

Ejemplo imputación múltiple con MICE

```
> library(mice)
> head(nhanes)
  age  bmi hyp chl
1   1  NA  NA  NA
2   2 22.7   1 187
3   1  NA   1 187
4   3  NA  NA  NA
5   1 20.4   1 113
6   3  NA  NA 184
```

Estos son los primeros 6 renglones de un total de 25 . Datos de la librería mice.

age Age group (1=20-39, 2=40-59, 3=60+)

bmi Body mass index (kg/m**2)

hyp Hypertensive (1=no,2=yes)

chl Total serum cholesterol (mg/dL)

```
> md.pattern(nhanes)
  age hyp bmi chl
13   1   1   1   1   0
3    1   1   1   0   1
1    1   1   0   1   1
1    1   0   0   1   2
7    1   0   0   0   3
    0   8   9  10  27
```

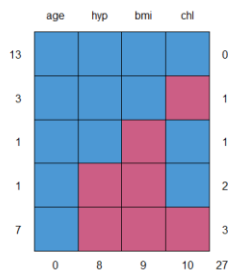
13 observaciones completas

3 que les falta medición de chl

1 que le falta bmi ,; 1 que le faltan hyp y bmi

7 que les faltan hyp, bmi y chl

$(7 \times 3) + (1 \times 2) + (3 \times 1) + (1 \times 1) = 27$



Este es el comando para correr la imputación múltiple. Fijamos la semilla inicial para que se pueda replicar .

```
> imp <- mice(nhanes, seed = 23109, print=FALSE)
```

```
> print(imp)
Class: mids
Number of multiple imputations: 5
Imputation methods:
  age  bmi  hyp  chl
    "" "pmm" "pmm" "pmm"
PredictorMatrix:
  age bmi hyp chl
age  0   1   1   1
bmi  1   0   1   1
hyp  1   1   0   1
chl  1   1   1   0
```

Habr  5 juegos de datos

Por ejemplo la variable

bmi ser  modelada a trav s de age+hyp+chl

```
> imp$imp$bmi
  1 2 3 4 5
1 28.7 27.2 22.5 27.2 28.7
3 30.1 30.1 30.1 22.0 28.7
4 22.7 27.2 20.4 22.7 20.4
6 24.9 25.5 22.5 21.7 21.7
10 30.1 29.6 27.4 25.5 25.5
11 35.3 26.3 22.0 27.2 22.5
12 27.5 26.3 26.3 24.9 22.5
16 29.6 30.1 27.4 30.1 25.5
21 33.2 30.1 35.3 22.0 22.7
```

Hay 9 valores faltantes de bmi

Aqu  est n los 5 juegos de 9 imputaciones

```
> imp$imp$hyp
```

```
1 2 3 4 5
1 1 2 1 1 1
4 2 2 2 1 1
6 2 2 2 1 1
10 1 2 1 1 1
11 1 1 1 1 1
12 1 1 1 1 1
16 1 1 1 1 1
21 1 1 1 1 1
```

Hay 8 valores faltantes de hyp

Aquí están los 5 juegos de 8 imputaciones

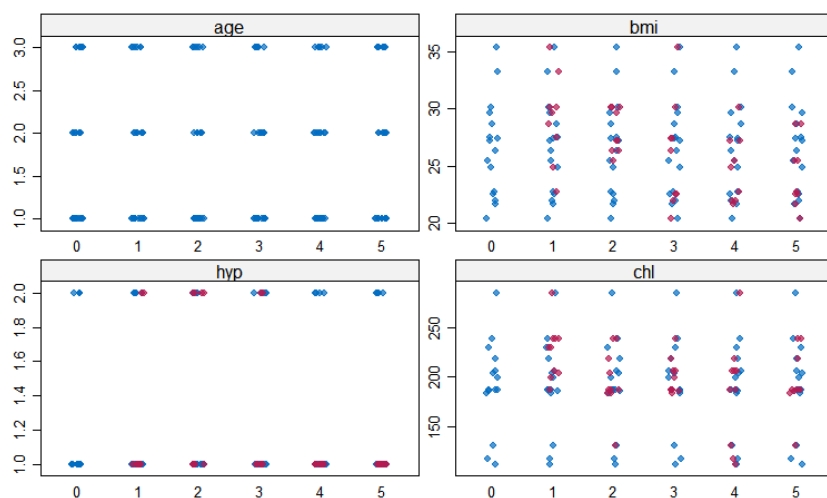
```
> complete(imp)
```

```
  age  bmi hyp chl
1   1 28.7  1 199
2   2 22.7  1 187
3   1 30.1  1 187
4   3 22.7  2 204
5   1 20.4  1 113
6   3 24.9  2 184
7   1 22.5  1 118
8   1 30.1  1 187
9   2 22.0  1 238
10  2 30.1  1 229
11  1 35.3  1 187
12  2 27.5  1 229
13  3 21.7  1 206
14  2 28.7  2 204
15  1 29.6  1 238
16  1 29.6  1 238
17  3 27.2  2 284
18  2 26.3  2 199
19  1 35.3  1 218
20  3 25.5  2 206
21  1 33.2  1 238
22  1 33.2  1 229
23  1 27.5  1 131
24  3 24.9  1 284
25  2 27.4  1 186
```

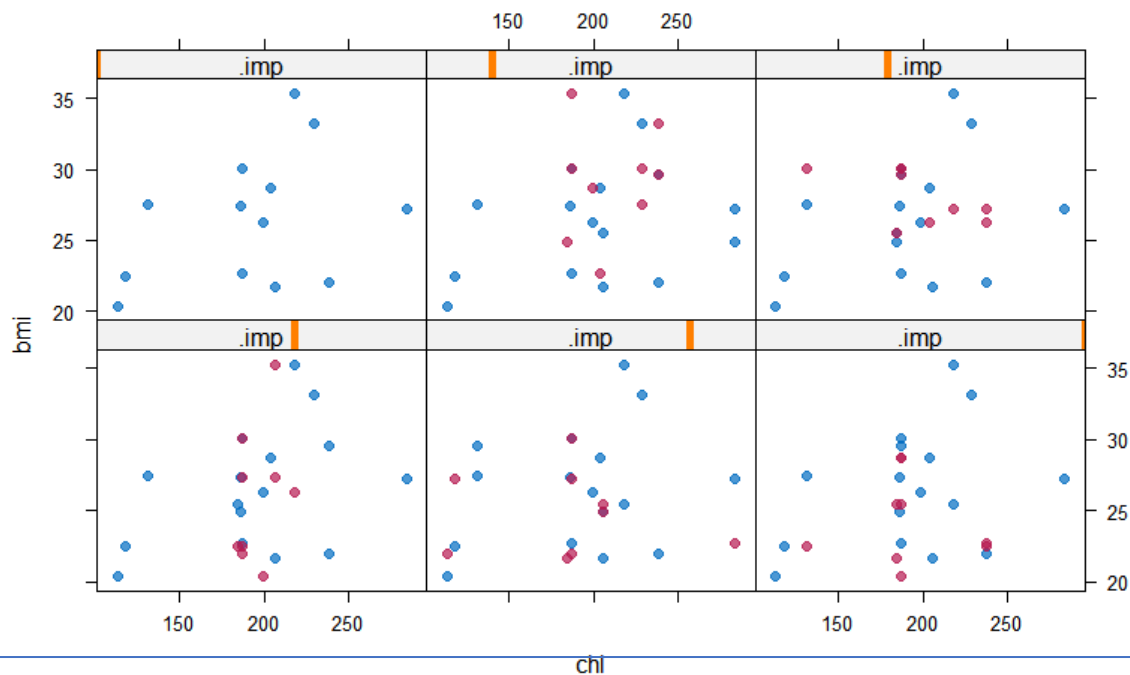
Este es un juego de datos completos. (el primero)

```
> library(lattice)
```

```
> stripplot(imp, pch = 20, cex = 1.2)
```



```
> xyplot(imp, bmi ~ chl | .imp, pch = 20, cex = 1.4)
```



Si la idea es correr un análisis de regresión para explicar los niveles de colesterol en función de la edad y el índice de masa corporal, este paquete *mice* permite hacerlo así.

```
> fit <- with(imp, lm(chl ~ age + bmi))
```

```
> print(pool(fit))
```

Class: mipo m = 5

	term	m	estimate	ubar	b	t	dfcom	df	riv	lambda	fmi
1	(Intercept)	5	5.958468	3575.717813	1649.43830	5555.043768	22	9.216954	0.5535465	0.3563115	0.4616878
2	age	5	29.725360	82.553435	115.72029	221.417779	22	4.331856	1.6821147	0.6271599	0.7288640
3	bmi	5	5.138790	3.686724	0.91985	4.790544	22	12.907767	0.2994040	0.2304164	0.3271721

```
> summary(pool(fit))
```

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	5.958468	74.532166	0.07994492	9.216954	0.93799003
2	age	29.725360	14.880114	1.99765683	4.331856	0.11101260
3	bmi	5.138790	2.188731	2.34783949	12.907767	0.03549532

$$\bar{\gamma} = \sum_{i=1}^D \gamma_i / D$$

Este es el estimador combinado

$$T_D = \bar{U}_D + (1 + 1/D)B_D$$

Esta es la varianza del estimador combinado

con

$$\bar{U}_D = \sum_{i=1}^D U_i / D \quad (\text{Promedio de varianzas})$$

y

$$B_D = \sum_{i=1}^D (\hat{\gamma}_i - \bar{\gamma})^2 / (D - 1) \quad (\text{Varianza entre imputaciones})$$