R & python: devoir maison

Contexte

[Connors *et al.*](The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators - PubMed) ont examiné l'association entre le recours au cathétérisme cardiaque droit (RHC pour *right heart catheterization*) pendant les 24 premières heures en soins intensifs (USI) et la mortalité à court terme (à 30 jours, résultat binaire), la durée des séjours en USI (jours, résultat discret) et l'utilisation des ressources (coût total du séjour, résultat continu) de 5 735 patients adultes sévèrement malades. Les données des patients recevant des soins intensifs ont été recueillies dans cinq hôpitaux universitaires des États-Unis entre 1989 et 1994. Les patients souffrant d'insuffisance respiratoire (IRA), de bronchopneumopathie chronique obstructive (BPCO), d'insuffisance cardiaque congestive (ICC), de cirrhose, de coma non traumatique, de cancer du côlon métastatique au foie, de cancer du poumon non à petites cellules (stade III ou IV) et de défaillance multiviscérales (MOSF) avec malignité ou septicémie étaient éligibles et ont été suivis pendant 6 mois.

Données

Une version prête à l'emploi des données peut être chargée en utilisant les commandes suivante dans Python ou dans R :

```
# dans Python
import pandas as pd

urlPath <- "https://hbiostat.org/data/repo/rhc.csv"
rawData <- pd.read_csv(urlPath)
```

```
# dans R
urlPath <- "https://hbiostat.org/data/repo/rhc.csv"
rawData <- read.csv(urlPath, header = TRUE)
```

la variable d'intérêt est la variable *Swang1* (qui vaut "RHC" ou "No RHC")

Objectif:

L'objectif de ce devoir est d'apprendre à utiliser et à interpréter le SMD (Standardized Mean Difference) comme une métrique des déséquilibres des covariables entre les

groupes dans les données observationnelles.

On dit qu'une co-variable est déséquilibrée si ses valeurs sont très différentes entre deux groupes

Pour rappel:

- pour une variable quantitative, le SMD se calcule selon (pour 2 groupes indépendants):

$$SMD = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{2}}}$$

où μ et σ sont les moyennes et les écart-types des groupes 1 et 2

- pour les variables qualitatives binaires, le SMD se caclule selon (pour 2 groupes indépendants):

$$SMD = \frac{100 \times (p_1 - p_2)}{\sqrt{\frac{(p_1(1-p_1)+p_2(1-p_2)}{2}}}$$

où p_1 et p_2 sont les proportions mesurées dans les groupes 1 et 2

- pour les variables qualitatives à plus de 2 classes, c'est plus compliqué :-)

Questions:

Q1. Identifier quinze variables quantitatives et cinq variables qualitatives binaires dans ce jeu de données (vous choisissez ces variables)

Q2. Pour toutes ces variables, construire un tableau présentant:

- pour les variables continues: les moyennes (m) et écart-types estimés (s) par groupe
- pour les variables binaires: les effectifs (n) et les proportions (%) par groupe
- le SMD correspondant

Ce tableau peut être du type:

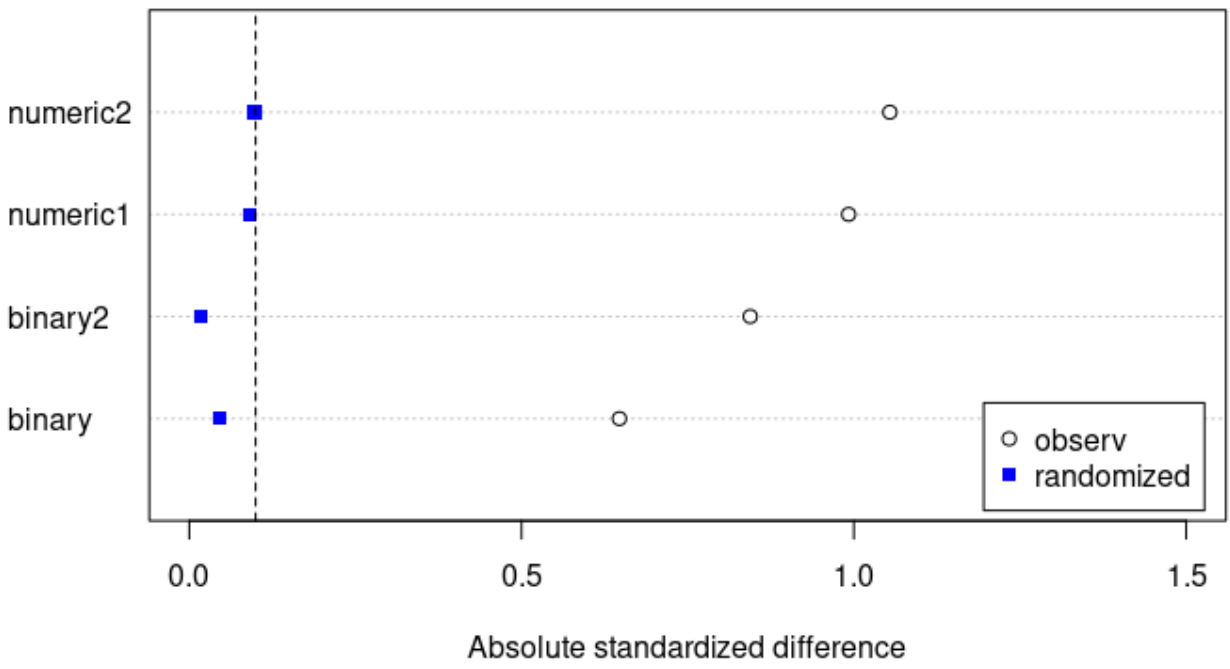
	RHC	No RHC	SMD
n	245	255	
numeric1 (mean (SD))	9.96 (10.17)	20.01 (10.09)	0.992
binary (n (%))	50 (20.4)	127 (49.8)	0.647
numeric2 (mean (SD))	10.16 (9.40)	20.66 (10.49)	1.054
binary2 (n (%))	55 (22.4)	155 (60.8)	0.844

Q3. Proposer une représentation graphique pour représenter les SMD des différentes variables, triés du plus grand au plus petit

Q4. Proposer un code permettant de randomiser le traitement RHC chez ces patients (vous pouvez utiliser la fonction `sample` dans R ou `resample` de Scikit learn dans python)

Q5. Recalculer les SMD pour les données après avoir randomisé le traitement (c'est à dire après avoir attribué au hasard pour chaque patient soit la valeur RHC, soit la valeur No RHC). Ajouter ces nouvelles valeurs sur le graphique de la question 3

Le graphique peut ressembler à ça:



Q6. Que peut-on conclure de l'utilisation des SMD pour vérifier l'équilibre des covariables (caractéristiques des patients) dans les études observationnelles?

Rendu et date limite

Un fichier pdf (au format **NomPrenom_DM.pdf**) avec vos réponses aux questions 1 à 6, et contenant une annexe présentant votre script R et/ou python vous ayant servi à répondre aux questions.

Il s'agit d'un travail **individuel** !

La deadline pour me rendre ce travail est le vendredi **06/12/2024 à 18h** - à déposer sur la page moodle du cours.