Estimation de la probabilité d'avoir une donnée manquante

Denys Pommeret

Estimation de la probabilité d'avoir une donnée manquante

Denys Pommeret

Institut de Mathématiques de Luminy, Case 907, Université de la Méditerranée 13288 Marseille cedex 9. Email: pommeret@iml.univ-mrs.fr

Abstract

En présence d'une variable de loi connue dont les données manquantes ne sont pas dues au hasard (No Missing At Random) nous proposons une méthode pour estimer les probabilités d'avoir une donnée manquante suivant les valeurs observées de la variable. Nous en déduisons une statistique pour tester si les données manquent au hasard ou non.

In this paper we first propose an estimation of the probability of missing for No Missing At Random data when the distribution of the valid variable is known. Second we construct a statistic for testing the random character of the missing; that is, Missing At Random versus NMAR.

Mots clés: données manquantes ; estimation non paramétrique ; test d'indépendance.

1 Introduction

Le problème des valeurs manquantes pour une ou plusieurs variables est fréquemment rencontré lorsque l'on analyse un jeu de données. Il est alors important de connaître le processus qui engendre ces données manquantes. On distingue en général trois cas (voir [2]) : Missing At Random (MAR) ce qui signifie que la donnée manque indépendamment de la valeur (manquante) de la variable. Missing Completely At Random, ce qui signifie que la donnée manque au hasard, indépendamment de toutes les valeurs des variables (s'il y en a d'autres). No Missing At Random signifie que les données manquantes sont directement liées à la valeur prise par la variable (manquante).

Nous nous plaçons dans ce dernier cas : NMAR avec seulement une seule variable (manquante ou non) observée. Une généralisation de ce travail à la dépendance avec d'autres covariables est envisagée. Notre objectif est de proposer un estimateur de la probabilité d'avoir une donnée manquante, puis construire un test pour décider du caractère aléatoire ou non du processus engendrant les données manquantes. Nous testons donc MAR contre NMAR. Pour cela nous proposons une modélisation particulière qui s'adapte à l'étude d'une variable continue Y. Notre idée est de poser $X = YW$, avec W de Bernoulli prenant la valeur 0 lorsque la valeur est manquante et 1 sinon. La loi de

Y est supposée connue. Il est clair que l'indépendance entre Y et W va coïncider avec le modèle MAR. Cette indépendance se traduit également par une probabilité constante $P(W=1|Y)$. Nous allons estimer cette probabilité. Nous proposons ensuite un test sur le mécanisme générant les données manquantes.

2 Estimation de la probablité d'une donnée manquante

Considérons des variables continues $Y_1, \cdots, Y_n$, i.i.d., à valeur dans un intervalle I (éventuellement $\mathbb{R}$). Supposons que l'on observe

$$X_i \;=\; Y_i W_i,$$

où W_i indique une donnée manquante en prenant la valeur 0. Sinon W_i vaut 1. On a

$$W_i|Y_i \;=\; \left\{ \begin{array}{lll} 1 & \text{avec} & p(Y_i) = P(W_i=1|Y_i) \\ 0 & \text{avec} & 1-p(Y_i) = P(W_i=0|Y_i) \end{array} \right.$$

Nous avons donc un modèle NMAR, sauf lorsque p est constante. On suppose que la variable Y a une loi de probabilité connue, notée μ. Il est facile alors d'exprimer p (qui est bornée et donc de carré intégrable par rapport à μ) dans une base orthonormée $\mathcal{B} = \{Q_n; n = 0, 1, \cdots\}$ de $L^2(\mu)$. Le fait que $\mathbb{E}(X^k) = \mathbb{E}(Y^k W^k) = \mathbb{E}(Y^k p(Y))$ nous incite alors à choisir une base de polynômes.

PROPOSITION 2.1 *Pour $y \in I$ on a :*

$$p(y) \;=\; \mathbb{E}(p(Y)) + \sum_{k>0} \big(\mathbb{E}(Q_k(X)) + Q_k(0)(\mathbb{E}(p(Y)) - 1)\big) Q_k(y).$$

Démonstration : p admet le développement suivant :

$$\begin{aligned} p(y) &= \sum_{k\in\mathbb{N}} \int p(t) Q_k(t) \mu(dt) Q_k(y) \\ &= \sum_{k\in\mathbb{N}} \mathbb{E}\big(Q_k(Y) p(Y)\big) Q_k(y) \end{aligned}$$

En utilisant le fait que pour $k \in \mathbb{N}^*$, $\mathbb{E}(X^k) = \mathbb{E}(Y^k W) = \mathbb{E}(Y^k p(Y))$ nous obtenons

$$\begin{aligned} \mathbb{E}\big(Q_k(X)\big) &= \mathbb{E}\big(Q_k(X) - Q_k(0)\big) + Q_k(0) \\ &= \mathbb{E}\big((Q_k(Y) - Q_k(0)) p(Y)\big) + Q_k(0), \end{aligned}$$

ce qui donne le résultat. □

Comme $\mathbb{E}(p(Y)) = \mathbb{E}(W)$ nous pouvons réécrire le résultat de la Proposition 2.1 :

$$p(y) \;=\; \sum_{k\geq 0} \big(\mathbb{E}(Q_k(X)) + Q_k(0)(\mathbb{E}(W)-1)\big)Q_k(y).$$

Ainsi, pour $K \in \mathbb{N}^*$ fixé, un estimateur d'ordre K de p est donné par :

$$\widehat{p}_K(y) \;=\; \sum_{k=0}^{K} \big(E_k + Q_k(0)(C-1)\big)Q_k(y),$$

avec

$$E_k = \frac{1}{n}\sum_{i=1}^{n} Q_k(X_i), \qquad C = \frac{1}{n}\sum_{i=1}^{n} W_i.$$

PROPOSITION 2.2 *Posons*

$$p_K(y) \;=\; \sum_{k\leq K} \big(\mathbb{E}(Q_k(X)) + Q_k(0)(\mathbb{E}(W)-1)\big)Q_k(y).$$

et notons $\|\|$ *la norme dans* $L^2(\mu)$. *On a*

$$\|p(y) - \widehat{p}_K(y)\|^2 \;\leq\; \|p(y) - p_K(y)\|^2 + \frac{K+1}{n}$$

Démonstration : Il suffit d'utiliser les propriétés d'othogonalité des polynômes Q_n pour avoir

$$\|p(y) - \widehat{p}_K(y)\|^2 \;=\; \|p(y) - p_K(y)\|^2 + \frac{1}{n}\sum_{i\leq K} Var(Q_i(Y)p(Y)),$$

et le résultat en découle. □

On en déduit un estimateur convergent de p, $\widehat{p}_K$, en choisissant $K = K(n)$ avec $\lim_{n\to\infty} K(n) = \infty$ et $K(n) = o(n)$. On peut voir également que $(\sqrt{n}(\widehat{p}_i(y) - p_i(y)))_{i=1,\cdots,K}$ converge vers un vecteur Gaussien.

3 Test du caractère aléatoire des données manquantes

On suppose que la loi de Y est connue et on s'intéresse aux hypothèses

$$H_0 : Y, W \text{ indépendantes} \quad VS \quad H_1 : Y, W \text{ dépendantes}$$

ce qui revient à décider du caractère MAR ou NMAR des données manquantes. L'hypothèse H_0 revient à écrire $p(y) = cste = \mathbb{E}(p(Y))$, ou encore de manière équivalente, $\mathbb{E}(Q_n(X)) = Q_n(0)(1 - \mathbb{E}(p(Y)))$, $\forall n = 1, 2, \cdots$. Considérons le vecteur

$$U_k \quad = \quad (a_1, \cdots, a_k),$$

avec

$$a_k \quad = \quad 1/\sqrt{n} \sum_{i=1}^{n} (Q_k(X_i) + W_i - 1)/Q_k(0).$$

Par le TCL, nous avons la convergence suivante sous H_0:

$$T_k = \Sigma_k^{-1/2} U_k^T \quad \longrightarrow_{\mathcal{L}} N(0, I),$$

où Σ_k est la matrice de variance covariance $k \times k$: $Var(U_k)$. On peut alors montrer que sous H_0

$$\Sigma_{ij} \quad = \quad \{\mathbb{E}(W)\,[\delta_{ij} - 1 + Q_i(0) + Q_j(0) - Q_i(0)Q_j(0)] + Q_i(0) + Q_j(0) - 1\}/(Q_i(0)Q_j(0)),$$

que l'on estime simplement en remplaçant $\mathbb{E}(W)$ par C. Remarquons que la matrice Σ dépend de Y à travers le choix de la base $\mathcal{B}$. Ici le choix de k est arbitraire. Pour construire notre test nous nous inspirons de l'idée de sélection automatique proposée par [3]. Nous allons considérer un nombre $k(n)$ fonction de la taille de l'échantillon tel que $\lim_{n\to\infty} k(n) = \infty$. Ensuite nous utilisons le critère de Schwarz en posant

$$S_n = \min \Big\{ \operatorname*{argmax}_{1 \le k \le k(n)} (T_k - k \log(n)) \Big\}.$$

La statistique de test est alors T_{S_n}.

PROPOSITION 3.1 *Soit $\lambda_{k(n)}$ la plus petite valeur propre de $W_{k(n)}$. Supposons que $\frac{\log k(n)}{\lambda_{k(n)}} = o_{\mathbb{P}}(\log n)$ et $\frac{k(n)}{\lambda_{k(n)}} = o_{\mathbb{P}}(n^{1/2})$. Alors sous H_0, T_{S_n} converge en loi vers un Khi-deux à un degré de liberté.*

Démonstration : La démonstration s'inspire de [1] et de [2]. □

En pratique, il suffit de fixer $k(n)$ assez grand pour mettre en oeuvre la statistique T_{S_n}.

4 Quelques extension

- On peut s'intéresser également au problème inverse suivant : tester l'adéquation de la loi de Y à une loi connue lorsque la probabilité p est connue. La méthode consiste alors à reconstruire la densité de X dans une base de $L^2(\mu)$ et à la comparer à la densité sous H_0.

- On peut également s'intéresser à un modèle logistique :

$$p(y) = \frac{\exp\{\beta_0 + \beta y\}}{1 + \exp\{\beta_0 + \beta y\}}$$

  et en déduire des estimations des paramètres.

$$\log\left(\frac{\widehat{p}(y)}{1-\widehat{p}(y)}\right) = \widehat{\beta}_0 + \widehat{\beta} y$$

- Un travail futur serait de considérer des covariables associées à Y et liées (ou non ?) au mécanisme des données manquantes.

Bibliographie

[1] Ghattas, B. Pommeret, D. Reboul, L. et Yao, A.F. (2009) Smooth test for paired population. Soumis.
[2] Janic-Wróblewska, A. et Ledwina, T. (2000) Data Driven Rank Test for Two Sample Problem, *Scandinavian Journal of Statistics*, 27, 281–297.
[3] Ledwina, T. (1994) Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, 89, 1000–1005.
[4] Little and Rubin (1987) *Statitical Analysis with Missing Data*, New York: John Wiley & Sons.