Chapitre 2 : Données manquantes.

Introduction

De ce qui précède, nous constatons qu'il est quasiment impossible dans une étude réelle d'avoir des données totalement complètes, d'où le besoin de traiter ces données de façon à ne pas biaiser les résultats.

Dans ce chapitre, nous présentons les caractéristiques des données manquantes, leurs mécanismes, la notion d'ignobilité et les différentes configurations des données manquantes. Puis nous décrivons les différentes méthodes de gestion des données manquantes.

Caractéristiques des données manquantes

Pour analyser les données manquantes, il faut commencer par connaitre les raisons de l'absence de certaines données.

Les données manquantes peuvent être classées en deux types :

Intentionnelles ; prévues par l'enquêteur, comme des réponses du type oui ou non ? si oui une autre sous question

Non intentionnelles ; hors de contrôle de l'enquêteur, un patient oublie ou refuse de répondre à une question.

Notations :

Nous utilisons les notations de Van Buuren [2012] :

$Y\ une\ matrice\ n.p\ contenant\ les\ valeurs\ de\ p\ variable\ pour\ tout\ les\ n\ individues\ de\ l'échantillon.$

$$R\ l'indicatrice\ des\ réponses, une\ patrice\ n \times p\ de\ 0\ et\ 1\ telle\ que$$

$r_{ij} = 1\ si\ y_{ij}\ est\ observée, et\ r_{ij} = 0\ sinon.$

$$Les\ données\ observées\ sont\ désignées\ par\ Y_{obs}\ et\ celles\ manquantes\ par\ Y_{mis}.$$

$Y = (Y_{obs}, Y_{mis})\ contient\ toutes\ les\ valeurs\ de\ données. Cependant\ les\ valeurs\ de\ Y_{mis}\ nous$

$sont\ inconues\ ellesont\ indiquées\ par\ R.$

$$Si\ Y = Y_{obs}\ c'est\text{à}\ dire\ que\ l'échantillonest\ complètement\ observé$$

$Si\ aucune\ données\ n'a\ été\ obtenue\ pour\ l'individu\ i,\ \ la\ i^{ème}\ ligne\ de\ Y\ contiendra\ uniquement\ l'identifiant\ de\ i\ et, éventuellement, des\ données\ administratives\ (cas\ de\ non-éponse\ totale).$

Mécanismes des données manquantes (MCAR, MAR, MNAR) :

Le processus qui régit les probabilités d'absence des données est appelé mécanisme des données manquantes ou mécanisme de non-réponse.

Les emplacements des données manquantes de Y sont stockés dans la matrice R. La distribution de R peut dépondre de $Y = (Y_{obs}, Y_{mis})$. Cette relation est décrite par le modèle de données manquantes dont les paramètres sont contenus dans ψ .

L'expression générale du modèle des données manquantes est : $P(R/Y_{obs}, Y_{mis}, \psi)$.

On dit que les données sont on MCAR (Missing Completly At Random) si : $P(R/Y_{obs}, Y_{mis}, \psi) = P(R = 0/\psi)$

Les causes des données manquantes ne sont pas reliées au données, la probabilité d'avoir une observation manquante y_{ij} dépend uniquement de certain paramètres ψ, et pas des données Y. Par exemple, des données perdues à cause d'une panne de matériel. MCAR est souvent irréaliste en pratique.

On dit que les donnes sont MAR (Missing At Random) si : $P(R/Y_{obs}, Y_{mis}, \psi) = P(R = 0/Y_{obs}, \psi)$

La probabilité des données manquantes peut dépendre des données observées, y compris les paramètres de conception. Par exemple si à partir d'un certain temps on a des données manquantes aléatoirement. MAR est plus général et plus réaliste que MCAR. Les méthodes modernes commencent souvent par une hypothèse MAR.

On dit que les données sont MNAR (Missing Not At Random) si $P(R/Y_{obs}, Y_{mis}, \psi)$ ne se simplifie pas.

La probabilité d'avoir des données manquantes dépend aussi des informations non observées Y_{mis}, est varient pour des raisons inconnues. Par exemple si les données manquantes dépendent de l'enquêteur qui les récolte.

Ignorables et non ignorables :

Soient θ les paramètres dont dépend la fonction de densité conjointe des données réellement observées ; $Y_{obs}\ et\ R$ pour les données complètes de Y, et les paramètres ψ pour l'indicatrice de réponses R.

La densité conjointe $f(Y_{obs}, R/\theta, \psi)$ est proportionelle à la vraisemblance de $\theta\ et\ \psi$.

$$l(\theta, \psi/Y_{obs}, R) \propto P(Y_{obs}, R/\theta, \psi)$$

Il est intéressant de pouvoir déterminer θ sans conaitre ψ, en ignorant le mécanisme de données manquantes.

On dit que le mécanisme des données manquantes est ignorable pour l'inférence probabiliste si :

Les données manquantes sont MAR.

L'espace des paramètres joints (θ, ψ) est le produit de l'espace des paramètre θ, et de l'espace des paramètre ψ. $\theta\ et\ \psi$ sont distincts.

Configuration des données manquantes :

Les données manquantes peuvent être classées en trois catégories :

Univariée

Dans ce cas une seule variable possède des données manquantes. Ce qui est rare en pratique.

Figure 6: Données manquantes univariée

Monotone

Dans ce cas on peut réarranger les colonnes de la matrice Y de taille $n \times p$ tel que si la variable Y_j est manquante pour un individue $i, i \in \{1, \dots, n\}$, alors toutes les variables suivantes $Y_k, k > j$ sont absentes pour ce même individue.

Figure 7: Données manquantes monotones

Arbitraire

Dans ce cas les données manquantes sont réparties uniformément dans le jeu de données, elles n'ont pas de structures.

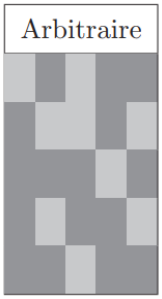

Figure 8: Données manquantes arbitraires.

Flux entrant et flux sortant

Le flux entrant et le flux sortant sont deux mesures qui permettent de savoir comment chaque variable se connecte aux autres.

Le coefficient de flux entrant I_j est défini par :

$$I_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - r_{ij}) r_{ik}}{\sum_k^p \sum_i^n r_{ik}}$$

I_j est le nombre de variables paires (Y_j, Y_k) avec Y_j manquant et Y_k observé, devisé par le nombre totale de cellules observées. Une variable complétement observée a un flux entrant égal à 0, tandis qu'une variable complétement manquante en a un égal à 1. Pour deux variables ayant la même proportion des valeurs manquantes, celle avec le flux entrant le plus élevé est la mieux connectée avec les variables observées, et est donc la plus facile à imputer.

Le coefficient de flux sortant O_j est défini par :

$$O_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - r_{ik}) r_{ij}}{\sum_k^p \sum_i^n (1 - r_{ij})}$$

O_j est le nombre de variables paires (Y_j, Y_k) avec Y_j observé et Y_k manquante, devisé par le nombre totale de cellules incomplètes. Une variable complétement observée a un flux sortant égal à 1, tandis qu'une variable complétement manquante en a un égal à 0. Pour deux variables ayant la même proportion des valeurs manquantes, celle avec le flux sortant le plus élevé est la mieux connectée avec les variables observées, et ainsi plus utile pour imputer les autres variables.

Méthodes de gestion de données manquantes :

Analyse des cas complets (Complete case analysis) :

La méthode par défaut de gestion des données manquantes était l'analyse des cas complets (CCA), elle consiste à éliminer les lignes (individus) ayant des données manquantes dans les variables à analyser.

A moins que les données soient MCAR, une situation peu réaliste en pratique, CCA peut fournir des estimations de la moyenne, des coefficients de régression et de corrélation biaisés.

Cependant cette méthode n'est pas toujours mauvaise, les conséquences des données manquantes peuvent varier selon les situations et l'endroit où elles se produisent ainsi que le taux de données manquantes.

Etude des cas disponibles (Pairwise deletion) :

Pour remédier à la perte d'information due à CCA, l'étude des cas disponibles consiste à ne considérer que les cas où les variables d'intérêt sont complètes. La moyenne d'une variable X_i est calculée sur tous les cas ayant des données observées sur X_i, et la corrélation et covariance de deux variables X_i et X_j sont calculer sur tous les cas où les deux variables sont observées à la fois.

Pour les données MCAR cette méthode utilise toutes les informations disponibles et ainsi fournit des estimations consistantes de la moyenne des corrélations et des covariances. Néanmoins, ces estimations ont d'importante lacunes et peuvent être biaisées si les données ne sont pas MCAR. La matrice de corrélation peut ne pas être définie positive ce qui complique les calcules pour la plupart des procédures multivariées

Procédure de modélisation de la distribution des données manquantes :

Pour cette approche il n'y a besoin d'imputer les données incomplètes ou les éliminer de l'analyse, en revanche les paramètres de la distribution du modèle sont estimés par maximum de vraisemblance (maximal likelyhood) par l'algorithme Expectation-Maximisation (EM) ou autres, pour ensuite faire des inférences et estimer les valeurs des données manquantes. Dans la plupart des cas ces inférences peuvent être évaluées en estimant leur erreur standard.

Cette méthode est à la base d'une autre plus intéressante ; l'Imputation Multiple, qui offre la possibilité d'évaluer l'effet des hypothèses de modélisation sur les inférences.

Imputation par la moyenne :

Cette méthode remplace les valeurs manquantes de chaque variable par sa moyenne, ou bien le mode pour les données catégoriques. Bien que ce soit une méthode simple et rapide, elle donne une variance sous-estimée et dérange les relations entre les variables et donne des estimations biaisées même pour la moyenne quand les données ne sont as MCAR. Ce qui restreint son utilisation uniquement pour les cas où seules quelques valeurs sont manquantes.

Imputation par régression :

Cette méthode repose sur la production d'un modèle à partir des données observées en incorporant les autres variables, puis remplacer les données manquantes par les prédictions du modèle. Avec la condition que les variables soient corrélées linéairement. Cette condition n'est pas toujours réaliste.

Sous la condition MCAR, cette méthode fournit des estimations non biaisées de la moyenne. Si les paramètres influençant les données manquantes font partie du modèle, les poids de la régression ne sont pas biaisés sous la condition MAR.

LOCF et BOCF :

La méthode de Dernière Observation Reportée (Last Observation Carried Forward), remplace les valeurs manquantes par la dernière valeur observée. Cette méthode est utilisée dans les essais cliniques ce qui la rend la méthode préférée de l'agence américaine FDA, mais elle peut donner des estimations biaisées même sous MCAR. Ce qui rend LOCF et BOCF moins recommandées si les hypothèses scientifiques ne le justifient pas.

Méthode de l'indicatrice :

Quand on veut effectuer une régression et que quelques valeurs des variables explicatives sont manquantes, la méthode de l'indicatrice remplace chaque valeur manquante par 0 et étend le modèle de régression par l'indicatrice de réponse et ainsi conserve le jeu de données complet, et tiens compte de la différence entre les données observées et celles estimées grâce à l'indicatrice. Or elle peut produire des résultats biaisés même sous la condition MCAR, surtout que les conditions sous lesquelles cette méthode est valide sont difficiles à obtenir en pratique, et qu'elle ne permet pas d'avoir des données manquantes dans les événements d'intérêt.

L'imputation multiple :

Généralement, la méthode de l'imputation multiple crée un nombre m de jeu de données ($m > 1$) , puis réalise l'étude statistique souhaitée sur chacun des jeux de données, ensuite les estimations de toutes ces études sont combinées en une seule avec une erreur standard calculée. Ceci peut se résumer en trois grandes étapes.

Etape 1 : l'imputation

La méthode crée m jeu de données complets en remplaçant chaque valeur manquante par une valeur plausible tirée spécifiquement d'une distribution modélisée à partir des données observées. Les m jeux de données sont identique pour les valeurs observées, mais différentes pour celle à imputer.

Etape 2 : l'analyse

Ensuite une analyse statistique est utilisée sur chaque jeu de données maintenant tous imputés, pour appliquer les méthodes voulues comme si les données étaient complètes et en estimer les paramètres d'intérêt. Ce qui donne m résultats différents. Ces différences coulent de l'incertitude sur les valeurs imputées.

Etape 3 : combinaison des estimations

Les m parametres estimés sont ensuite combinés en une seule dont on estimera la variance. Cette variance combine la variance d'échantillonnage classique et la variance supplémentaire causée par les données imputées. Sous les bonnes conditions, les valeurs estimées ne sont pas biaisées et ont des propriétés statistiques correctes.