protéine normalement sécrétée par les cellules de la prostate, mais une cellule cancéreuse sécrète 10 fois plus qu'une cellule normale. Espérons que cette propriété pourrait être utile pour détecter le cancer de la prostate. Cependant, le niveau de l'antigène spécifique de la prostate peut être augmenté par de nombreux autres facteurs tels que le volume de la prostate, les infections et/ou l'inflammation) voire diminué par certains traitements.

L'objectif de cette étude est de mieux comprendre les facteurs influençant le taux d'antigène prostatique spécifique. Une étude a été menée sur des hommes atteints d'un cancer de la prostate et ayant subi une prostatectomie radicale, c'est-à-dire une ablation chirurgicale complète de la prostate. Avant la chirurgie, le niveau d'antigène spécifique de la prostate a été déterminé par un test sanguin. Les tissus prélevés lors de l'opération ont été examinés plus précisément afin de caractériser le cancer.

Le jeu de données prostate.txt contient les variables suivantes :

— volume du cancer (vol).

— poids de la prostate (wwt).

— âge du patient (âge).

— quantité d'hyperplasie bénigne (bh). L'hyperplasie bénigne est une tumeur bénigne de la prostate.

— propagation dans les vésicules séminales (vs).

— pénétration capsulaire (pc). Plus la variable est élevée, plus le cancer traverse la capsule entourant la prostate pour atteindre les structures voisines.

— taux spécifique d'antigène de la prostate (psa).

La variable vs est binaire, les autres sont des variables réelles.

1.2. Première analyse : description du jeu de données

Importez le jeu de données prostate.txt et nommez-le prostate.

Familiarisez-vous avec les données en effectuant les tâches suivantes et répondre aux questions :

1. Combien y a-t-il d'observations ? Présentez les données et l'objectif de l'étude statistique.
2. Utilisez la commande summary() pour calculer des statistiques descriptives pour la variable cible psa. Interprétez les résultats.
3. Effectuer une première analyse de la variable psa basée sur les autres en calculant le coefficient corrélation entre psa et chacune des autres variables. Laquelle est la plus corrélée avec psa ?

En traçant les nuages de points entre psa et chacune des autres variables à l'aide de la commande plot(prostate) vous remarquerez que ces plots affichent de grandes accumulations de points dans de petites zones.

Pour résoudre ce problème, il est habituel de considérer le logarithme plutôt que les valeurs d'origine.

4. Effectuez une transformation logarithmique de tous les ensembles de données de variables à l'exception de l'âge et modifiez les noms des variables transformées en faisant précéder le nom de la lettre l (exemple lvol au lieu de vol). Visualisez à nouveau le Nuages de points et interpréter les résultats.

Attention : Par la suite, toute l'analyse sera effectuée en utilisant les valeurs transformées.

2. Analyse en composantes principales (ACP)

2.1. Question théorique :

Si deux variables sont parfaitement corrélées dans le jeu de données, serait-il approprié de les inclure toutes les deux dans l'analyse lors de la réalisation de l'ACP ? Justifiez votre réponse.

2.2. Application pratique : La fonction apply() permet d'appliquer une fonction à chaque ligne ou colonne du jeu de données. Pour exemple, la commande apply(prostate, 2, mean) permet de calculer la moyenne empirique de chaque variable. Calculez la variance de chaque variable et interprétez les résultats.

2.3. Pensez-vous qu'il est-il nécessaire de normaliser les variables avant d'effectuer l'ACP pour ce jeu de données ? Pourquoi ?

Pour réaliser l'ACP, la fonction PCA peut être utilisée par lignes de commandes à l'aide du package FactoMineR ou grâce à une interface graphique via le package Factoshiny.

2.4. Effectuez l'ACP à l'aide de la fonction PCA() avec les arguments et options appropriés en tenant compte de votre analyse précédente. Analysez la sortie de cette fonction.

2.5. Interpréter les valeurs des deux premiers vecteurs de chargement des composantes principales ?

Plusieurs solutions existent pour déterminer le nombre d'axes à analyser en ACP. La plus courante consiste à représenter le diagramme en barres des valeurs propres ou des inerties associées à chaque axe Grâce à la fonction barplot.

2.6. Tracez le PVE expliqué par chaque composant, ainsi que le PVE cumulé. Calculer le pourcentage de variance expliquée (PVE) par chaque composant ?

2.7. Combien de composants garderiez-vous ? Pourquoi ?

3. Régression linéaire

3.1. Question théorique : Supposons que l'on fait un modèle de régression linéaire simple pour expliquer Y comme une fonction linéaire de X.

Quelle est la relation entre, la corrélation coefficient entre ces deux variables r (X; Y) et le coefficient de détermination R^2 obtenu par adapter le modèle ? Quelle est la plage de valeurs que peut prendre r ?

Application pratique

3.2. Calculez la corrélation entre la variable lpsa et les autres variables existant dans le jeu de données.

Notons X la variable la plus corrélée avec lpsa et considérons le modèle de régression linéaire simple suivante :

$$\texttt{lpsa} = \beta_0 + \beta_1 X + \epsilon \quad (1)$$

Ajuster le modèle donné en (1) et répondre aux questions suivantes :

3.3. Quelles sont les estimations des coefficients ? Interpréter l'estimation du coefficient $\hat{\beta}_1$.

3.4. Élaborer le test d'hypothèse de pente nulle pour le coefficient β_1 et conclure s'il y a

3.5. Une relation entre lpsa et X. β_1 est-il significativement non nul ?

3.6. Quelle est la valeur du coefficient de détermination R^2 ? Interprétez ce résultat.

Ce modèle est-il adapté pour prédire le taux d'antigène spécifique de la prostate ?