Que faire des valeurs manquantes?

Quelques options possibles (Raghunathan-2004)





Balcone Thomas
DMS

Insee
Mesurer pour comprendre

1. L'approche standard

- Exclusion des valeurs manquantes

  → Problème les estimateurs sont souvent biaisés

  – Exemple: estimateurs des coefficients d'une régression logistique

$$\log it\,\Pr(D=1|\,x,E)=\text{int ercept}+\alpha\times E+\beta\times x \quad (1)$$

Avec :

$$x \sim N(0,1)$$

$$\text{int ercept}=-0,5$$

$$\alpha=0,5$$

$$\beta=0,5$$

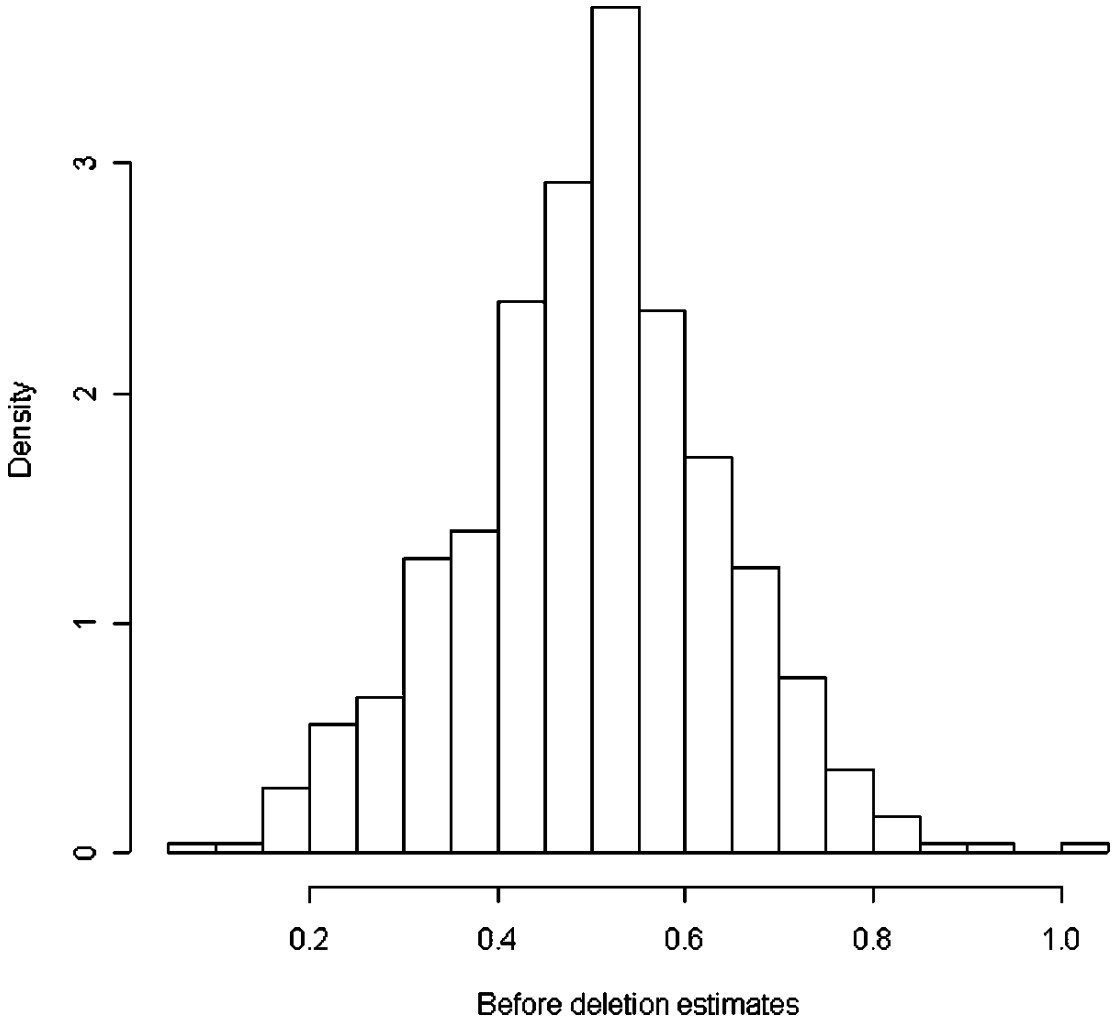
1. L’approche standard

- On génère 2500 échantillons de taille n=1000
- On crée 2500 échantillons correspondant en affectant une valeur manquante à x pour environ 15% des observations

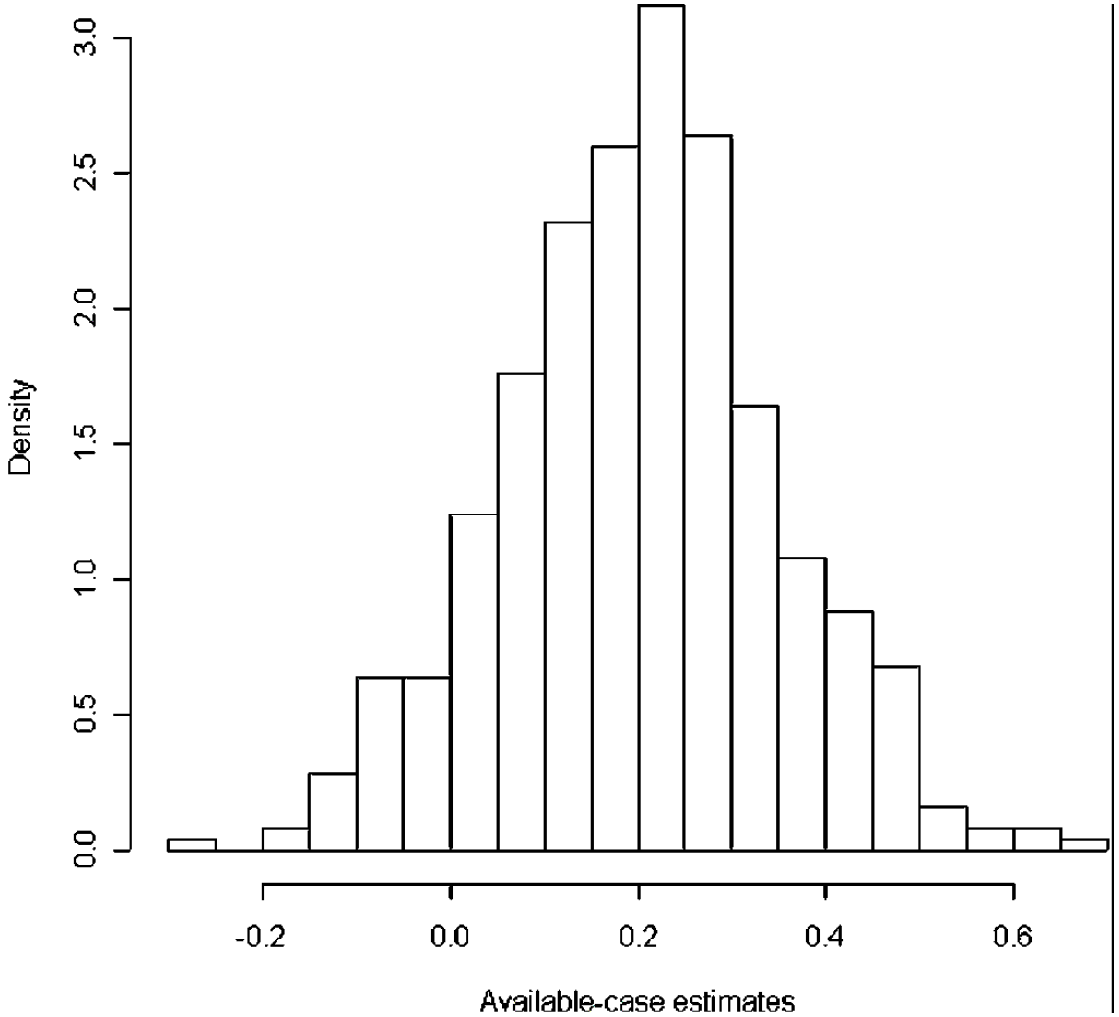
- Puis, on réestime le modèle (1) pour chacun de ces ensembles de données (données non manquantes vs données manquantes)

- Paramètre d’intérêt β

1. L’approche standard



1. L’approche standard



2. Les mécanismes de non-réponse

- 2 variables:
  - U qui peut être à valeurs manquantes,
  - V qui est toujours observée

- On définit une variable dichotomique R:

$$R = \begin{cases} 1 \text{ si } U \text{ est observée} \\ 0 \text{ sinon} \end{cases}$$

2. Les mécanismes de non-réponse

- Le MCAR (missing completely at random):

$$\Pr(R = 1 \mid U, V) = c$$

Hypothèse très forte et rarement vérifiée en pratique

- Le MAR (missing at random):

$$\Pr(R = 1 \mid U, V) = f(V)$$

Les valeurs manquantes de U peuvent être estimées à partir de la distribution de U|V

- Le NMAR (not-missing at random)

$$\Pr(R = 1 \mid U, V) = f(U, V)$$

3. Pondérer

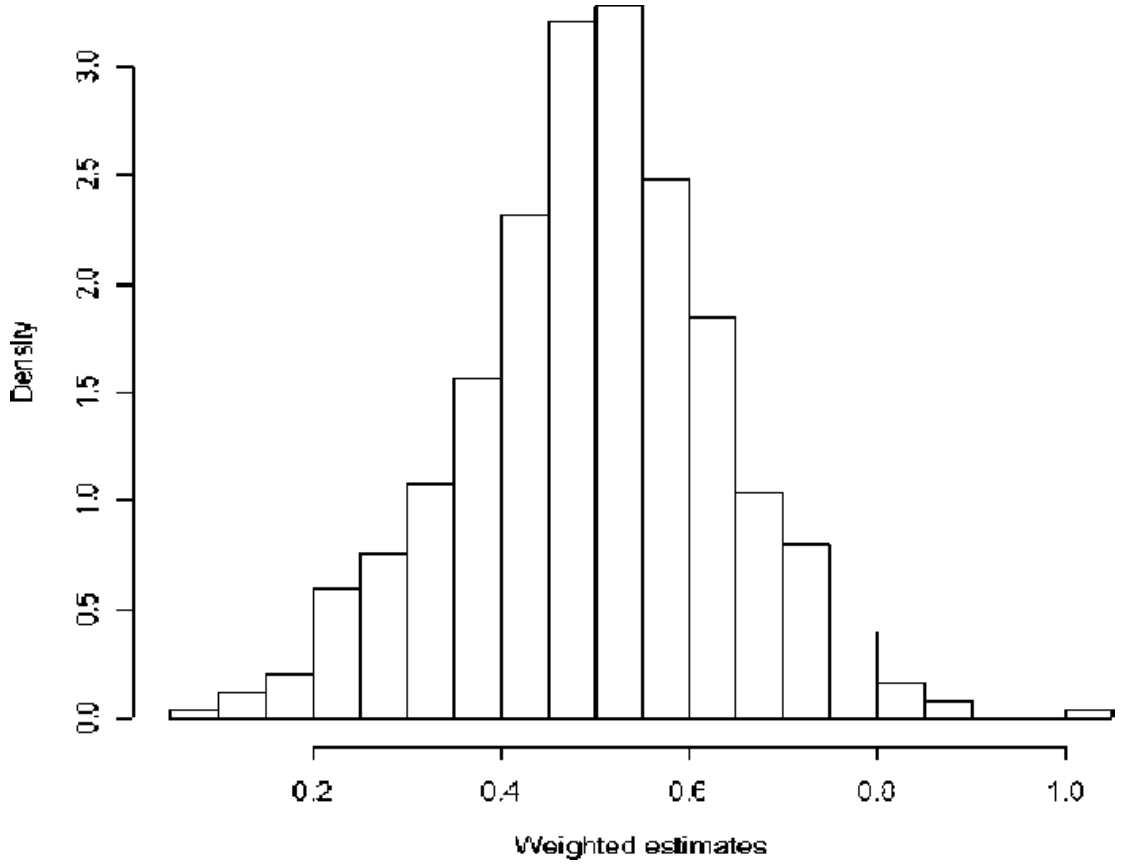
- On pondère les répondants pour prendre en compte la non-réponse

- Retour à l'exemple précédent:
  - Pour chaque croisement DxE, le poids est l'inverse du taux de réponse correspondant (the « adjustment-cell method »):

$$w_{de} = \frac{n_{de}}{r_{de}}$$

3. Pondérer

– On obtient un nouvel estimateur de β avec une régression logistique pondérée:



3. Pondérer

- Autre manière possible de définir les poids (the « response propensity method »):
  - On estime le modèle suivant:

$$\Pr(R=1|V)=\left[1+\exp\left(-\beta_0-V^t\beta_1\right)\right]^{-1}$$

  - On en déduit alors les poids des répondants:

$$w_j=1+\exp\left(-\hat{\beta}_0-V_j^t\hat{\beta}_1\right)$$

4. Imputation multiple

- Imputation simple:
  - Pas assez d'«incertitude »
    - Écarts-types trop petits
    - Intervalles de confiance trop « étroits »

- Imputation multiple:
  - Idée: on s'autorise plusieurs (M) valeurs possibles pour imputer les valeurs manquantes
    - On obtient alors M bases de données complètes (i.e. sans valeurs manquantes)
  - En pratique: 2<=M<=5

4. Imputation multiple

- On définit l'estimateur « multiple imputation »:

$$\bar{e}_{MI} = \frac{1}{M}\sum_{l=1}^{M} e_l$$

- Et l'erreur-type associée:

$$s_{MI} = \sqrt{variance\_intra + variance\_inter}$$

où

$$variance\_intra = \bar{u}_M = \frac{1}{M}\sum_{l=1}^{M} s_l^2$$

et

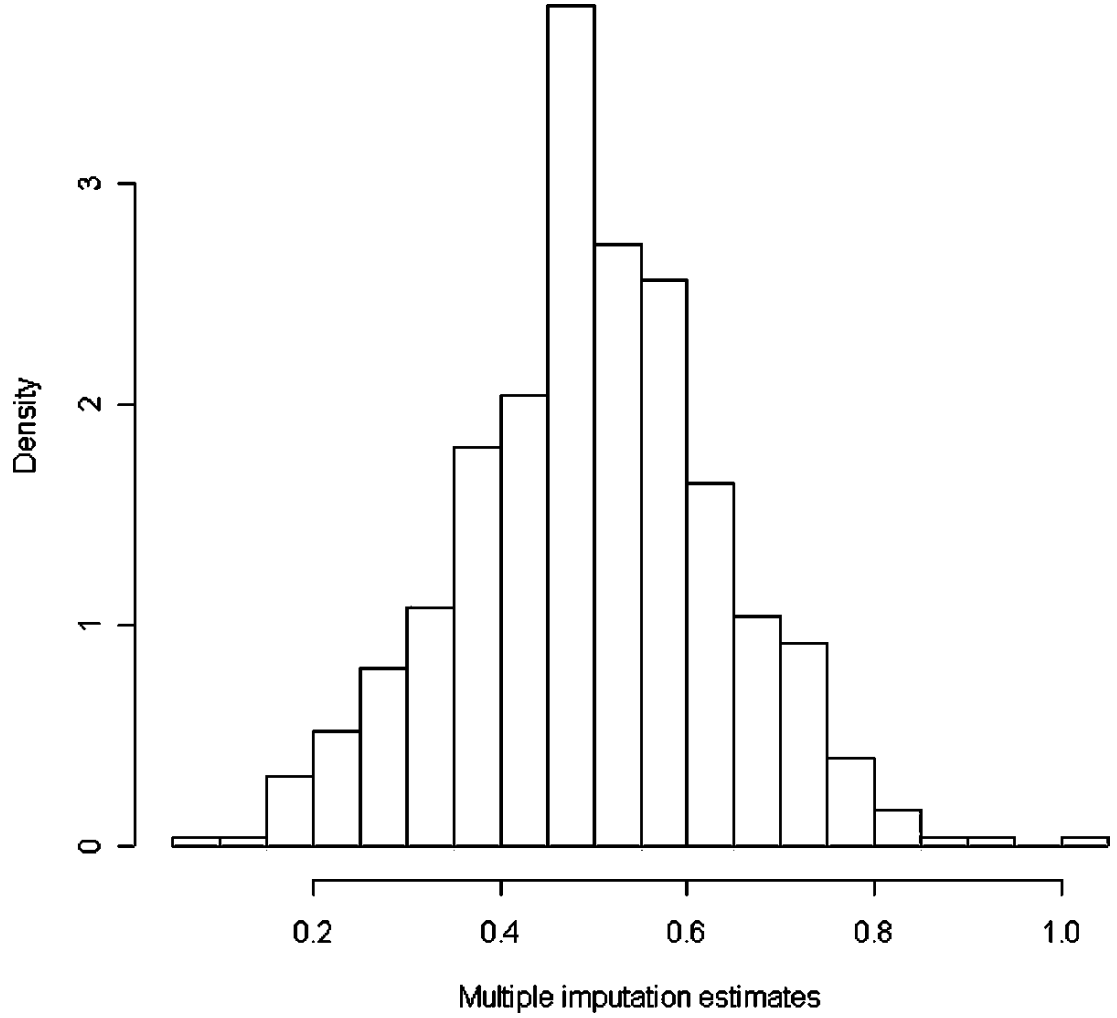
$$variance\_inter = \frac{M+1}{M} b_M = \frac{M+1}{M}\frac{1}{M-1}\sum_{l=1}^{M}(e_l - \bar{e}_{MI})^2$$

4. Imputation multiple

- Comment générer les différentes valeurs possibles pour les valeurs manquantes?
  - En utilisant l'approche Bayésienne:
    - Les valeurs sont tirées dans la distribution a posteriori des valeurs manquantes conditionnellement aux valeurs observées

- Exemple:
  - 2500 échantillons avec des valeurs manquantes pour x
  - 5 imputations différentes sont réalisées pour chacun de ces 2500 échantillons en utilisant l'approche Bayésienne
    - On en déduit alors 2500 estimateurs « imputation multiple » pour β

4. Imputation multiple



4. Imputation multiple

- Une alternative à l'approche Bayésienne: le SRMI (Sequential Regression Multiple Imputation):
  - X: variables sans valeurs manquantes
  - Y_1, Y_2,…, Y_k: variables à valeurs manquantes
  - C étapes:
    - 1ère étape:
      - On régresse Y_1 sur X et les valeurs manquantes de Y_1 sont imputées
      - On régresse Y_2 sur X et Y_1 et les valeurs manquantes de Y_2 sont imputées
      - …
      - On régresse Y_k sur X, Y_1 , …, Y_{k-1} et les valeurs manquantes de Y_k sont imputées
    - De la 2ème à la c-ième étape:
      - On procède de la même manière qu'à l'étape 1 sauf que toutes les variables excepté la variable qu'on impute sont maintenant incluses dans la régression
    - Après la c-ième étape:
      - On obtient les valeurs imputées « finales » pour les variables Y_1 , Y_2 ,…, $Y_{k....}$

5. Maximum de vraisemblance

- On estime les paramètres d'intérêt en maximisant la vraisemblance calculée sur les données observées

  – Exemple:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}\right]$$

  - Hypothèses:
    – p individus pour lesquels X et Y sont observées
    – q individus pour lesquels seule X est observée
    – r individus pour lesquels seule Y est observée

  - La vraisemblance « observée » L_{obs}:

$$L_{obs} = L_p \times L_q \times L_r$$

  - Maximisation en utilisant Newton-Raphson ou l'algorithme E-M

5. Maximum de vraisemblance

– Cas plus général:
  - i=1,2,…,n

$$Y_i = \left(Y_{i,obs}, Y_{i,miss}\right)$$

$$f\left(Y_i \mid \theta\right) = f\left(Y_{i,obs}, Y_{i,miss} \mid \theta\right)$$

  - La vraisemblance « observée »:

$$L_{obs}(\theta) = \prod_{i=1}^{n} L\left(\theta \mid Y_{i,obs}\right) \propto \prod_{i=1}^{n} \int f\left(Y_{i,obs}, Y_{i,miss} \mid \theta\right) dY_{i,miss}$$

- Mais la maximisation peut s'avérer complexe…

6. Discussion et limites

- Sous l'hypothèse d'un mécanisme de non-réponse MAR, 3 approches valides pour traiter la non-réponse:
  - Pondérer:
    - Estimateur non biaisé
    - Ne prend pas en compte la non-réponse partielle
  - L'imputation multiple:
    - La plus pratique
    - Relativement simple à implémenter dans la plupart des logiciels
      - STATA, IVEWARE, SAS, SOLAS
    - Gros travail en amont pour créer les différentes bases de données avec les valeurs imputées
  - Le maximum de vraisemblance:
    - L'approche la plus « sophistiquée »
    - Problème de la disponibilité des logiciels

6. Discussion et limites

- Le problème des données manquantes est inévitable…
  - Essayer d'avoir le maximum de variables auxiliaires qui pourront être utilisées dans l'imputation multiple
    - Données administratives
    - Variables à ajouter dans l'enquête
    - …

Que faire des valeurs manquantes?

Merci de votre attention !

Contact
M. Thomas Balcone
Tél. : 01 41 17 64 54
Courriel : thomas.balcone@insee.fr

Insee
18 bd Adolphe-Pinard
75675 Paris Cedex 14
www.insee.fr
Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00

Insee