Missing probability estimation

Denys Pommeret
Aix-Marseille University

ECODEP Ecology and Dependence Project

Sept. 2024

Granted by the Research Chair ACTIONS under the aegis of BNP Paribas Cardif

Institut de Mathématiques de Marseille
I2M - UMR CNRS 7373
Aix-Marseille Université • CNRS • École Centrale de Marseille

Chaire ACTIONS
Actuaries for Change in Technologies and Insurees Opportunities for Next Steps

M?s??ng probab?l?ty est??at??n

Den?s Pomm?ret
Aix-Mars??lle Univers???

?C?D?P Ec?l?gy and De??dence Pr?j?ct

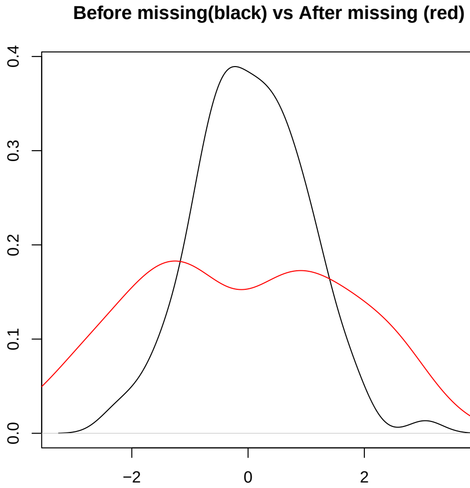
Sept. 2024

Grant?d by t?? Research Chair ?CT?ONS under the aeg?s of BNP Pa??bas Card?f

Institut de Mathématiques de Marseille
I2M - UMR CNRS 7373
Aix-Marseille Université • CNRS • École Centrale de Marseille

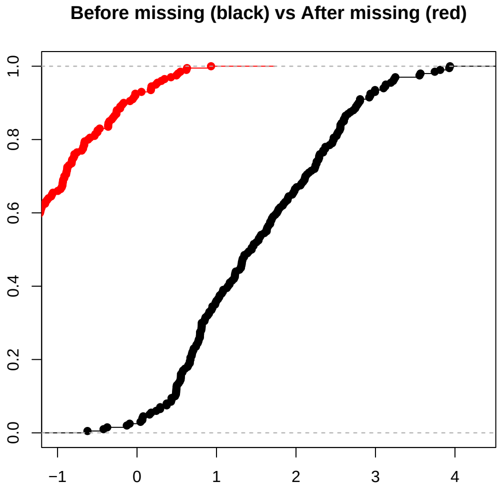
Chaire ACTIONS
Actuaries for Change in Technologies and Insurees Opportunities for Next Steps

Illustrative introduction



We observe before and after missing $\hookrightarrow$ can we use the difference between pdf to estimate missing probability?

Illustrative introduction



Can we use the difference between cdf to estimate missing cdf?

Illustration

$$\left\{\begin{array}{l}\text{One sample with } Y \text{ completely observed}\\ \text{One sample with } Y \text{ with missing values}\end{array}\right. \quad \Rightarrow \quad \mathbb{P}(Y \text{ missing})?$$

Comparing pdf or cdf seems too complex if the missing mechanism depends on other variables $X_1, X_2, \cdots$

Several questions

Several questions arise:

- Understanding the missing mechanism (is it due to the unknown value? to other variables?)
- Estimate the missing probability (there is little work, and only in the parametric case)
- Imput missing values (many works, depending on the missing mechanism)

Notation

We observe continuous random variables X, Y.
X and Y can be multidimensional.

There is a missing mechanism acting on Y.

Write C_Y the indicator of missing value:

- Y missing $(C_Y = 0)$
- or Y observed $(C_Y = 1)$.

Mechanism of missing value

Rubin (1976)

- MCAR (Missing Completely At Random)
- MAR (Missing At Random)
- MNAR (Missing Not At Random)

MCAR

The missing mechanism is not related to X or Y. C_Y depends neither on X nor Y.

In this case we can remove the missing values.

But there is a loss of information!

MAR

The missing mechanism is related to X, but not to Y. C_Y depends on X (observed), but not on Y.

We can try a regression model to reconstruct Y.

MNAR

The missing mechanism is related to both X and Y. C_Y depend on both X (observed) and Y (unobserved).

A major problem, but few solutions.

Missing probability

X and Y are continuous.

Copula transformation to get uniform distributions:

$$U = F_X(X),\ V = F_Y(Y),\ Z = C_Y V,$$

Z can be observed ($C_Y = 1$) or not ($C_Y = 0$).

$$Y \text{ missing} \quad \Leftrightarrow \quad Z \text{ missing.}$$

Assumption

We know (or we estimate) the cdf of Y.

$\hookrightarrow$ We need to have observed Y otherwise!!

Strong hypothesis, but realistic...

- We need to have information about Y through a survey, another sample, a signal before the loss of data, etc.
  $\hookrightarrow$ We then estimate the cdf of the missing variable Y.
- This is the price we have to pay.
- We also estimate the cdf of the (fully observed) variable X.

Missing probability

Copula transformation:

$$U = F_X(X),\ V = F_Y(Y),\ Z = C_Y V,$$

Associated missing probability:

$$p(u,v) = \mathbb{P}(C_Y = 1 | U = u, V = v) = \mathbb{E}(C_Y | U = u, V = v),$$

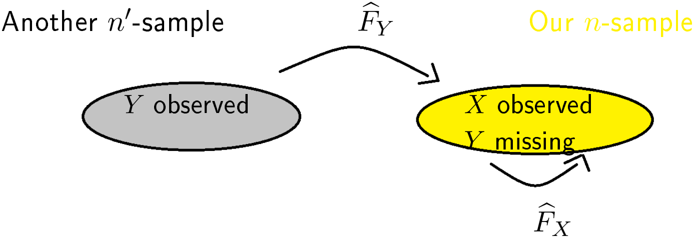
Original data:

$$\mathbb{P}(C_Y = 1 | X = x, Y = y) \quad = \quad p(F_X(x), F_Y(y)).$$

$L^2([0,1])$ basis

- Why copula transformation?
  $\hookrightarrow$ Uniform distributions
  $\hookrightarrow$ We know an orthonormal basis of $L^2([0,1])$: $\{L_k; k \in \mathbb{N}\}$ the set of Legendre orthonormal polynomials.
- Why orthonormal polynomials?
  $\mathbb{E}((C_Y V)^\ell) = \mathbb{E}(C_Y^\ell V^\ell) = \mathbb{E}(C_Y V^\ell) = \mathbb{E}(p(U,V)V^\ell)$

CDF estimations



Another n'-sample

$\widehat{F}_Y$

Our n-sample

Y observed

X observed

Y missing

$\widehat{F}_X$

- The empirical cdf of Y is based on an independent sample of size n' such that our sample size n satisfies $n/n' \to l < \infty$.
- The empirical cdf of X is based on our sample (since X is observed).

In the following, we can change a cdf F with its empirical estimator $\widehat{F}$ without modifying the asymptotic results.

MAR case

C_Y depends on the observed variable X, and is independent of Y. We have

$$U = F_X(X).$$

In that case we simply write

$$p(u) \quad = \quad \mathbb{P}(C_Y = 1 | U = u).$$

Proposition

For all $u \in (0,1)$*, we have:*

$$\begin{aligned} p(u) \quad &= \quad \mathbb{E}(C_Y) + \sum_{k>0} \{\mathbb{E}(L_k(U)C_Y)\} L_k(u) \\ &:= \quad \sum_{k\geq 0} \alpha_k L_k(u). \end{aligned}$$

Approximation & Estimation

Kth order approximation

$$p_K(u) \quad = \quad \sum_{k\leq K} \alpha_k L_k(u).$$

Kth order estimation

$$\widehat{p}_K(u) \quad = \quad \sum_{k\leq K} \widehat{\alpha}_k L_k(u),$$

where

$$\widehat{\alpha}_k \quad = \quad \frac{1}{n}\sum_{i=1}^{n} L_k(U_i) = \frac{1}{n}\sum_{i=1}^{n} L_k(\widehat{F}_X(X_i)).$$

$\hookrightarrow$ We need to add a constraint to stay in $]0,1[$.

Choosing the order: Part I

$$\widehat{p}_K(u) \;=\; \sum_{k \le K} \widehat{\alpha}_k L_k(u),$$

To choose (automatically) the order $K = K(n)$ we can use the asymptotic normality of the coefficients $\widehat{\alpha}_k$. A series of embedded test can be deployed.

Choosing the order: Part II

The choice of $K(n)$ can be based on a LASSO technique.

Indeed:

$$\begin{aligned} p_k(u) &= \sum_{k\leq K} \alpha_k L_k(u) \\ &\approx \mathbb{E}(C_Y = 1|u) \end{aligned}$$

and the α_k can be considered as regression coefficients on $L_1(u), \cdots, L_K(u)$.

Illustration (MAR case)

We consider a logit model:

$$\text{logit}\big(\mathbb{P}(C_Y = 1 | X = x, Y = y))\big) \quad = \quad ax + by + c,$$

with $a = -1$, $b = -1$, and $c = 1$.

MAR case $\hookrightarrow$ $b = 0$

Approximation

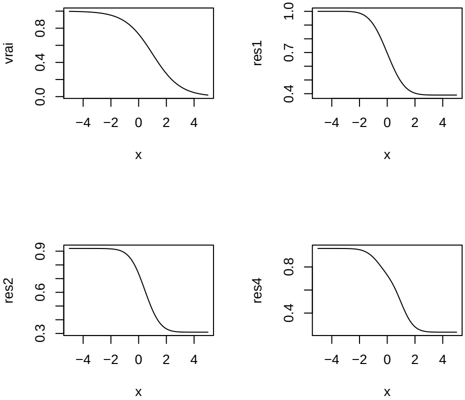


Figure – True and estimations for $K = 1, 2, 4$

Errors

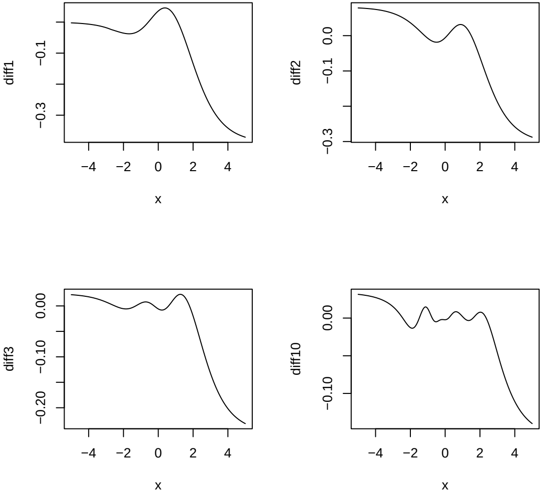


Figure – Errors for $K = 1, 2, 3, 10$

Univariate MNAR case

We first consider one variable Y. C_Y depends of Y. We have

$$V = F_Y(Y),\ Z = C_Y V,$$

and we simply write

$$p(v) \quad := \quad \mathbb{P}(C_Y = 1 | V = v).$$

Proposition

For all $v \in (0,1)$*, we have:*

$$\begin{aligned} p(v) &= \mathbb{E}(C_Y) + \sum_{k>0} \left\{ \mathbb{E}(L_k(Z)) + L_k(0)\mathbb{E}(C_Y - 1) \right\} L_k(v) \\ &:= \sum_{k \geq 0} \beta_k L_k(v) \end{aligned}$$

Approximation & Estimation

Kth order approximation

$$p_K(v) \quad = \quad \sum_{k \le K} \beta_k L_k(v).$$

Kth order estimation

$$\widehat{p}_K(v) \quad = \quad \sum_{k \le K} \widehat{\beta}_k L_k(v),$$

where

$$\widehat{\beta}_k \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \{L_k(Z_i)) + L_k(0)(C_{Y_i} - 1)\}.$$

Illustration (univariate MNAR case)

We consider a logit model:

$$\text{logit}\big(\mathbb{P}(C_Y = 1 | X = x, Y = y))\big) \quad = \quad ax + by + c,$$

with $a = -1$, $b = -1$, and $c = 1$.

Univariate MNAR case $\hookrightarrow a = 0$

Approximation

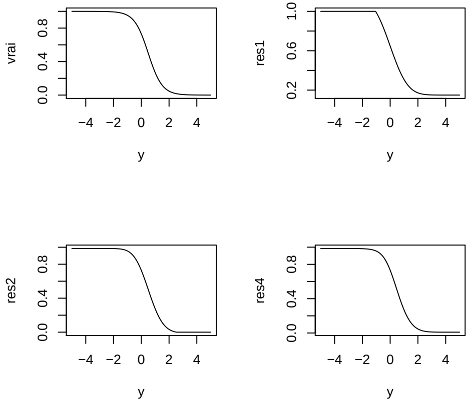


Figure – True and estimations for $K = 1, 2, 4$

Errors

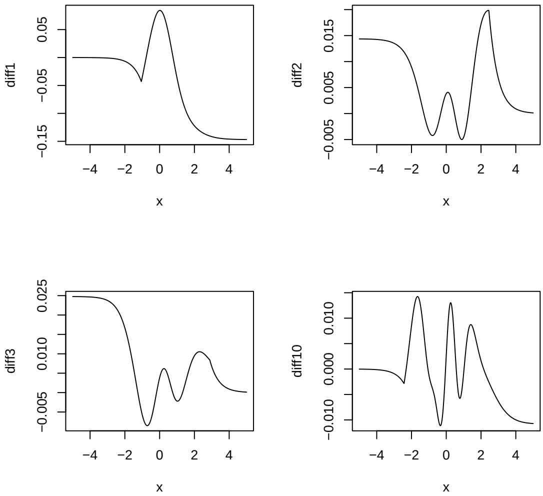


Figure – Errors for $K = 1, 2, 3, 10$

MNAR (general case)

C_Y depends on both X and Y. We have

$$U = F_X(X),\ V = F_Y(Y),\ Z = C_Y V,$$

Proposition

For all $(u,v) \in [0,1]^2$*, we have:*

$$p(u,v) = \mathbb{E}(C_Y) + \sum_{(k,\ell)\neq(0,0)} \mathbb{E}\big\{L_k(U)(L_\ell(Z) + L_\ell(0)(C_Y - 1))\big\} L_k(u) L_\ell(v).$$

Approximation & Estimation

To simplify we define the Kth order approximation as:

$$p_K(u) = \sum_{k\leq K;\ell\leq K} \widehat{\alpha}_{k,\ell} L_k(u) L_\ell(v),$$

and its associated estimator

$$\widehat{p}_K(u) = \sum_{k\leq K;\ell\leq K} \widehat{\alpha}_{k,\ell} L_k(u) L_\ell(v).$$

MISE

We consider the MISE (Mean Integrated Square Error) criterion to evaluate the behavior of the estimators:

$$MISE(\widehat{p}_K) \quad := \quad \mathbb{E}\big(\|p - \widehat{p}_K\|^2\big),$$

where

$$\|p\|^2 \quad := \quad \int_{[0,1]^2} p(u,v)^2 du dv.$$

Corollary

Let $K = K(n) = o(n^{1/4})$, such that $K(n) \to \infty$ as n tends to infinity. Then

$$MISE(\widehat{p}_{K(n)}) \to 0, \text{as } n \text{ tends to infinity.}$$

Illustration

$$\mathbb{P}(C_Y = 1 | X = x, Y = y)) \quad = \quad \frac{|a * x + b * y|}{|a * x + b * y| + c}$$

with $a = 1$, $b = -1$, and $c = 1$.

We fix x or y to represents a plot of probabilities.

Estimation

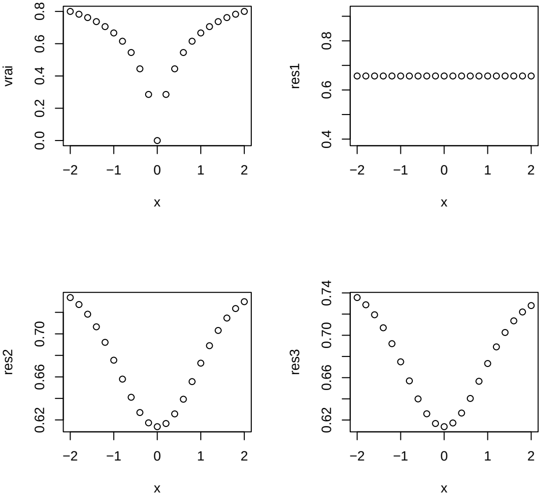


Figure – True and estimations for $K = 0, 2, 3$

Conclusion

In conclusion, this approach can be used to understand the underlying non-response mechanism when the variable of interest has been observed (independently) elsewhere. This non-response can also be seen as presence or absence, life or death, and ultimately as censorship. For example, in ecology, if we observe organisms that have survived a certain environment, or species that have migrated. We can estimate the probability of migration, or death, as a function of individual characteristics.

Perspective: multiple imputation

We want to apply a model to (X, Y).

Given v (that is, X), we use the probabilities of $\widehat{p}_K(u, v)$ to run M simulations and apply M models to obtain M intermediate results, which we combine to obtain a final result.