

# A stacked approach for chained equations multiple imputation incorporating the substantive model

Juan David Duitama  
Juan Felipe Moreno



# Antecedentes

# MAR: Missing at Random

El proceso de datos faltantes se dice MAR si condicionando a las variables observadas, el proceso es independiente a las no observadas, la falta de datos depende sólo a través de los componentes observados.

# MCAR: Missings completely at random

Un proceso de datos faltantes se define MCAR si la ausencia de datos no depende de los valores de los datos, ya sean faltantes u observados.

Si para todos los  $i$  y para cualquier conjunto de valores distintos  $y_i$  e  $y_i^*$  en el espacio muestral de  $Y$ ,

$$f_{M|Y}(m_i \mid y_i, \phi) = f_{M|Y}(m_i \mid y_i^*, \phi)$$


# MI: Multiple Imputation

Implica completar valores para los datos faltantes al extraer de distribuciones obtenidas a partir de una distribución conjunta asumida para todas las variables de interés.  $f(Y_i^{miss} | Y_i^{obs}, X_{i1}, X_{i2}, \dots, X_{ik})$



# MICE: Multiple imputation by chained equations

Implica especificar las distribuciones condicionales para cada variable con datos faltantes directamente (Raghunathan, 2001; Van Buuren et al., 2006).



## SMC-FCS (substantive model compatible fully conditional specification)

“Utiliza directamente la relación supuesta  $Y|X$  para incorporar  $Y$  en las distribuciones de imputación. En particular, la variable faltante  $X_p$  se imputa a partir de una distribución proporcional al modelo de resultado  $f(Y|X)$  multiplicado por una relación supuesta entre  $X_p$  y las demás covariables.” (Bartlett and Morris, 2015).

# Complete data set

Sea  $Y = (y_{ij})$  que denota un conjunto de datos rectangular ( $n \times K$ ) sin valores faltantes, es decir, un conjunto de datos completo,  $y_i = (y_{i1}, \dots, y_{iK})$  con la fila  $i$ -ésima, donde  $y_{ij}$  es el valor de la variable  $Y_j$  para la unidad  $i$ ." (Little & Rubin, 2019)



# Imputation Stacking

Múltiples imputaciones de los datos faltantes se apilan una sobre otra para crear un conjunto de datos grande (Robins y Wang, 2000; Van Buuren, 2018).

1. Generar múltiples imputaciones para los valores faltantes
2. Combinar las imputaciones en un conjunto de datos **apilado**, donde cada fila de en el conjunto de datos representa una imputación diferente y así mismo una estimación de los valores faltantes.
3. Se crea un conjunto de datos más grande y completo.
4. Se realiza un modelo con el caso completo.

ID	$y_i$	$x_{i1}$	$x_{i2}$
1	2.5	-	1.8
2	-	3.2	2.1
3	1.9	2.7	-

Conjunto de datos original  
con datos faltantes

ID	$y_i$	$x_{i1}$	$x_{i2}$
1	2.5	3.0	1.8
1	2.5	2.8	1.8
2	1.7	3.2	2.1
2	2.1	3.2	2.1
3	1.9	2.7	2.2
3	1.9	2.7	2.0

Tall-Stack

# Modelos de riesgos proporcionales de Cox

Con un riesgo de base  $\lambda_0$

$$\lambda(t, X_1, X_2, \dots, X_n) = \lambda_0(t) \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)$$

```
library("survival")
```

```
library("survminer")
```

```
coxph(formula,data,method)
```

- formula: is linear model with a survival object as the response variable. Survival object is created using the function *Surv()* as follow: *Surv(time, event)*.
- method: is used to specify how to handle ties. The default 'efron' is generally preferred.

Riesgo base



Método propuesto

# Paso 1: MICE

Imputar valores faltantes en las covariables sin considerar Y.

1. Utilizar MICE para obtener múltiples imputaciones de  $X_i$  basándose en una distribución asumida para  $f(X^{miss} | X^{obs})$ .
2. Implementar modelos de regresión para cada covariable con datos faltantes, excluyendo el resultado.

En casos faltantes en Y, realizar imputaciones de X ignorando Y primero y luego imputar valores faltantes de Y a partir de  $f(Y | X)$  para cada conjunto de datos imputado.

## Paso 2: Stack Imputations

Obtenemos una versión apilada de los datos, donde cada uno de los  $M$  conjuntos de datos imputados de tamaño  $n \times p$  se apilan uno encima del otro para formar un conjunto de datos  $Mn \times p$ , llamado el 'conjunto alto' o 'tall stack'. Se asume que los Missings son MAR.

# Paso 3 Asignación de Pesos

Aumentamos el conjunto de datos apilado con pesos definidos para cada fila como 1 dividido por el número de veces que aparece ese sujeto en el conjunto de datos apilado. En nuestra aproximación de apilamiento de imputación modificada, aumentamos el conjunto de datos apilado con una columna de pesos, donde los pesos están definidos de manera proporcional a  $f(Y_i|X_i)$ .

$$w_{im} = \frac{f(Y_i | X_{im}; \hat{\theta}_{cc})}{\sum_{j=1}^M f(Y_i | X_{ij}; \hat{\theta}_{cc})}$$

ID	$y_i$	$x_{i1}$	$x_{i2}$	$w_{im}$
1	2.5	3.0	1.8	$w_{1,1}$
1	2.5	2.8	1.8	$w_{1,2}$
2	1.7	3.2	2.1	$w_{2,1}$
2	2.1	3.2	2.1	$w_{2,2}$
3	1.9	2.7	2.2	$w_{3,1}$
3	1.9	2.7	2.0	$w_{3,2}$



# Paso 4

Estimar theta mediante un modelo  $Y|X$  en los datos apilados usando los pesos

# Método

$$I_{obs}(\hat{\theta}) \approx \sum_i E_{\hat{\theta}}[J_{com}^i(X_i, Y_i) \mid X_i^{obs}, Y_i] - \sum_i Var_{\hat{\theta}}[U_{com}^i(X_i, Y_i) \mid X_i^{obs}, Y_i]$$

Donde  $J_{com}$  es la contribución del i-ésimo individuo a la matriz de información de fisher.

$U_{com}$  la contribución del individuo  $i$  a la matriz de score.



SIMULACIONES



# Simulaciones

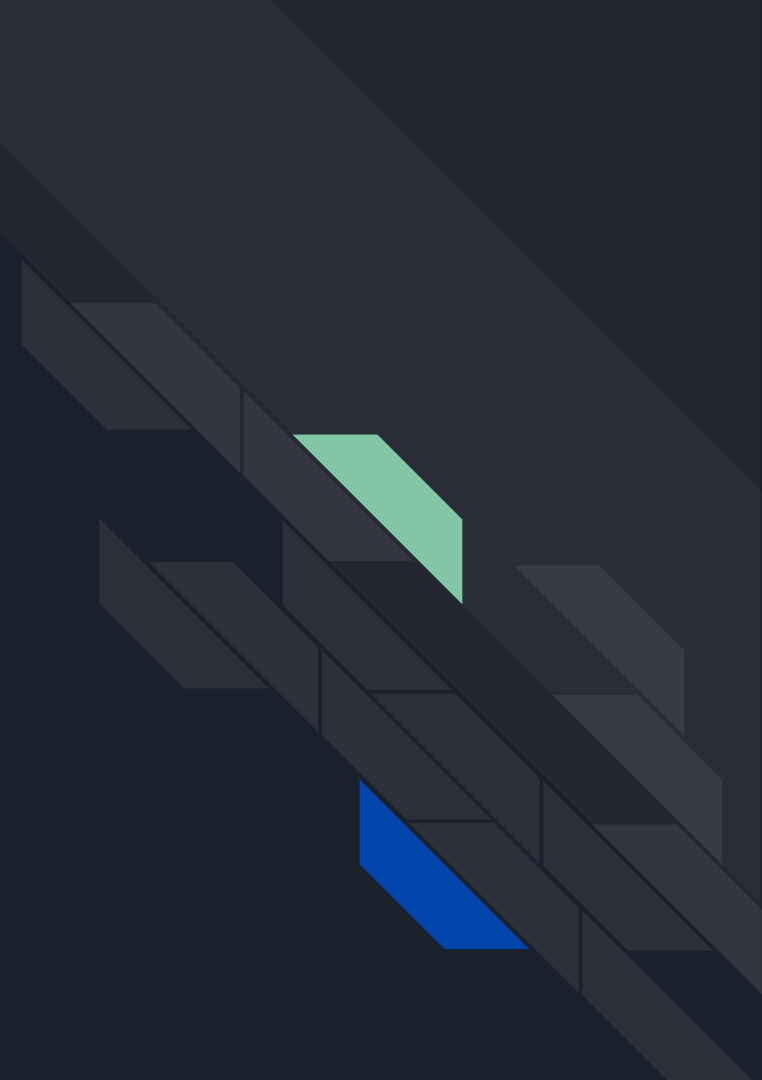
Exploración del desempeño de la estrategia de imputación propuesta y del estimador del error estándar.

Cuatro escenarios de estudio:

- Gaussiano Y con falta en una sola covariable X.
- Binario Y con falta en dos covariables.
- Gaussiano Y con falta en una sola covariable e interacciones en el modelo de resultados.
- Tipo de supervivencia censurado Y con ausencia en una sola covariable.

Cuatro mecanismos de falta diferentes: MCAR, MAR dependiente de X, MAR dependiente de Y, y MAR dependiente de X y Y.

# CONFIGURACIÓN DE LA SIMULACIÓN





## ESCENARIO 1

Se generaron las covariables  $X_1$  y  $X_2$  de una distribución normal multivariada con media 0, varianzas 0.49 y 0.09 y covarianza de 0.12.

$Y$  fue generada de una normal con media  $0.53X_1 + 1.25X_2$  y varianza 0.55.



# ESCENARIO 1

Cerca del 50% de las faltas en  $X_2$  fueron generadas usando el modelo  $\text{logit}(P(X_2 \text{ observado} | X_1, Y)) = t_0 + t_1 X_1 + t_2 Y$  con  $t = \{(0, 0, 0), (0, 1, 0), (0, 0, 1), (0, 1, -1)\}$ .

Estos valores de  $t$  corresponden a MCAR, MAR dependiente de  $X_1$ , MAR dependiente de  $Y$ , MAR dependiente de  $X_1$  y  $Y$ .



## ESCENARIO 2

Se generaron tres covariables  $X_1$ ,  $X_2$  y  $X_3$  de una distribución normal multivariada de medias 0, varianzas 1 y covarianza 0.3.

Y fue generado utilizando la relación

$$\text{logit}(P(Y=1|X_1, X_2, X_3)) = 0.5 + 0.5X_1 + 0.5X_2 + 0.5X_3$$



## ESCENARIO 2

Las ausencias de  $X_2$  fueron generadas utilizando el modelo del escenario 1 utilizando  $t = \{(0.5, 0, 0), (0.5, 1, 0), (0.5, 0, 1), (0.5, 1, -1)\}$  e independiente de  $X_3$ .

Cerca del 30% de las ausencias de  $X_3$  fueron generadas utilizando MCAR.

Esto produjo que alrededor del 40% de los individuos tuvieran toda la información completa, es decir, con información en  $X_1$ ,  $X_2$  y  $X_3$ .



## ESCENARIO 3

Se generaron 2 covariables  $X_1$  y  $X_2$  de una distribución normal multivariada con medias 0, varianzas 0.81 y 1.21, y covarianza 0.59.

Y fue generada de una normal con media  $X_1 + X_2 + X_1X_2$  y varianza 1.

Las ausencias de  $X_2$  fueron generadas como en el escenario 1.





## ESCENARIO 4

Las dos covariables  $X_1$  y  $X_2$  fueron generadas de una distribución normal multivariada de medias 0 y varianzas 1, con covarianza 0.5.

$T$  fue generada de una distribución exponencial con parámetro de escala  $e^{(0.5X_1 + 0.5X_2)}$ .

La censura Uniforme(0.2,3) fue impuesta en  $T$ .



## ESCENARIO 4

Alrededor del 50% de las ausencias de X2 fueron generadas bajo el modelo  $\text{logit}(P(X2 \text{ observado} | X1, Y)) = t_0 + t_1 X1 + t_2 h$  con

$$t = \{(0, 0, 0), (0, 1, 0), (-0.7, 0, 1), (-0.7, 1, -1)\}$$

h corresponde al indicador de censura o evento que hace parte de Y.



# IMPUTACIÓN

- 500 conjuntos de datos, 2000 individuos, 50 múltiples imputaciones.
- Reglas de combinación de Rubin o método de apilamiento propuesto.
- Uso de tres estimadores para el cálculo de las estimaciones del error estándar.
- Para el cálculo de los pesos en el escenario 4 se utilizó un modelo Cox.

# IMPUTACIÓN

	Standard MICE	Bartlett et al. (2014)	Stacked, 1/M weighted	Stacked, $f(Y X)$ weighted
Covariate Imputation	$f(X_p X_{-p}, Y)$ , specified as regression model	$f(X_p X_{-p}, Y) \propto f(Y X)f(X_p X_{-p})$ , where $f(X_p X_{-p})$ is a regression model	Often, same as MICE. Could also apply other imputation methods.	$f(X_p X_{-p})$ , specified as regression model
Point Estimation	Fit model to each imputed dataset separately	Fit model to each imputed dataset separately	Fit single weighted model to stacked imputations. *	Fit single weighted model to stacked imputations. Weights $\propto f(Y X)$
Standard Errors	Rubin's rules	Rubin's rules	Previously, unclear how to estimate. ** We propose new approach in Eq. 3.	We propose new approach in Eq. 3.
Comments	<ul style="list-style-type: none"> <li>↳ Easy to implement</li> <li>↳ Tricky to specify imputation regressions</li> </ul>	<ul style="list-style-type: none"> <li>↳ Limited outcome models supported by current software</li> <li>↳ Easy to implement for supported models</li> <li>↳ Outcome model built into imputation</li> </ul>	<ul style="list-style-type: none"> <li>↳ Inherits properties of imputation approach chosen</li> <li>↳ Different data analysis</li> <li>↳ Proposed new standard errors</li> </ul>	<ul style="list-style-type: none"> <li>↳ Imputation ignores <math>Y</math>. Easy to implement.</li> <li>↳ Imputation and analysis separated. Easy to compare outcome models.</li> </ul>
R Packages	<i>mice</i>	<i>smcfcs</i>	<i>mice</i> , <i>StackImpute</i> <sup>†</sup>	<i>mice</i> , <i>StackImpute</i> <sup>†</sup>

# RESULTADOS DE LA SIMULACIÓN



# SESGO - ESCENARIO 1

Missingness: <sup>†</sup>	Bias × 100 in effect of $X_1$				Bias × 100 in effect of $X_2$			
	MCAR	$X_1$	$Y$	$X_1, Y$	MCAR	$X_1$	$Y$	$X_1, Y$
Scenario 1: Linear Regression								
Full Data	0.02	0.01	0.14	0.28	-0.05	-0.15	-0.17	-0.20
Complete Case	-0.03	-0.05	-5.18	5.29	-0.16	0.18	-13.11	-13.59
MICE with $Y^*$								
↳ Rubin's rules	0.08	0.03	0.28	0.36	-0.41	0.02	-0.75	-0.30
↳ Stacked, 1/M weighted	0.11	0.07	0.32	0.39	-0.53	-0.12	-0.88	-0.41
MICE without $Y^*$								
↳ Rubin's rules	16.1	16.1	18.48	18.0	-62.6	-62.3	-69.09	-69.4
↳ Stacked, $f(Y X)$ weighted	0.32	0.27	0.60	0.66	-1.36	-0.88	-1.85	-1.46
Bartlett et al. (2014) <sup>✗</sup>	0.14	0.11	0.47	0.47	-0.61	-0.21	-1.38	-0.72

# SESGO - ESCENARIO 2

Missingness: <sup>†</sup>	Bias × 100 in effect of $X_1$				Bias × 100 in effect of $X_2$			
	MCAR	$X_1$	$Y$	$X_1, Y$	MCAR	$X_1$	$Y$	$X_1, Y$
Scenario 2: Logistic Regression								
Full Data	0.34	-0.03	0.09	0.13	0.24	-0.09	0.22	0.12
Complete Case	0.75	0.37	-0.12	21.0	0.18	-0.09	0.56	0.32
MICE with $Y$								
↳ Rubin's rules	0.35	-0.08	0.05	-0.07	-0.17	-0.60	0.17	-0.53
↳ Stacked, 1/M weighted	0.35	-0.08	0.04	-0.09	-0.26	-0.73	0.10	-0.72
MICE without $Y$								
↳ Rubin's rules	5.85	5.87	5.01	6.49	-18.49	-20.8	-14.5	-26.6
↳ Stacked, $f(Y X)$ weighted	0.49	0.11	0.13	0.30	-0.25	-0.61	0.12	-0.43
Bartlett et al. (2014)	0.42	0.05	0.09	0.08	0.12	-0.31	0.30	-0.19

# SESGO - ESCENARIO 3

Missingness: <sup>†</sup>	Bias × 100 in effect of $X_1$				Bias × 100 in effect of $X_2$			
	MCAR	$X_1$	$Y$	$X_1, Y$	MCAR	$X_1$	$Y$	$X_1, Y$
Scenario 3: Linear Regression with Interaction								
Full Data	0.10	0.10	0.29	-0.22	-0.14	-0.04	-0.30	0.26
Complete Case	0.21	-0.10	-8.97	-0.58	-0.36	-0.09	-9.90	-14.88
MICE with $Y$								
↳ Rubin's rules	-2.12	-13.9	-4.73	-7.99	-12.28	13.14	-1.35	-3.97
↳ Stacked, 1/M weighted	-2.07	-13.95	-4.70	-7.82	-12.40	13.11	-1.38	-4.29
MICE with $Y$ + interaction *	-2.75	18.93	-10.05	-17.52	-10.28	21.35	5.93	-10.14
MICE without $Y$								
↳ Rubin's rules	36.8	24.13	16.84	81.70	-50.20	-32.75	-35.32	-70.16
↳ Stacked, $f(Y X)$ weighted	0.05	0.05	-1.22	-1.24	-0.10	-0.08	-1.37	0.01
Bartlett et al. (2014)	0.38	0.19	0.35	0.40	-0.49	-0.22	-0.50	0.16



# SESGO - ESCENARIO 4

Missingness: <sup>†</sup>	Bias $\times 100$ in effect of $X_1$				Bias $\times 100$ in effect of $X_2$			
	MCAR	$X_1$	$Y$	$X_1, Y$	MCAR	$X_1$	$Y$	$X_1, Y$
Scenario 4: Cox Proportional Hazards Regression								
Full Data	0.12	0.04	-0.07	0.21	0.18	0.10	-0.01	0.15
Complete Case	0.12	0.07	-5.69	-9.07	0.07	0.26	-5.29	-4.31
MICE with $Y$								
↳ Rubin's rules	-1.62	-1.65	-2.04	-1.83	-4.18	0.37	-3.42	0.94
↳ Stacked, 1/M weighted	-1.61	-1.59	-2.02	-1.75	-4.30	0.27	-3.54	0.83
MICE without $Y$								
↳ Rubin's rules	0.48	1.58	0.95	2.59	-27.2	-25.02	-29.69	-27.47
↳ Stacked, $f(Y X)$ weighted	0.15	0.56	-0.18	0.91	-0.30	-2.43	-1.26	-2.47
Bartlett et al. (2014)	0.15	-0.05	-0.08	0.12	0.03	0.25	0.11	0.22

# VARIANZA EMPÍRICA RELATIVA - ESCENARIO 1

Missingness: <sup>†</sup>	Relative variance for effect of $X_1$				Relative variance for effect of $X_2$			
	MCAR	$X_1$	$Y$	$X_1, Y$	MCAR	$X_1$	$Y$	$X_1, Y$
Scenario 1: Linear Regression								
Full Data	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Complete Case	2.06	2.07	1.87	1.85	1.88	2.09	1.75	1.73
MICE with $Y^*$								
↳ Rubin's rules	1.35	1.37	1.45	1.31	1.70	1.85	1.98	1.90
↳ Stacked, 1/M weighted	1.35	1.37	1.45	1.31	1.70	1.85	1.97	1.90
MICE without $Y^*$								
↳ Rubin's rules	0.86	0.87	0.85	0.86	0.55	0.54	0.48	0.48
↳ Stacked, $f(Y X)$ weighted	1.34	1.37	1.45	1.31	1.69	1.83	1.95	1.89
Bartlett et al. (2014) <sup>✗</sup>	1.39	1.45	1.50	1.33	1.74	1.95	2.07	1.99

# VARIANZA EMPÍRICA RELATIVA - ESCENARIO 2

Missingness: <sup>†</sup>	Relative variance for effect of $X_1$				Relative variance for effect of $X_2$			
	MCAR	$X_1$	$Y$	$X_1, Y$	MCAR	$X_1$	$Y$	$X_1, Y$
Scenario 2: Logistic Regression								
Full Data	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Complete Case	2.52	2.29	2.02	4.08	2.36	2.46	2.15	3.66
MICE with $Y$								
↳ Rubin's rules	1.08	1.08	1.04	1.13	1.64	1.64	1.45	2.35
↳ Stacked, 1/M weighted	1.08	1.07	1.04	1.12	1.63	1.63	1.45	2.33
MICE without $Y^*$								
↳ Rubin's rules	0.93	0.95	0.92	0.94	0.54	0.45	0.60	0.43
↳ Stacked, $f(Y X)$ weighted	1.09	1.08	1.03	1.14	1.78	1.82	1.55	2.77
Bartlett et al. (2014)	1.09	1.09	1.05	1.14	1.73	1.74	1.52	2.58

# VARIANZA EMPÍRICA RELATIVA - ESCENARIO 3

Missingness: <sup>†</sup>	Relative variance for effect of $X_1$				Relative variance for effect of $X_2$			
	MCAR	$X_1$	$Y$	$X_1, Y$	MCAR	$X_1$	$Y$	$X_1, Y$
Scenario 3: Linear Regression with Interaction								
Full Data	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Complete Case	2.14	2.13	1.78	2.37	2.11	2.04	1.83	2.50
MICE with $Y$								
↳ Rubin's rules	2.85	2.12	1.34	5.20	3.16	3.35	1.62	4.02
↳ Stacked, 1/M weighted	2.85	2.12	1.34	5.21	3.16	3.35	1.62	4.05
MICE with $Y$ + interaction *	2.92	2.45	1.81	4.40	4.96	4.51	2.79	5.81
MICE without $Y$								
↳ Rubin's rules	2.25	1.69	1.16	4.54	1.03	0.77	0.86	0.85
↳ Stacked, $f(Y X)$ weighted	1.50	1.40	1.26	2.07	1.74	1.71	1.60	2.06
Bartlett et al. (2014)	1.52	1.46	1.29	2.07	1.75	1.60	1.55	1.99

# VARIANZA EMPÍRICA RELATIVA - ESCENARIO 4

Missingness: <sup>†</sup>	Relative variance for effect of $X_1$				Relative variance for effect of $X_2$			
	MCAR	$X_1$	$Y$	$X_1, Y$	MCAR	$X_1$	$Y$	$X_1, Y$
Scenario 4: Cox Proportional Hazards Regression								
Full Data	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Complete Case	1.85	2.20	1.64	2.05	2.13	1.81	2.12	1.79
MICE with $Y$								
↳ Rubin's rules	1.06	1.13	1.02	1.17	1.62	1.57	2.02	1.95
↳ Stacked, 1/M weighted	1.07	1.24	1.02	1.17	1.62	1.64	2.01	1.94
MICE without $Y$								
↳ Rubin's rules	0.97	1.01	0.95	0.99	0.42	0.42	0.45	0.44
↳ Stacked, $f(Y X)$ weighted	1.14	1.21	1.08	1.17	1.91	1.61	2.18	1.81
Bartlett et al. (2014)	1.15	1.27	1.11	1.19	2.02	1.83	2.39	2.01

# DISCUSIÓN





# DISCUSIÓN

- En este trabajo se propusieron cuatro cosas importantes a comparación de anteriores modelos:
  1. Imputar ausencias en las covariables sin tener en cuenta la variable respuesta  $Y$ .
  2. Apilar las múltiples imputaciones para tener un solo conjunto de datos.
  3. Asignar pesos a las observaciones teniendo en cuenta el modelo  $f(Y|X)$ .
  4. Analizar lo anterior con un nuevo estimador para los errores estándar.
- La imputación de datos y el modelado de resultados están separados

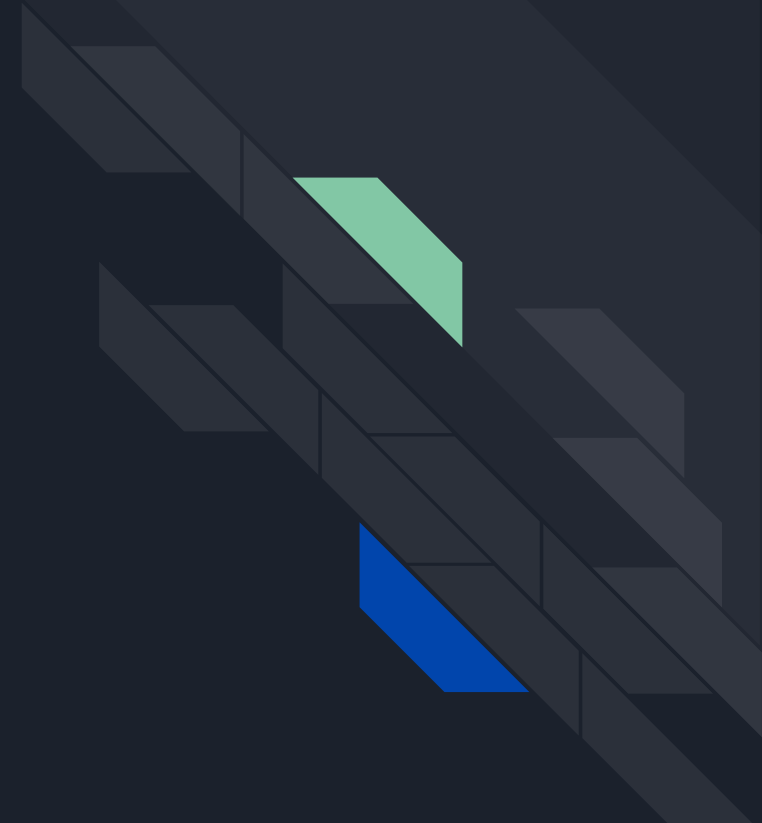


# DISCUSIÓN

- El estimador presentado para el cálculo de los errores estándar en imputaciones múltiples apiladas puede ser también aplicado para el análisis de datos generales de datos imputados múltiples como una alternativa a las reglas de Rubin.
- Una ventaja del análisis propuesto es que se pueden imponer restricciones fácilmente en las estimaciones del modelo a través de las imputaciones múltiples.
- Una desventaja de este enfoque es que requiere el cálculo de la puntuación y las matrices de información para un modelo paramétrico determinado.
- Este estimador se puede implementar fácilmente utilizando el paquete StackImpute.



# EJEMPLO DE APLICACIÓN



# Breast cancer in Wisconsin

Cromatina: Es importante en el diagnóstico de enfermedades, como el cáncer, ya que la cromatina puede mostrar características anormales en las células cancerosas, como la presencia de cromatina más densa o irregular.

Uniformidad del tamaño de la célula


Uniformidad de la forma de la célula

Available at

<https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original>

Variables Table

Variable Name	Role	Type
Sample_code_number	ID	Categorical
Clump_thickness	Feature	Integer
Uniformity_of_cell_size	Feature	Integer
Uniformity_of_cell_shape	Feature	Integer
Marginal_adhesion	Feature	Integer
Single_epithelial_cell_size	Feature	Integer
Bare_nuclei	Feature	Integer
Bland_chromatin	Feature	Integer
Normal_nucleoli	Feature	Integer
Mitoses	Feature	Integer



Title: Statistical Analysis with Missing Data  
Series: Wiley Series in Probability and Statistics  
Author(s): Roderick J. A. Little, Donald B. Rubin  
Publisher: Wiley  
Year: 2019

ISBN: 0470526793; 9780470526798; 9781118596012; 1118596013; 9781118595695; 1118595696

Department of BioStatistics (2017). Título del conjunto de datos.

[<https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original>]. Repositorio de Machine Learning de la Universidad de California, Irvine.