Données manquantes

STT-7330 Analyse de données

Véronique Tremblay

Département de mathématiques et de statistique

Hiver 2019

Valeurs manquantes

Une valeur manquante implique qu'il y a une valeur pour cette observation, mais qu'elle n'a pas été capturée.

À ne pas confondre avec l'absence de valeur.

Mécanisme de non-réponse

- Données manquantes complètement au hasard (MCAR)
- Données manquantes au hasard (MAR)
- Données manquantes pas au hasard (NMAR)

Notation

Soit Y une matrice de données de dimension n par p.

L'entrée Y_{ij} donne la valeur de la variable j pour l'observation i.

On divise cette matrice en deux parties

$$Y = (Y_{obs}, Y_{mis}).$$

Y_{obs} contient les données observées,

Y_{mis} contient les données manquantes.

On définit également une matrice indicatrice R de taille n par p nommée **matrice de réponse** où

$$R_{ij} = \begin{cases} 1 \text{ si } Y_{ij} \text{ est observé} \\ 0 \text{ si } Y_{ij} \text{ est manquant} \end{cases}$$

On s'intéresse à la distribution de R, soit le mécanisme de réponse. On peut écrire

$$f(R|Y, \theta)$$

où θ est un vecteur de paramètres.

MCAR (Missing completely at random)

Lorsque la probabilité de réponse ne dépend pas de Y, i.e.

$$f(R|Y,\theta) = f(R|\theta).$$

Exemple: On perd au hasard 20% des valeurs mesurées. Alors

$$f(R_{ij}|\theta) \sim \text{Bernoulli}(0,8)$$

pour tout $i = 1,\ldots,n$, $j = 1,\ldots,p$.

MAR (Missing At Random)

Lorsque la probabilité de réponse dépend uniquement de la valeur des variables qui ont été observées, i.e.

$$f(R|Y,\theta) = f(R|Y_{obs},\theta)$$

Exemple: On fait remplir un questionnaire à plusieurs individus, en leur demandant leur sexe et leur salaire. Si les femmes ont une plus grande probabilité que les hommes d'accepter de fournir leur salaire, alors on aurait un mécanisme MAR.

NMAR (Not Missing At Random)

Lorsque la probabilité de réponse dépend de Y_{mis} également. Dans ce cas, on ne peut pas simplifier

$$f(R|Y, \theta)$$

NMAR (Not Missing At Random)

- La probabilité de réponse dépend de la variable elle-même.
  - Les individus avec des salaires plus élevés ont une plus grande probabilité de refuser de déclarer leur salaire.
- La probabilité de réponse dépend d'une variable non-observée.
  - La probabilité de déclarer son salaire dépend de l'âge de l'individu, mais on n'a pas mesuré cette variable.

Quoi faire?

- Documenter la non-réponse.
  - Faire des statistique descriptives
- Identifier quel mécanisme de non-réponse serait défendable.
  - Tests pour MCAR
  - Connaissance du domaine
- Choisir une méthode de traitement de la non-réponse.

Identifier le mécanisme de non-réponse

Test de Welch (test de t)

Adaptation du test de t de Student dans le cas où les deux populations ont des variances inégales.

Technique : Séparer les cas observés et manquants pour une variable et tester les différences de moyennes pour les autres variables.

Identifier le mécanisme de non-réponse

Test de Welch (test de t)

- Implique beaucoup de tests; problème de comparaisons multiples.
- Ne tient pas compte des corrélations entre les variables.
- Ne garantie pas l'hypothèse MCAR
- Peut nous aider à identifier les variables pour lesquelles ajuster dans les procédures de traitement de données manquantes.

Identifier le mécanisme de non-réponse

Test de Little

Compare en un seul test les moyennes et les variances à l'intérieur de chaque patron de non-réponse.

Statistique de test :

$$d^2 = \sum_{j=1}^{J} n_j (\hat{\mu}_j - \hat{\mu}_j^{ML}) \hat{\Sigma}_j^{-1} (\hat{\mu}_j - \hat{\mu}_j^{ML})$$

où $j = 1, \ldots, J$ dénote le patron de non-réponse, μ_j est la moyenne pour les cas du patron j, $\hat{\mu}_j^{ML}$ est l'estimateur du maximum de vraisemblance global de la moyenne et Σ_j l'estimateur du maximum de vraisemblance global de la matrice de covariance. Le nombre de variables change d'un patron à l'autre, d'où l'index j.

Identifier le mécanisme de non-réponse

Test de Little (suite)

Sous l'hypothèse nulle que les données sont MCAR, on a

$$d^2 \overset{approx}{\sim} \chi^2 \text{ avec } \sum_{j=1}^{J} k_j - k \text{ degrés de liberté}$$

où k_j est le nombre de variables complètement observées dans le patron j et k est le nombre de variables du jeu de données.

Identifier le mécanisme de non-réponse

Test de Little

- Le test ne nous indique pas quelles variables sont problématiques.
- Le test suppose la même matrice de covariance pour tous les patrons de non-réponse.
- Des études de simulations ont mis en évidence un manque de puissance de ce test.
- Ne permet pas de garantir l'hypothèse MCAR.

Identifier le mécanisme de non-réponse

Traiter la non-réponse

- Retirer les variables qui contiennent trop de données manquantes
  - Perte d'information
- Retirer les observations avec des données manquantes
  - Risque de biais si MAR ou NMAR

Traiter la non-réponse

- Si la variable est catégorielle : créer une catégorie «inconnu»
- Imputer

Traiter la non-réponse - Imputer

- Par la moyenne, la médiane ou le mode
  - Modifie la distribution
  - Altère la structure de corrélation

Traiter la non-réponse - Imputer

- Par la distribution
  - On remplace par une autre valeur pigée au hasard dans les données
  - Préserve la distribution, mais on perd la structure de corrélation

Traiter la non-réponse - Imputer

- Avec un modèle prédictif
  - Régression, arbre de décision
  - Préserve la distribution et la structure de dépendance
  - Réduit la variabilité (on surestimera la précicion)

Quelques solutions

- Ajouter une erreur au modèle de régression
- Faire de l'imputation multiple