Valores faltantes

El análisis de los valores faltantes es parte del análisis descriptivo de los datos.

Ejemplo MCAR

Para los n individuos se registra la presión arterial y para una submuestra aleatoria n' se registra el peso.

Ejemplo MAR

Para los n individuos se registra la presión arterial y sólo para aquellos que tienen presión alta se registra el peso.

Ejemplo NMAR.

Para los n individuos se registra la presión arterial y sólo para aquellos que tienen sobrepeso se registra el peso.

Ejemplo. Datos de la superficie del mar

5 Variables más año.

Table 3.1. Missing values on each variable

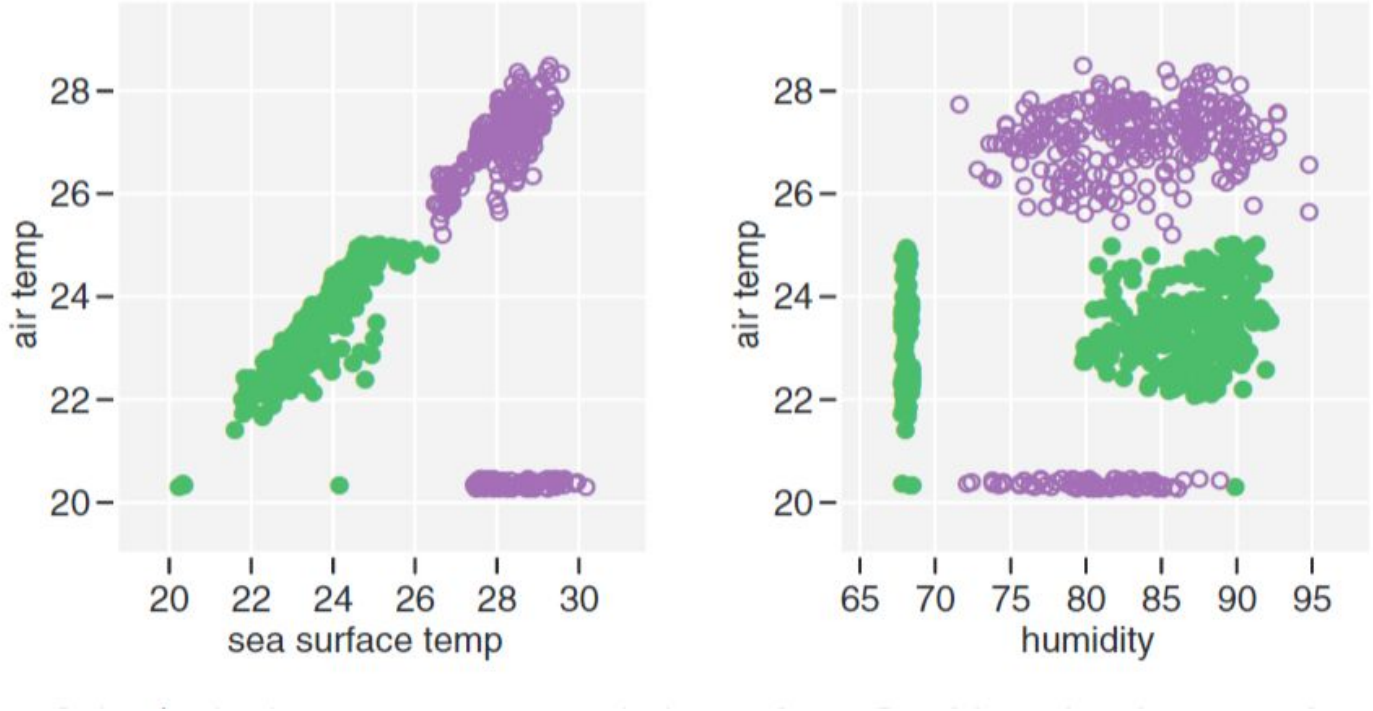
Variable	Number of missing values	
	1993	1997
sea surface temp	3	0
air temp	4	77
humidity	93	0
uwind	0	0
vwind	0	0

Una manera de reportar los datos faltantes

Table 3.2. Distribution of the number of missing values on a case.

No. of missings on a case	1993		1997	
	No. of cases	%	No. of cases	%
3	2	0.5	0	0
2	2	0.5	0	0
1	90	24.5	77	20.9
0	274	74.5	291	79.1

Asignar NA=0.1*min(xj)



Qué se ve en la gráfica.

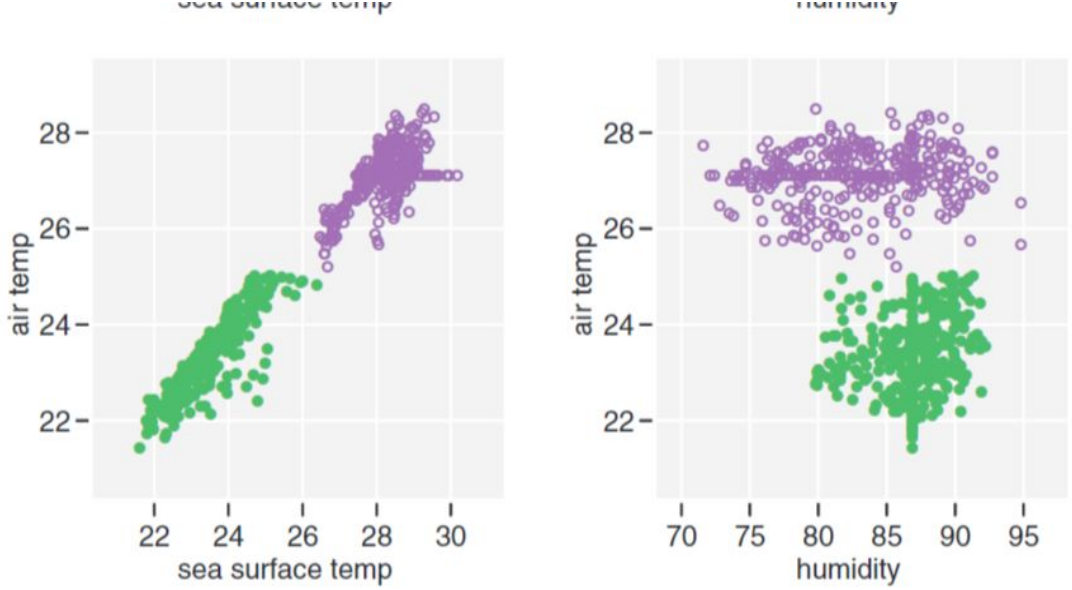
Gráfica izquierda.
Casi no hay faltantes de sea surface temp.
Aquí para que los 3 casos no se amontonen en un solo punto, se les sumó un error aleatorio pequeño.
Hay muchos casos faltantes para air temp.
La variable air temp se distribuye diferente para cada año.

Gráfica derecha.
Hay 93 faltantes de 368 mediciones de humidity en 1993, ninguna para 1997.
La distribución de humidity es muy distinta en cada año. Para 1993 tiene menos dispersión.

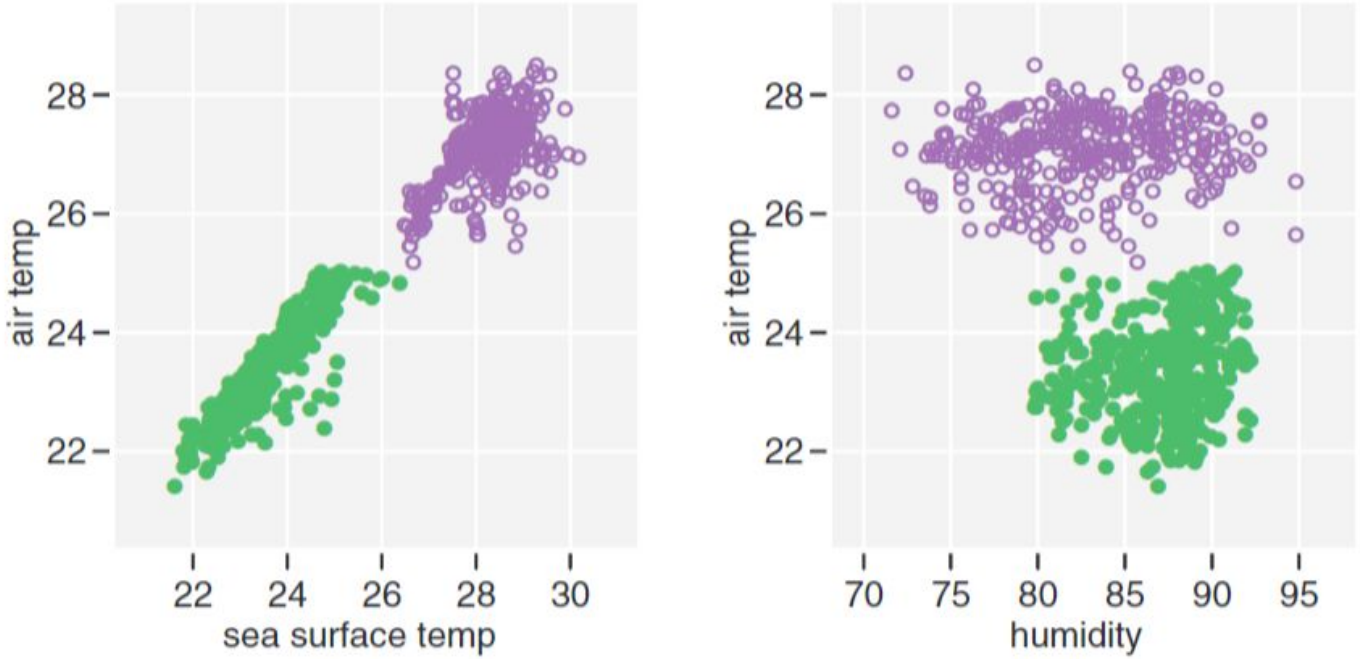
Pudimos ver que la distribución de los valores faltantes al menos en humidity y air temp está correlacionada con el año del registro, estos datos tienen más bien faltantes tipo MAR (el patrón de pérdidas depende de lo observado).

Podemos usar estas correlaciones para hacer imputación y no perder tanta información.

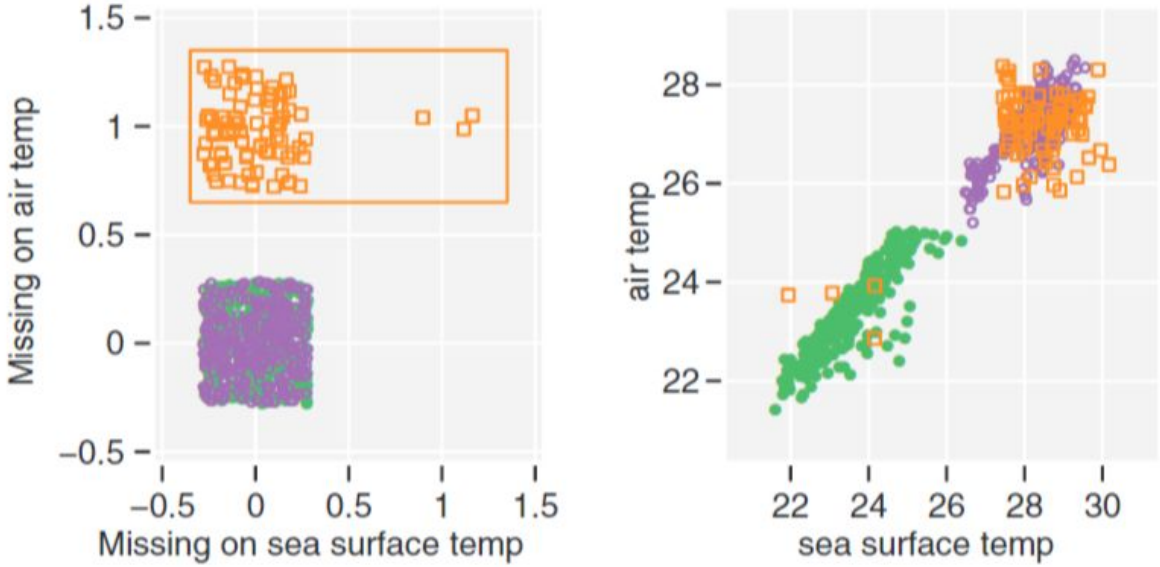
Imputación de NA con los medias



Imputación de NA con casos elegidos aleatoriamente, pero condicionando con el año



Ubicar dónde quedaron los casos imputados



Journal of Statistical Software

December 2011, Volume 45, Issue 3. http://www.jstatsoft.org/

mice: Multivariate Imputation by Chained Equations in R

Stef van Buuren
TNO

Karin Groothuis-Oudshoorn
University of Twente

$f(y|\theta)$ $f(y|\theta)$ $f(y|\theta)$

Paquete MICE hace imputación múltiple.

Usa el GIBBS Sampler.

Si **y** son las observaciones generadas por la distribución f(**y**|Θ) y $\boldsymbol{\Pi}(\Theta)$ es la distribución apriori en el espacio de parámetros de $\boldsymbol{\Theta}$. Uno de los objetivos en estadística bayesiana es encontrar la distribución aposteriori

$$\pi(\theta|y) = \frac{f(y|\theta) \cdot \pi(\theta)}{m(y)} \qquad m(y) = \int_{\Theta} f(y|\theta) \cdot \pi(\theta) d\theta$$

Algoritmo del *Gibbs sampler*

Initialize: pick arbitrary starting value $\theta^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, \cdots, \theta_i^{(1)}, \theta_{i+1}^{(1)}, \cdots, \theta_K^{(1)})$

Iterate a Cycle:

Step 1. draw $\theta_1^{(s+1)} \sim \pi(\theta_1 | \theta_2^{(s)}, \theta_3^{(s)}, \cdots, \theta_K^{(s)}, y)$

Step 2. draw $\theta_2^{(s+1)} \sim \pi(\theta_2 | \theta_1^{(s+1)}, \theta_3^{(s)}, \cdots, \theta_K^{(s)}, y)$

$\vdots$

Step i. draw $\theta_i^{(s+1)} \sim \pi(\theta_i | \theta_1^{(s+1)}, \theta_2^{(s+1)}, \cdots, \theta_{i-1}^{(s+1)}, \theta_{i+1}^{(s)}, \cdots, \theta_K^{(s)}, y)$

Step i+1. draw $\theta_{i+1}^{(s+1)} \sim \pi(\theta_{i+1} | \theta_1^{(s+1)}, \theta_2^{(s+1)}, \cdots, \theta_i^{(s+1)}, \theta_{i+2}^{(s)}, \cdots, \theta_K^{(s)}, y)$

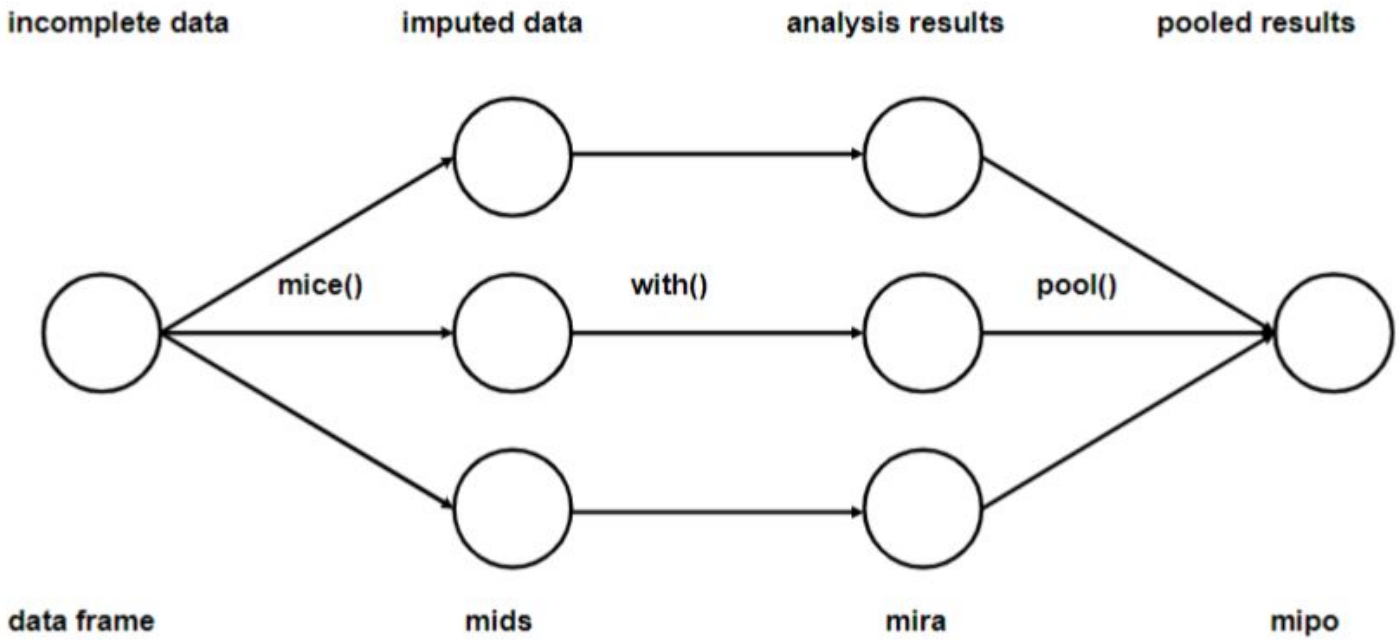
$\vdots$

Step K. draw $\theta_K^{(s+1)} \sim \pi(\theta_K | \theta_1^{(s+1)}, \theta_2^{(s+1)}, \cdots, \theta_{K-1}^{(s+1)}, y)$

end Iterate

La **imputación múltiple** es una propuesta que permite imputar los faltantes con varios valores plausibles, es decir considera varios escenarios. Estos valores plausibles son **extraídos de una distribución específicamente modelada** para cada entrada.

Con la información observada 1) se rellenan varios juegos de datos, 2) con cada juego de datos se hace el análisis de interés y 3) se combinan o integran los resultados.


incomplete data
imputed data
analysis results
pooled results
mice()
with()
pool()
data frame
mids
mira
mipo

Intentaré explicarles que hace el paquete MICE

Notación

Let Y_j with $(j = 1, \ldots, p)$ be one of p incomplete variables, where $Y = (Y_1, \ldots, Y_p)$. The observed and missing parts of Y_j are denoted by Y_j^{obs} and Y_j^{mis}, respectively, so $Y^{\text{obs}} = (Y_1^{\text{obs}}, \ldots, Y_p^{\text{obs}})$ and $Y^{\text{mis}} = (Y_1^{\text{mis}}, \ldots, Y_p^{\text{mis}})$ stand for the observed and missing data in Y. The number of imputation is equal to $m \geq 1$. The hth imputed data sets is denoted as $Y^{(h)}$ where $h = 1, \ldots, m$. Let $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$ denote the collection of the $p - 1$ variables in Y except Y_j. Let Q denote the quantity of scientific interest (e.g., a regression coefficient). In practice, Q is often a multivariate vector. More generally, Q encompasses any model of scientific interest.

Aquí suponen normalidad multivariada para Y.

Let the hypothetically complete data Y be a partially observed random sample from the p-variate multivariate distribution $P(Y|\theta)$. We assume that the multivariate distribution of Y is completely specified by θ, a vector of unknown parameters. The problem is how to get the multivariate distribution of θ, either explicitly or implicitly. The MICE algorithm obtains the posterior distribution of θ by sampling iteratively from conditional distributions of the form

$$P(Y_1|Y_{-1}, \theta_1)$$
$$\vdots$$
$$P(Y_p|Y_{-p}, \theta_p).$$

The parameters $\theta_1, \ldots, \theta_p$ are specific to the respective conditional densities and are not necessarily the product of a factorization of the 'true' joint distribution $P(Y|\theta)$. Starting from a simple draw from observed marginal distributions, the tth iteration of chained equations is a Gibbs sampler that successively draws

$$\theta_1^{*(t)} \sim P(\theta_1|Y_1^{\text{obs}}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)})$$

$$
\begin{aligned}
Y_1^{*(t)} &\sim P(Y_1 | Y_1^{\text{obs}}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}, \theta_1^{*(t)}) \\
&\vdots \\
\theta_p^{*(t)} &\sim P(\theta_p | Y_p^{\text{obs}}, Y_1^{(t)}, \ldots, Y_{p-1}^{(t)}) \\
Y_p^{*(t)} &\sim P(Y_p | Y_p^{\text{obs}}, Y_1^{(t)}, \ldots, Y_p^{(t)}, \theta_p^{*(t)})
\end{aligned}
$$

Este método MICE ha funcionado bien en varios estudios de simulación.

Brand 1999; Horton and Lipsitz 2001; Moons et al.2006;van Buuren et al.2006b; Horton and Kleinman 2007; Yu et al.2007; Schunk 2008; Drechsler and Rassler 2008; Giorgi et al.2008)