Regression With Missing X's : A Review

Roderick J. A. Little

Journal of the American Statistical Association

Dec. 1992

Plan

- Énoncé du problème
- Schémas et Mécanisme des valeurs manquantes
- Les méthodes d'estimation tenant compte des données manquantes
- Autres pistes
- Conclusions

Énoncé du problème

- Problème de régression linéaire

$$E(Y \mid X_1, \ldots, X_p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j; \qquad \beta = (\beta_1, \ldots, \beta_p)$$

$$\beta = \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1}; \qquad \beta_0 = \mu_y - \sum_{j=1}^{p} \beta_j \mu_j;$$

$$\mu = (\mu_1, \ldots, \mu_p, \mu_y) \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \sigma_{yy} \end{pmatrix}$$

- Avec valeurs manquantes (VM) sur les X's
- Les observations avec VM sur Y n'apportent pas ou peu d'information à la régression – qd elles sont aléatoires

Schémas des VM

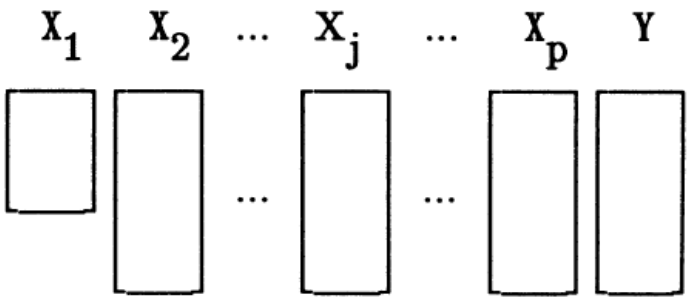

Figure 1. Pattern of Univariate Missing Data.

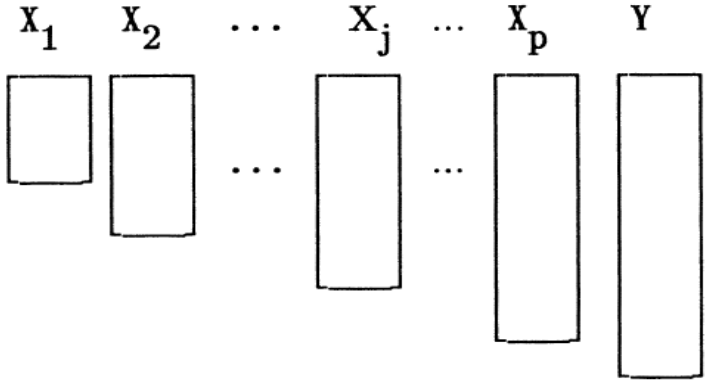

Figure 2. Pattern of Monotone Missing Data.

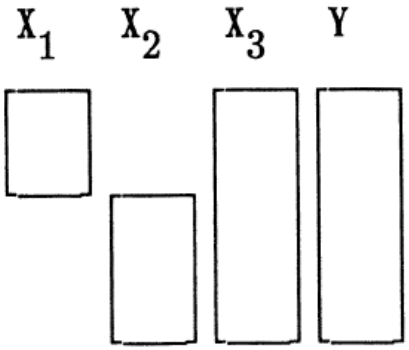

Figure 3. Special Pattern of Missing Data with Unidentified Parameters.

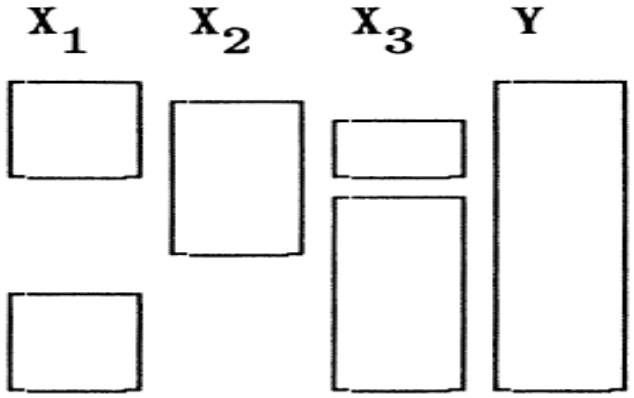

Figure 4. General Pattern of Missing Data.

Mécanisme des VM

- La probabilité qu'une donnée soit manquante est :
- (a) : indépendante de sa valeur et des valeurs des autres variables
- (b) : dépendante de sa valeur (non observée)
- (c) : dépendante des valeurs des autres X
- (d) : dépendante des valeurs des autres X et Y
- Mécanisme MCAR : proba d'apparition des VM est indépendante des variables étudiées, (a) par exemple
- Mécanisme MAR : proba d'apparition des VM est dépendante des valeurs observées, (c) et (d) par exemple lorsqu'on est dans le schéma 1

Application aux données (MHS)

- Régression de Y = la dépression, par 4 régresseurs : revenu (-, 1), âge (-, 4), santé (+, 3), bed-days (+, 2)
- Jeu complet de données de 300 observations
- Simuler 3 jeux de données avec 50% de VM sur la variable revenu (S1) :

  - VM générées de manière aléatoire, MCAR

  - VM corrélées aux revenus X1, sélection à partir de X

  - VM corrélées aux valeurs de Y, sélection à partir de Y

- Exemple pour illustrer les méthodes, mais non pour tirer des conclusions générales

Les méthodes d'estimation

- 6 classes :

  1 - Observations complètes (CC)

  2 - Données disponibles (AC)

  3 - Moindres carrés sur données imputées (LS)

  4 - Maximum de vraisemblance (ML)

  5 - Méthodes Bayésiennes

  6 - Imputations multiples (MI)

1 - Observations complètes (CC)

- Régression classique (par MC), restreinte au jeu d'observations pour lequel toutes les variables sont renseignées : méthode standard (base de comparaison)
- Avantages :

  - simple à mettre en œuvre

  - inférence valide quand VM dépend de sa valeur (b)
- Inconvénients :

  - perte d'information, d'autant plus que le nbre de X est important

  - sinon supprime les X qui comportent un gd nbr de VM

1 – CC bis

- A noter :

  Supprimer les observations incomplètes revient à choisir judicieusement les valeurs imputées de telle sorte que le résidu soit nul

- Application MHS :

  - Inférence reste valide dans le cas MCAR et sélection à partir de X, mais la précision des estimateurs décroit de 30 % à 50 % qui reflète la perte en observations de 50 %

  - Le coeff relatif à X1 est biaisé quand la sélection est faite à partir de Y

2 – Données disponibles (AC)

- Estimer Espérance et Variance avec valeurs disponibles de Xj
- Covariance (Xj, Xk) avec les observations dont on dispose de valeurs pour les 2 variables
- Meilleur que CC, sauf quand les X sont très corrélés
- Matrice var-covar pas tjr définie - positive (convergence estimateurs)
- MHS : ici résultats meilleurs que CC, car corrélation faible des X

3 - MC sur données imputées (LS)

- Imputer les VM + faire une régression MCO ou MC Pondérées sur les données

- **Imputation par une moyenne** : matrice var-covar biaisée (dénominateur), méthode non recommandée
- **Imputation par moyenne conditionnée sur les X** : régression de Xj (VM) par les autres X sur les CC (MAR).

3 – LS bis

- 2 conséquences :

  - Augmente variance résiduelle → pour diminuer l'influence des VM, régression MC Pondérées $w^* = \sigma_{yy \cdot 1s} / \sigma_{yy \cdot s} = 1 - \rho^2_{1y \cdot s}$

  - Augmente corrélation entre X sur les observations incomplètes, affecte estimation de variance des estimateurs, d'autant plus quand VM importantes → choix de poids qui tient compte de la part des VM $w = \frac{(1 - \rho^2_{1y \cdot s}) m/n}{\rho^2_{1y \cdot s} + (1 - \rho^2_{1y \cdot s}) m/n}$

- MHS : OLS et WLS similaires car X1 peu corrélé avec autres X, Pour X1 même coeff que CC, autres coeffs proches AC, avec précision meilleure, mais sous-estimée car non prise en compte de la variance de l'imputation

3 – LS ter

- **Imputation par moyenne conditionnée sur Y et les X** (méthode Buck) :

  - Pour faire des imputations, régression CC

  - On régresse par Y pour ensuite faire une régression sur Y

  - Estimation biaisée, mais il existe des méthodes de correction des biais dans la matrice var-covar

  - Méthode valable pour le schéma général des VM

  - Précision est sous-estimée, erreurs d'imputation non prises en compte (voir MI)

4 - Maximum de vraisemblance (ML)

- Loi de distribution jointe des X et Y (normale multivariée) déterminer des paramètres de la loi par MV (espérance et matrice var-covar)
- **Factorisation de la vraisemblance** en produit de deux vraisemblances. une avec *m* données l'autre avec *n* $L(\varphi_1, \varphi_2) = L_1(\varphi_1)L_2(\varphi_2)$
- Dans MCAR, par rapport au CC, réduction variance liée à la prise en compte des observations incomplètes $\text{var}(\hat{\theta}) \simeq \text{var}(\tilde{\theta}) - \hat{\theta}_2^T\{(\mathbf{I}_m^{-1}(\hat{\varphi}_1) - \mathbf{I}_n^{-1}(\hat{\varphi}_2)\}\hat{\theta}_2$
- Méthode de factorisation applicable pour le schéma général

4 – ML bis

Table 3. Regression of Y on X_1 and X_2, X_1 Observed for m Cases and Missing for n − m Cases: Proportional Decrease in Variance of Estimators from OLS, WLS, and ML Relative to CC Estimate

	Parameter	
Method	$\beta_{y1 \cdot 12}$	$\beta_{y2 \cdot 12}$
OLS	0	$\left(1 - \frac{m}{n}\right) \frac{(1 - \rho_{12}^2)(1 - 2\rho_{1y \cdot 2}^2)}{1 - \rho_{1y \cdot 2}^2}$
WLS	0	$\left(1 - \frac{m}{n}\right)(1 - \rho_{1y \cdot 2}^2)(1 - \rho_{12}^2)$
ML	$\left(1 - \frac{m}{n}\right)2\rho_{1y \cdot 2}^2(1 - \rho_{1y \cdot 2}^2)$	$\left(1 - \frac{m}{n}\right)(1 - \rho_{1y \cdot 2}^2)[1 - \rho_{12}^2(1 - 2\rho_{1y \cdot 2}^2)]$

Proportionnelle à la part des VM et corrélations partielles,

Si X1 est peu corrélé avec X et Y, réduc var maximale, X1 influe peu sur régression de Y sur X2

Qd corré partielle X1,Y sup à 1/2, OLS moins bon que CC, WLS meilleur, et ML encore meilleur

4 – ML ter

- Pour le schéma général, la méthode ML utilise une démarche itérative EM, Estimation – Maximisation
- Point de départ, les coeffs issus de CC
- E correspond à la phase d'imputation par la méthode de Buck avec correction de biais
- M est la phase de MV pour déterminer les nouveaux coeffs
- L'algorithme repart avec les nouveaux coeffs, jusqu'à convergence

5 - Méthodes Bayésiennes

- Partie succincte (1 colonne contre 2 pages pour ML)
- 2 messages :

  - Plus efficace que ML quand petit échantillon

  - Méthode assez peu développée pour traiter le prob de VM sur les X

6 - Imputations multiples (MI)

- MI solution au problème de sous-estimation variance erreur dans le cas LS.
- On fait ***I*** imputations de VM et ensuite ***I*** régressions, l'estimation finale des paramètres est le moyenne des ***I*** estimations
- Variance de l'estimateur est la somme de 2 termes :

  - moyenne des ***I*** variances de chaque estimateur

  - variance inter-estimateurs (aléa liés aux imputations)

$$\hat{v}^2 = s_w^2 + (1 + I^{-1})s_b^2, \qquad s_w^2 = \Sigma\hat{v}_m/I \qquad s_b^2 = \Sigma(\hat{\theta}_m - \hat{\theta})^2/(I-1)$$

6 – MI bis

- Les estimateurs entrant dans la composition de la moyenne sont de types Buck ou ML (EM)
- Dans ce cas conditionner les imputations par Y ne biaise pas l'estimateur final, conditionner par les X seules biaise (S.1. coeff reg de X1 atténué, on ajoute du bruit sur VX1, sans modifier la corré partielle entre X1 et Y)
- Si imputations avec modèle explicite, MI converge vers ML quand *n* et *I* augmentent
- MI peut aussi être élaboré sur la base de modèle implicite (hot deck)

Modèle non-normal

- Méthode ML, l'hypothèse de normalité n'est pas nécessaire sur les variables complètement observées, cas S1 par exemple l'hypothèse de normalité sur la distribution jointe Y, X1 suffit (convergence vers les paramètres de distribution multi-normales)
- Sous MCAR, estimation ML peut être valide en l'absence de normalité, mais pas forcément efficace (ex de WLS meilleur sur une distribution aplatie)
- Certains auteurs montrent que EM peut être adapté aux méthodes ML de distribution ***t***

Pistes d'investissement

- Les méthodes suppose un mécanisme MAR, non-MAR plus sensible aux erreurs de spécification
- Méthodes d'inférence sur les petits échantillons (méthodes Bayésiennes)
- Méthodes pour les données non-normales
- Méthodes pour les modèles non linéaires et linéaires généralisés

Conclusions

- CC méthode la plus commune et simple, acceptable quand le nombre de VM faible
- AC ne peut pas être généralisée, car dépend des données (corrélation entre X)
- Méthodes fondées sur les modèles comme ML, MI et MB préférables car exploitent toute l'information et sur lesquelles on peut établir l'inférence stat